Exploring the assessment and appraisal of regenerative medicines and cell therapy products

Produced by CRD and CHE Technology Assessment Group, University of York, Heslington, York YO10 5DD

Authors Robert Hettle, Honorary Research Fellow, CHE

Mark Corbett, Research Fellow, CRD

Sebastian Hinde, Research Fellow, CHE

Robert Hodgson, Research Fellow, CRD

Julie Jones-Diette, Research Fellow, CRD

Nerys Woolacott, Senior Research Fellow, CRD

Stephen Palmer, Professor of Health Economics, CHE

Correspondence to Stephen Palmer, CHE, University of York, York YO10 5DD

Email: stephen.palmer@york.ac.uk

Date completed 21/12/2015

Source of funding

This report was commissioned by the NIHR HTA Programme as project number 14/151/06.

Declared competing interests of the authors

All authors have no personal or pecuniary conflict of interests.

Acknowledgements

We acknowledge and thank Professor Robert Hawkins (Cancer Research UK Professor of Medical Oncology, University of Manchester and Christie Hospital) and Dr Beki James (Consultant Paediatric Haematologist, Leeds Teaching Hospitals NHS Trust) for their advice on clinical issues; Natalie

Mount and Panos Kefalas from Cell Therapy Catapult for their advice and their comments on the report; and Kath Wright for her help in searching for studies and managing references.

Rider on responsibility for report

The views expressed in this report are those of the authors and not necessarily those of the NIHR HTA Programme. Any errors are the responsibility of the authors.

Contributions of authors

Robert Hettle contributed to the project protocol and conducted the analysis and modelling for the exemplar and wrote sections of the report.

Mark Corbett contributed to the project protocol and was lead reviewer for the clinical issues sections and the clinical review of the exemplar and wrote sections of the report.

Sebastian Hinde contributed to the protocol, conducted the review of issues for cost-effectiveness, contributed to the analysis and modelling for the exemplar, and wrote sections of the report.

Robert Hodgson contributed to the protocol and prepared the clinical issues section on study biases.

Julie Jones-Diette prepared the clinical issues section on surrogate outcomes.

Nerys Woolacott had overall responsibility for the clinical sections of the report. She contributed to the protocol, the preparation of all clinical sections, and the writing of the report.

Stephen Palmer had overall responsibility for the cost-effectiveness sections. He contributed to the protocol and to all aspects of the cost-effectiveness work including the writing of the report.

Abstract

In response to the House of Lords' inquiry into regenerative medicines, an expert group (RMEG) was convened to develop an action plan for the NHS. RMEG proposed that the National Institute for Health and Care Excellence (NICE) commission a "mock technology appraisal" to assess whether changes to its methods and processes are needed. This report presents the findings of independent research commissioned to inform this appraisal and the deliberations of a panel convened by NICE to evaluate the mock appraisal.

The specific objective of our research was to investigate the application of existing NICE appraisal methodology to regenerative medicines, identifying challenges and areas where adaptation may be appropriate. Our research included reviews to identify conceptual differences and the relevance of alternative decision frameworks, alongside the development of an exemplar case study of CAR T-cell therapy for treating acute lymphoblastic leukaemia.

An assessment of previous evaluations of regenerative medicines by NICE and other bodies found that while previous assessments were associated with a number of evidential challenges, none were unique to regenerative medicines or beyond existing methods used to conceptualise uncertainty of a decision.

Regarding the clinical evidence for regenerative medicines, it is not universally the case that they are trialled using non-randomised study designs. However, there may be high variation in response across both individuals and centres. Also regenerative medicines may be subject to continuing development, posing problems when evaluating long-term efficacy and safety. Where single-arm trials are used to assess efficacy the relative treatment effect generated is likely to be optimistic, unless the historical control data are very certain. Pivotal trials may use surrogate endpoints, which, on average, overestimate treatment effects. Also, surrogate primary outcomes are likely to be associated with immature overall survival data. To reduce overall uncertainty, multivariate meta-analysis methods to analyse all available data should be considered. Incorporating indirectly relevant but more reliable (more mature) data into the analysis can also be considered.

For the exemplar case of CAR T-cell therapy, Target Product Profiles (TPPs) were developed which considered the 'curative' and 'bridging to stem-cell transplantation' treatment approaches separately. Within each TPP, three 'hypothetical' evidence sets (minimum, intermediate and mature) were generated to simulate the impact of alternative levels of precision and maturity in the clinical evidence. Subsequent assessments of cost-effectiveness were undertaken, employing the existing NICE reference case alongside additional analyses suggested within alternative frameworks. The

additional exploratory analyses were undertaken to demonstrate how assessments of cost-effectiveness and uncertainty could be impacted by alternative Managed Entry Agreements (MEAs), including price discounts, performance related schemes and technology leasing.

The Panel deliberated on the range of TPPs, evidence sets and MEAs presented, commenting on the recommendations they would be likely to make for each scenario. The Panel discussed a wide range of challenges associated with the exemplar and regenerative medicines more broadly, focussing on the need for a robust quantification of the level of uncertainty in the cost-effective estimates and the potential value of MEAs in limiting the exposure of the NHS to high upfront costs and loss associated with a wrong decision.

Table of Contents

Table of Tables

Table of Figures

List of abbreviations

ACI autologous chondrocyte implantation

ADAPT SMART Accelerated Development of Appropriate Patient Therapies: a Sustainable, Multi-stakeholder Approach from Research to Treatment-outcomes

AE adverse event

ALL acute lymphocytic (lymphoblastic) leukaemia

ASCT Allogeneic stem cell transplantation

ATMP advanced-therapy medicinal product

AYA Adolescent and Young Adult

BiTE therapy, bi-specific T cell engaging therapy

BNF British National Formulary

CAR T-cell Chimeric antigen receptor T-cell

CAT Committee for Advanced Therapies

CEA cost-effectiveness analysis

CHMP Committee for Medicinal Products for Human Use

CMA cost-minimisation analysis

CPI consumer price index

CR Complete remission (or response)

CRD Centre for Reviews and Dissemination

CrI Credible interval

CRS Cytokine release syndrome

CUA cost–utility analysis

DFS Disease free survival

DIC deviance information criterion

EAMS Early Access to Medicines Scheme

EFS Event free survival

EMA European Medicines Agency

EPAR European Public Assessment Report

EQ-5D EuroQol 5 Dimensions

ERG Evidence Review Group

EU European Union

FDA Food and Drug Administration (US)

G-BA Gemeinsamer Bundesausschuss (German Federal Joint Committee for healthcare regulation)

GVHD Graft versus host disease

HRQoL Health related quality of life

HSCT Hematopoietic stem cell transplantation

HTA Health technology assessment

HUI Health utility index

ICER Incremental cost-effectiveness ratio

IMI Innovative Medicines Initiative

IPD Individual patient data

ISPOR International Society for Pharmacoeconomics and Outcomes Research

ITT intention-to-treat

IVIG intravenous immunoglobulin

MACI Matrix applied characterised autologous cultured chondrocyte implant

MAPP Medicines Adaptive Pathways to Patients

MHRA Medicines & Healthcare products Regulatory Agency

MRD Minimal residual disease

NA not applicable

NICE National Institute for Health and Clinical Excellence

NR not reported

NSAID non-steroidal anti-inflammatory drug

OLS ordinary least squares

OMERACT outcome measures in rheumatology

OR odds ratio

OS Overall Survival

p.a. per annum

PFS Progression free survival

Ph+ ALL Philadelphia positive acute lymphoblastic leukaemia

PIMs promising innovative medicines

PPP purchasing power parity

PR partial response/remission

PSA probabilistic sensitivity analysis

QALY quality-adjusted life-year

QoL quality of life

RCT randomised controlled trial

REA Relative effectiveness assessment

RFS Relapse free survival

RM regenerative medicine

RMEG Regenerative Medicine Expert Group

RR relative risk

SA sensitivity analysis

SAE Serious adverse event

SCT stem cell transplant

SD standard deviation

SMR standardised mortality ratio

TKI tyrosine kinase inhibitor

Glossary

Adverse effect

An abnormal or harmful effect caused by and attributable to exposure to a drug which is indicated by some result such as death, a physical symptom or visible illness. An effect may be classed as adverse if it causes functional or anatomical damage, causes irreversible change in the homeostasis of the organism, or increases the susceptibility of the organism to other chemical or biological stress.

Antigen CD19

A protein present on B-cell leukaemias (as well as on healthy B cells)

Aplasia

The failure of an organ or tissue to develop or to function normally

Autologous

Derived from an individual's own cells

Between-study variance

Between-study variance is a measure of statistical heterogeneity that depends on the scale of the outcome measured. It represents the variation in reported study effects over and above the variation expected given the within-study variation.

Biologic therapies (biological)

Medical preparations derived from living organisms. Includes anti-TNF drugs and other new drugs which target pathologically active T cells.

CAR-T cell

Chimeric antigen receptor T-cell

Clinical trial phases

Phase I studies

Researchers test a new drug or treatment in a small group of people for the first time to evaluate its safety, determine a safe dosage range, and identify side effects.

Phase II studies

The drug or treatment is given to a larger group of people to see if it is effective and to further evaluate its safety.

Phase III studies

The drug or treatment is given to large groups of people to confirm its effectiveness, monitor side effects, compare it to commonly used treatments, and collect information that will allow the drug or treatment to be used safely.

Phase IV studies
Studies are done after the drug or treatment has been marketed to gather information on the drug's effect in various populations and any side effects associated with long-term use.

Consolidation chemotherapy
Chemotherapy given once a remission is achieved, to sustain a remission.

Cost-benefit analysis
An economic analysis that converts the effects or consequences of interventions into the same monetary terms as the costs and compares them using a measure of net benefit or a cost-benefit ratio

Cost-effectiveness analysis
An economic analysis that expresses the effects or consequences of interventions on a single dimension. This would normally be expressed in 'natural' units (e.g. cases cured, life-years gained, additional strokes prevented). The difference between interventions in terms of costs and effects is typically expressed as an incremental cost-effectiveness ratio (e.g. the incremental cost per life-year gained).

Cost-utility analysis
The same as a cost-effectiveness analysis but the effects or consequences of interventions are expressed in generic units of health gain, usually quality-adjusted life years (QALYs).

Credible Interval
In Bayesian statistics, a credible interval is a posterior probability interval estimation which incorporates problem-specific contextual information from the prior distribution. Credible intervals are used for the purposes similar to those of confidence intervals in frequentist statistics.

Fixed-effect model
A statistical model that stipulates that the units under analysis (e.g. people in a trial or study in a meta-analysis) are the ones of interest, and thus constitute the entire population of units. Only within-study variation is taken to influence the uncertainty of results (as reflected in the confidence interval) of a meta-analysis using a fixed effect model.

Graft rejection
Rejection of transplanted organs as a result of humoral and cell-mediated responses by the recipient to specific antigens present in the donor tissue.

Haematologic cancers
Cancer of blood cells which can be sub-divided into three main diseases: leukaemia, lymphoma and myeloma.

Heterogeneity
In systematic reviews heterogeneity refers to variability or differences between studies in the estimates of effects. A distinction is sometimes made between "statistical heterogeneity" (differences in the reported effects), "methodological heterogeneity" (differences in study design) and "clinical heterogeneity" (differences between studies in key characteristics of the participants, interventions or outcome measures).

I-squared (I^2)
I-squared (I^2) is a measure of "statistical heterogeneity" (differences in the reported effects). It varies between 0 and 1, where 0 indicates that the differences in reported effects are entirely consistent with the within-study uncertainty, and 1 indicates that the differences in reported effects are entirely explained by study characteristics that vary across studies.

Immune reconstitution
A condition where the patient's immune system begins to recover after treatment

Immune reconstitution inflammatory syndrome
A condition where the patient's immune system begins to recover after treatment but then reacts later with an overwhelming inflammatory response

Immunoconjugates
Antibodies joined to a second molecule, usually a toxin, radioisotope or label for use in immunotherapy. .

Immunotoxins
A protein that consists of a targeting portion linked to a toxin which will bind to a cell and causes endocytosis allowing the toxin to kill the cell.

Immunophenotype
The protein type expressed by cells.

Immunotherapy
A treatment designed to boost the body's natural defences to fight cancer by utilising material either from the body or produced in vitro to improve, target, or restore immune system function.

Intention-to-treat
An intention-to-treat analysis is one in which all the participants in a trial are analysed according to the intervention to which they were allocated, whether they received it or not.

Medical devices directive
The Medical Device Directives is a directive relating to the safety and performance of medical devices which were harmonized in the EU in the 1990s.

Monoclonal antibody
An antibody produced in a laboratory from a single clone that recognizes only one antigen.

Open-label study
A type of study in which both participants and researchers know which treatment is being administered.

Orphan designation/status
Based on the EMA criteria, a medicine can qualify for orphan status if;

it is intended for the treatment, prevention or diagnosis of a disease that is life-threatening or chronically debilitating; the prevalence of the condition in the EU must not be more than 5 in 10,000 or it must be unlikely that marketing of the medicine would generate sufficient returns to justify the investment needed for its development; no satisfactory method of diagnosis, prevention or treatment of the condition concerned can be authorised, or, if such a method exists, the medicine must be of significant benefit to those affected by the condition.

Pharmacodynamic effects
The study of how a drug behaves in the body.

Pharmacokinetic effects
The study of the effect the body has on a drug.

Placebo
An inactive substance or procedure administered to a patient, usually to compare its effects with those of a real drug or other intervention, but sometimes for the psychological benefit to the patient through a belief that s/he is receiving treatment.

Persistence

In treatment intended for direct in vivo administration, persistence may describe how long the product is effective in treating a targeted disease. It may also be used to refer to the persistence of the product, e.g. gene expression or any permanent changes, within the patient as a result of treatment with the product.

Quality Adjusted Life Year (QALY)

An index of health gain where survival duration is weighted or adjusted by the patient's quality of life during the survival period. QALYs have the advantage of incorporating changes in both quantity (mortality) and quality (morbidity) of life.

Quality of Life

A concept incorporating all the factors that might impact on an individual's life, including factors such as the absence of disease or infirmity as well as other factors which might affect their physical, mental and social well-being.

Random effects model

A statistical model sometimes used in meta-analysis in which both within-study sampling error (variance) and between-studies variation are included in the assessment of the uncertainty (confidence interval) of the results of a meta-analysis.

Randomised controlled trial (RCT)

An experiment in which investigators randomly allocate eligible people into intervention groups to receive or not to receive one or more interventions that are being compared.

Refractory

Disease which does not respond to attempted forms of treatment.

Regenerative medicine

A field of research and clinical applications dealing with the process of replacing or regenerating human cells, tissues or organs to restore or establish normal function

Relative Risk (RR) (synonym: risk ratio)

The ratio of risk in the intervention group to the risk in the control group. The risk (proportion, probability or rate) is the ratio of people with an event in a group to the total in the group. A relative risk of one indicates no difference between comparison groups. For undesirable outcomes an RR that is less than one indicates that the intervention was effective in reducing the risk of that outcome.

Salvage chemotherapy

Chemotherapy given to a patient when all other treatment options are exhausted

Sensitivity analysis

An analysis used to determine how sensitive the results of a study or systematic review are to changes in how it was done. Sensitivity analyses are used to assess how robust the results are to uncertain decisions or assumptions about the data and the methods that were used.

Time to relapse

Length of first remission

Weighted mean difference (in meta-analysis)

A method of meta-analysis used to combine measures on continuous scales, where the mean, standard deviation and sample size in each group are known. The weight given to each study is determined by the precision of its estimate of effect and, is equal to the inverse of the variance. This method assumes that all of the trials have measured the outcome on the same scale.

Plain English Summary

Regenerative medicines replace or regenerate human cells, tissues or organs to restore or establish normal function. Potential breakthroughs in this area of clinical research are eagerly anticipated and expectations are often high due to the possibility of cures (or substantial improvements) for diseases which are currently deemed chronic or fatal. However, the assessment of the long term costs and benefits of such therapies is more difficult than for conventional treatments.

In response to a House of Lords' inquiry into regenerative medicines, an expert group (RMEG) was set up to develop an action plan for the NHS. RMEG proposed that the National Institute for Health and Care Excellence (NICE) commission a "mock technology appraisal" to assess whether changes to its methods and processes are needed.

This report presents the findings of independent research commissioned to inform this appraisal. We reviewed evaluations of regenerative medicines by NICE and other groups as well as conducting reviews of the existing literature concerned with the challenges of assessing the therapies. In addition an exemplar case study of CAR T-cells for acute leukaemia was constructed to inform the deliberations of an expert panel set up by NICE.

Our research found that, while evidence about regenerative medicines is expected to be associated with much uncertainty in determining the long term costs and benefits to patients and the NHS, the existing methods available to estimate the implications of this uncertainty are sufficient. Ways of sharing the risks between the NHS and the therapy manufacturers should be investigated further.

1 Scientific Summary

1.1 Background

Regenerative medicines replace or regenerate human cells, tissues or organs to restore or establish normal function. Potential breakthroughs in this area of clinical research are eagerly anticipated and expectations are often high due to the possibility of cures (or substantial improvements) for diseases which are currently deemed chronic or fatal. However, efficacy, safety and cost-effectiveness evaluations of regenerative medicines may be difficult compared with evaluations of conventional pharmaceutical treatments. This was highlighted by the Regenerative Medicine Expert Group (RMEG), which was tasked to develop an NHS regenerative medicine delivery readiness strategy and action plan. It reported that application of NICE appraisal methods, based on cost utility analysis, to products whose true value may not be known for many years, might be challenging. This is mainly because of the inherent uncertainty of estimating long-term benefit from evidence which may have been derived from small, short-term, non-randomised studies which may also have used surrogate primary outcomes (rather than real clinical outcomes). Such study methods are often an inevitable consequence of the seriously ill, very small populations with unmet medical needs which are often the initial target of new regenerative medicines. Important potential challenges to assessing cost-effectiveness may include claims of cure and lifetime benefit; longer-term safety issues caused by persistence and possible product changes within the body; organisational and scaling issues; and significant up-front costs. Investigation is therefore needed to determine whether the conceptual differences between regenerative medicines and other types of health technology require different approaches to the conduct and assessment of efficacy, safety and cost-effectiveness.

1.2 Objectives

- To test the application of NICE appraisal methodology to regenerative medicines, identifying challenges and any areas where methods research and/or adaptation of methodology is appropriate.
- To identify specific issues related to the appraisal of regenerative medicines using the current NICE appraisal process and decision framework.
- To develop a framework for those developing regenerative medicines to facilitate understanding of how NICE evaluates clinical and cost-effectiveness and to identify the most important evidence areas to develop before cost-effectiveness can be reasonably estimated.

1.3 Potential issues for the evaluation of clinical effectiveness

Two different approaches were taken to identify and explore issues and challenges which may be associated with NICE evaluations of regenerative medicines:

- Perform a broad exploration of the applicability of NICE technology appraisal methods to regenerative medicines;
- Undertake an exemplar NICE appraisal of a hypothetical regenerative medicine product.

Several reviews were undertaken to identify and discuss technology appraisal methodology issues which may be particularly relevant to regenerative medicines: a review of EMA, NICE and FDA assessments of regenerative medicines licensed in the EU; a review of the use of surrogate endpoints in clinical research; and a review of the biases likely to affect results of non-randomised studies (with a particular focus on the challenges of using results from single-arm trials to estimate efficacy). These reviews were performed pragmatically, i.e. they were not intended to be systematic, since the aim was to identify key issues, rather than to identify *all* relevant studies and papers.

Several broad issues which may affect uncertainty were apparent from these reviews:

- It is not universally the case that regenerative medicines are trialled using non-randomised study designs: submitted pivotal studies may well in fact be randomised, notably when levels of unmet need are low and diseases/conditions are not rare; in such cases the maturity of data (which would be available at the time of a NICE appraisal) has been up to five years' duration.
- With single-arm trials a key consideration when judging levels of uncertainty should be the likelihood of cure or improvement *without* experimental treatment. However, it may be very difficult to identify published prognostic studies which have suitable historical control data. Other strategies for obtaining historical data may therefore be needed, such as seeking access to national patient databases.
- When single-arm trials or case series *do* form the basis of a regulatory submission, unless the historical control data are very certain, the relative treatment effect generated is likely to be optimistic.
- Where single-arm trial data are compared with historical data and appropriate methods to adjust for confounding are employed, the selection of the method used must be explicit and be based on sound reasoning; despite advances in statistical techniques, clear challenges remain in generating accurate unbiased estimates of effect from non-randomised data.
- Pivotal trials in regulatory submissions are likely to report primary endpoints which are surrogates for real clinical endpoints. On average, trials using surrogates report larger treatment effects than trials using final patient-relevant outcomes. This has implications for effect estimate

uncertainty, especially when *only* surrogate endpoints are reported. The choice of surrogate outcomes must be researched, explicit and justified, preferably via a systematic review of the evidence for the validation of the surrogate-final outcome relationship. Nevertheless, to maximise the use of *all* available data, and to reduce overall uncertainty, multivariate meta-analysis methods to analyse data should also be considered - whatever the maturity of the evidence base.

- Related to the issue of surrogates as primary outcomes is that of duration of follow-up: use of intermediate shorter-term outcomes avoids the need for long follow-up. The consequence of this is that even where overall survival data are recorded, these data are immature at the point of regulatory approval.
- The high technology status of regenerative medicines may imply greater potential for variation in response across both individuals and centres. This is likely to have implications in terms of the generalisability of efficacy and safety estimates obtained from small single-centre (probably expert centre), single-arm studies. Single centre studies may produce larger effect estimates than multi-centre studies. In the absence of larger or more varied trials, this issue might only be addressed by access to individual patient data so that potential predictors of response or effect modifiers may be investigated.
- Although more mature evidence - such as confirmatory RCTs - may sometimes be viable in the specific population being evaluated, such evidence might also only be expected in larger, similar populations (for example, B-ALL patients in *first* relapse). This raises the possibility of incorporating indirectly relevant but more reliable (and possibly more mature) data into the analysis, to reduce uncertainty.
- Most regenerative medicines are, by their nature, innovative products which may be subject to continuing development, with new generations of product having improved efficacy. This may pose problems when evaluating long-term efficacy and safety; for example, to what extent can the long-term safety data from a first-generation product be used to inform long-term safety of a related newly-licensed second-generation product? This may mean that as well as bioavailability-type studies, key trials conducted earlier in the development process may have to be replicated or adjustments be made in the analyses of trial data to account for their indirectness.

1.4 Potential issues for the evaluation of cost-effectiveness

We undertook a pragmatic review of potential cost-effectiveness issues to identify possible conceptual differences between regenerative medicines and more conventional medicines to identify potential methodological challenges for cost-effectiveness assessments.

NICE has previously evaluated two regenerative medicines within the existing TA process: autologous chondrocyte implantation (ACI) for the treatment of cartilage defects in the knee joints

and sipuleucel-T (Provenge®) for treating asymptomatic or minimally symptomatic metastatic hormone-relapsed prostate cancer. Several issues and common themes emerged from our review of these:

- While the innovative nature of both products was acknowledged, neither committee considered there to be any evidence supporting a distinct benefit that would not be appropriately reflected in the reference case measure of QALYs;
- The high levels of uncertainty were highlighted a key issues in both appraisals leading to important concerns regarding the robustness of the cost-effectiveness results;
- The generalisability and transferability of evidence across different generations appears an important issue, particularly regarding longer-term outcomes which may not be sufficiently mature for newer generations;
- Important uncertainties were also identified surrounding the costs that would be incurred by the NHS, due to limited experience with their use and also the complex pathway in which they were used.

Many of the challenges identified do not appear unique to these types of therapies and are also likely to be faced by manufacturers of more conventional pharmaceuticals, biologics and devices. However, it is clear that these challenges may be faced more routinely for regenerative medicines and cell therapies. Concerns surrounding the potential high-upfront costs of regenerative medicines and affordability to health care systems have also received particular attention in the literature, leading several authors to conclude that alternative financing approaches may also have to be considered.

Many of the issues associated with regenerative medicines will inevitably impact on the level of uncertainty associated with the cost-effectiveness of the technology when introduced into clinical practice. Even where products have significant potential to confer important clinical advances over current therapies, this may not be known with a high level of certainty at the time a regenerative medicine is licensed. Inevitably a new technology's cost-effectiveness may be more difficult to determine in these circumstances and schemes that allow the development of further evidence or entail a risk-sharing component may be required. Managed entry agreements (MEAs) are an increasingly common policy response to dealing with uncertainty in the evidence base of new health technologies entering the market.

How to determine when efficiency is sufficiently weak or uncertain, such that MEAs are appropriate policy responses remains a key methodological issue that has important implications both for policy

making and research investments made by the regenerative medicine industry. Several studies have concluded that reimbursement decisions and the possible use of MEA should be based not only on the expected value of a technology but also the value of further research, the anticipated effect of coverage on further research, and the costs associated with reversing the decision (i.e. irrecoverable costs).

Importantly, there already exists provision within NICE's methods guide to accommodate some of these considerations, although potential challenges may arise in ensuring these are consistently applied between committees and understood by manufacturers. NICE will also need to consider whether further amendments to their processes and methods are required. Broader consideration will also need to be given to approaches which may extend beyond NICE's existing remit e.g. alternative funding approaches. Consequently other bodies and manufacturers themselves may also have an important role in identifying more innovative approaches to seeking reimbursement which recognise the inherent uncertainties and lead to a more efficient sharing of associated risk.

1.5 Exemplar NICE appraisal of CAR T-cells for relapsed/refractory B-ALL

For the exemplar appraisal, the chosen hypothetical product was CAR (chimeric antigen receptor) T-cell therapy specific to the antigen CD19, for treating relapsed (two relapses or more) or refractory B-cell acute lymphoblastic leukaemia (B-ALL). This combination was selected based on the existence of relatively mature data sets – in the context that none of the currently available CAR T-cell products are licensed. The exemplar began with a review of both the completed, and ongoing, trials of CAR T-cell therapies for B-ALL. This was followed by a review of other treatments for relapsed/refractory B-ALL which have been licensed by the EMA or FDA.

1.5.1 Clinical evidence for the efficacy and safety of CAR T-cells for relapsed/refractory B-ALL

The available trial data for CAR T-cells is limited to small, single-arm studies. The therapies have the potential to offer patients a 'bridge' to a stem cell transplant or possibly a cure (without transplant), depending on the particular type of CAR T-cell therapy. However, potentially serious adverse effects are also possible. The relapsed/refractory B-ALL population is narrowly defined with extremely poor prognosis and limited alternative therapy options. The length of persistence of CAR T-cells within patients will have implications for both efficacy and safety; the ideal product would balance the trade-off between persistence being long-enough to eradicate malignant cells, and short enough to prevent B-cell aplasia being a problematic adverse effect. Length of persistence may therefore dictate whether CAR T-cells will be administered with either curative intent, or as a bridge to a stem cell transplant.

The target product profile (TPP) and hypothetical evidence sets

Based on the available clinical evidence for CAR T-cell and licensed non-regenerative medicines for relapsed B-ALL, two target product profiles (TPPs) were developed to be considered as part of the exemplar appraisal:

- CAR T-cell therapy used as a "*bridge to HSCT*", where the primary goal of treatment is to induce short-term remission of disease in order to maximise the opportunity for successful HSCT;
- CAR T-cell therapy used with "*curative intent*", where the primary goal of CAR T-cell treatment is long-term remission/cure of disease (with or without HSCT).

These two approaches to treatment with CAR T-cell therapy imply two potentially different contexts in which therapy may be appraised. Consequently, there are separate implications arising from the different applications that require their consideration as two distinct scenarios.

To explore the impact of different levels of precision and maturity in the evidence base, three hypothetical data sets were constructed for each TPP:

- The minimum set (60-80 patients, median follow-up approx. 10 months): the minimum data considered potentially sufficient for CAR T-cell therapy to be granted conditional regulatory approval;
- The intermediate set (60-80 patients, maximum follow-up of 5-years): a variant of the minimum set, where the efficacy and safety of CAR T-cell therapy has been assessed over a longer follow-up period;
- The mature set (120-140 patients, maximum follow-up of 5-years): a variant of the intermediate set where the efficacy and safety of CAR T-cell therapy has been assessed in a larger clinical study but with a similar follow-up period as the intermediate set.

In total, six evidence sets were developed spanning the separate TPPs (three sets for *"bridge to HSCT"*, and three sets for *"curative intent"*). Each of the three evidence sets included hypothetical efficacy and safety data for CAR T-cell therapy and for a historical control. For consistency across TPPs and evidence sets, the same historical control and data was considered in all scenarios.

1.5.2 Development of exemplar cost-effectiveness model – summary of approach and key findings

Two *de novo* decision models were developed to assess the cost-effectiveness of CAR T-cell therapy within the two separate TPPs ("*bridge to HSCT"* and "*curative intent"*) across each of the separate evidence sets. Although, a number of common inputs and assumptions were employed across both

models, the two models had important structural differences which led to differences both in the underling modelling approach as well as in the use of external evidence.

In the *"bridge to HSCT"* scenario, the primary health benefits of treatment with CAR-T cell therapy were assumed to be driven by an increase in the proportion of patients receiving HSCT and the subsequent success of HSCT itself (based on remission and MRD status). The introduction of an epidemiological "link" between a potential established surrogate outcome/process (i.e. MRD and HSCT status) and final health benefits (i.e. overall survival and QALYs) also enabled the use of external evidence to be utilised alongside the separate hypothetical evidence sets generated. A landmark responder model was developed, incorporating evidence from the hypothetical evidence sets to inform short-term outcomes on remission, HSCT and MRD status and external evidence to estimate overall survival conditional on these shorter-term outcomes.

The key assumption employed within this scenario is that external evidence substantiating the relationship between MRD and HSCT status in studies in which CAR T-cells have not been used can be generalised to patients in whom CAR-T cells have been used. Importantly, results of our validation work appears to demonstrate that, with minor calibration and adjustment, the combination of trial reported evidence on short-term outcomes (remission, HSCT and MRD status) and external evidence appeared to closely match the overall survival estimates directly reported within the studies used to generate the evidence sets for CAR T-cell therapy and the comparator (clofarabine).

In the "*curative intent*" model, a different assumption was employed; specifically that the CAR T-cell therapy itself potentially confers longer-term and potentially curative benefits without the need to bridge to HSCT. In this context, the case for use a structural link between final health benefits and a surrogate outcome or process such as HSCT appears more limited. Instead, a simple three state partitioned survival model was developed to model long-term outcomes via the direct extrapolation of overall survival data from the evidence sets.

An important consideration within the *"curative intent"* model was whether the use of conventional parametric survival functions (e.g. exponential, Weibull, log-normal etc.) would adequately capture a less conventional hazard function that might be observed for a potentially curative treatment; and how this might be affected by different levels of precision and maturity of evidence. Consequently, our work considered the goodness of fit of conventional survival functions and more flexible "spline-based" survival models. A key finding was that the more flexible survival models appeared to more closely approximate the hazard function across each of the evidence sets. Although the use of these more flexible survival models are briefly discussed within existing NICE technical support documents, we are not aware of any examples of their use to date within the NICE TA process.

Consequently, further research may be required to more formally consider the appropriateness of alternative survival modelling approaches to regenerative medicines and cell-based therapies, including more flexible models and cure fraction models.

The importance of the level of data maturity in deriving robust survival projections for the economic model was evident in our results from the *"curative intent"* model. Whilst the spline models appeared to generate a robust fit to the data over the first 3 months of the KM estimate used in the minimum dataset, the functions were not able to accurately predict the tail of distribution. Furthermore, considerable variation was evident in the predicted long-term survival of the modelled cohort employing different parametric functions. We concluded that it was unlikely that a single survival distribution could adequately characterise uncertainties over the longer-term extrapolation period in the minimum evidence set.

To more formally account for the uncertainty surrounding choice of survival distribution, a model averaging approach was adopted. This technique involves the parameterisation of uncertainty surrounding the choice of distribution, combining results from a series of alternative survival functions as part of a weighted distribution. This approach samples both the parametric uncertainty associated within each distribution and the uncertainty (or weights) surrounding the choice of preferred method. Through the probabilistic analysis, it is therefore possible to estimate the joint distribution of uncertainty around the parameter estimates and the choice of survival function.

In contrast to the minimum set, the additional data maturity in the intermediate and mature evidence sets results in greater certainty over the long-term survival benefits of treatment. This leads to reduced variability in the potential trajectories for the survival benefits of treatment. In addition, with more mature evidence, the fitted survival models are better able to predict the tail of the KM. Therefore, unlike the *"bridge to HSCT"* model, additional evidence maturity in the *"curative intent"* model leads to different projections of survival benefit, as well as impacting on the parametric uncertainty surrounding model extrapolations.

Given the inevitable uncertainties which are likely to exist regarding the longer-term benefits of regenerative medicines and cell-based therapies, further methodological research could be usefully undertaken to help inform how these uncertainties might be appropriately quantified in a transparent manner to inform subsequent decisions. A key consideration here would be the extent to which potential weights can be defined prior to the Committee's deliberations or should be more directly informed by them. Given the potential complexity in both undertaking these analyses and communicating the results, more efforts should be made to ensure models are developed to ensure that informal judgements can be more explicitly incorporated in a timely and transparent manner.

Assumptions, strengths, limitations

A key assumption employed within both models is that from year 5 onwards in the model, all patients who remained alive were assumed to experience a similar mortality risk profile consistent with a long-term survivor of ALL. This assumption reduced some of the longer-term uncertainties that inevitably arise from the extrapolation of the data beyond the maximum reported follow-up across the evidence sets considered for CAR T-cell therapies. Additional follow-up data could be used to test the validity of such an approach any claims of longer-term benefits.

Our searches to inform other model parameters identified further important uncertainties. The existing HRQoL data in ALL was limited and several assumptions were required. Importantly, no existing CAR T-study had incorporated measures of HRQoL that could be considered directly in the model. In the absence of this data, assumptions were made based on external studies to account for the possible magnitude of HRQoL benefits of achieving remission, alongside any negative impacts due to the mode of therapy (i.e. HSCT, chemotherapy) and specific adverse events, specifically CRS and B-cell aplasia.

Finally, our research also identified important uncertainties regarding key elements of the CAR T-cell process (e.g. leukapheresis, conditioning therapies, level of hospitalisation required for different aspects such as conditioning, subsequent administration and monitoring etc.) and associated costs. Additional evidence needs to be provided by manufacturers to enable these potential costs to the NHS to be calculated more accurately.

The costs of HSCT and any additional costs that may arise due to longer-term management of patients were also an important source of uncertainty. A variety of possible sources were identified in our review and important differences observed across these. Further studies would be useful to more formally cost the short and longer-term implications of HSCT in paediatric populations and to also determine the generalisability of studies reporting estimates from outside the UK.

Although the existence of possible learning curves was identified as an important issue in the conceptual review, these were not directly considered within the exemplar. Some aspects of these may become more apparent as larger studies report, particularly involving centres with different levels of expertise. As experience with using CAR T-cell therapies develops, this may have important implications for both the identification and management of potential adverse events, as well as provision of the therapy itself. An assumption is made in the exemplar model is that the different stages of the process for CAR T-cells would require separate hospitalisation (i.e. for the initial conditioning therapy and later for the subsequent administration of the CAR T-cells and subsequent monitoring). However, aspects of the process may evolve over time such that subsequent

administration and monitoring may be undertaken in a less resource intensive setting. Although the existence of learning curves has received significant attention in the clinical literature, to date the implication for and application within cost-effectiveness analysis remains limited and warrants further investigation.

Finally, an important assumption made within the exemplar relates to the acquisition cost of CAR T-cell therapy itself. In the absence of a commercially available product and published price, an assumption was made that the manufacturer would employ a value-based approach to their decision such that the resulting cost-effectiveness (ICER) estimate was close to NICE's cost-effectiveness threshold. In the context of the exemplar, this was assumed to be based on the maximum range of the threshold considered by NICE assuming the existing *"End of Life"* criteria are met. Importantly, this price is not considered to be indicative of the final acquisition cost that might be set when commercially available products are available. Neither are we presuming that NICE's current *"End of Life"* criteria would apply. Instead, the basis for setting the price on the basis of an existing cost-effectiveness threshold was to enable different interested parties to better understand the potential impact of other uncertainties (e.g. precision and maturity of evidence) within NICE's current decision making process, identifying potential trade-offs that may exist and illustrating how these uncertainties might be more explicitly addressed within different MEAs (i.e. evidence generation and/or pricing schemes). Although it is clearly possible to examine a range of different possible prices for the CAR T-cell therapies within the exemplar, it was considered that this approach may result in the subsequent Panel decision process becoming unmanageable (i.e. multiple pricing scenarios) and would lessen the generalisability learning which the exemplar was developed to highlight.

1.5.3 Assessment of cost-effectiveness, uncertainty and the value of alternative policy options – summary of approach and key findings

The primary purpose of these assessments was to estimate the potential cost-effectiveness of CAR T-cell therapy within the separate scenarios considered and to highlight key uncertainties surrounding these results. An important aspect of this work was also to consider how these estimates could be presented and communicated to the Panel to inform their deliberations. In doing this we considered analyses based on approaches routinely requested within NICE's existing methods guide. We also undertook additional analyses that may provide useful additional insights to help inform subsequent committee deliberations and the potential nature of such analyses.

The sequence of assessments presented started with a conventional assessment of cost-effectiveness at the patient level based on the minimum evidence set. Disaggregated estimates of the costs and outcomes were estimated, together with resulting cost-effectiveness estimates based on the ICER. These results were also expressed using Net Health Effects (NHEs), representing the difference

between any health gained with the intervention and health foregone elsewhere in the health care system, expressed either in monetary and QALY terms.

The impact of uncertainties was explored using conventional one-way sensitivity analyses (i.e. varying individual parameters or specific assumptions) and probabilistic approaches (i.e. exploring the impact of joint uncertainty across all parameters). Conventional scatter-plots and acceptability curves were utilised to graphically show the impact of parameter uncertainties and other methodological uncertainties (e.g. the appropriate discount rate). The analyses also explored the potential impact if the Panel were to consider the criteria met for applying the non-reference-case discount rate of 1.5% for costs and health effects.

Impact on the NHS

In addition to the analyses undertaken using the conventional reference case approaches, a series of more exploratory analyses were also undertaken. In particular, the per-patient assessments were subsequently scaled up to population assessments, requiring an estimate of the number of potentially eligible patients (assumed to be approximately 38 patients per annum) and an assessment of the *"technology time-horizon"* i.e. the period over which the therapy might be utilised within clinical practice (assumed to be 10 years in the exemplar). Although the presentation of population-level analyses are not formally requested within the existing NICE Methods Guide for reporting cost-effectiveness results, an assessment of population impact is required within Section 5.12 (Impact on the NHS). Hence, these exploratory analyses were considered to be consistent with the NICE Methods Guide's requirement to consider population impact and the specific requests within Section 6.4.1 (Research recommendations) for the committee to balance the potential NHEs of current and future NHS patients when considering making research recommendations.

The results of the population based analyses were summarised in terms of incremental NHE (both in terms of QALYs and equivalent monetary value) together with an assessment of the probability that CAR T-cells were cost-effective. Alongside these more conventional assessments, an assessment of the scale of the likely consequences of uncertainty was considered to be potentially informative to the Panel, particularly in deliberations related to possible research recommendations. An estimate of the consequences of existing decision uncertainty was subsequently derived, reflecting the potential magnitude of NHEs that could be gained if uncertainty surrounding this decision could be resolved immediately.

Impact of alternative pricing scenarios

Using the different analyses, the impact of alternative pricing scenarios were also explored, including conventional PAS type schemes (i.e. equivalent to a fixed price reduction) as well as more

sophisticated schemes based on pay for performance and leasing approaches. Similarly, the impact of the alternative evidence sets was explored to establish the implications of increased precision and maturity assumed in the intermediate and mature evidence sets.

Quantifying potential uncertainties

An important consideration within this work is the extent to which current NICE methods and processes are likely to appropriately quantify the potential uncertainties surrounding regenerative medicines and cell-based therapies to ensure that appropriate policy decision are made regarding adoption and spread of potentially promising technologies. Our findings show that the conventional assessments requested within the current TA process may not be sufficient. Estimates of the ICER and associated probability that CAR-T therapy is cost-effective at a specific threshold, were shown to be virtually identical in one of the TPPs despite being based on 3 different evidence sets with varying levels of precision and maturity. Similarly, across both TPPs, several of the alternative pricing schemes again reported similar estimates of the ICER and associated probabilities that CAR-T were cost-effective. Consequently, it is unclear how these differences would be reflected within the current deliberative process. Whilst it is acknowledged that different conclusions might be reached by the Panel based on informal judgements, the importance of ensuring transparency in subsequent decisions remains a key principle of the Institute and appears critical for manufacturers in developing appropriate R&D and pricing strategies.

Presentation of the scale of consequences using population NHE appeared to provide a clearer distinction between the different evidence sets and an assessment of the impact of alternative pricing schemes. Consequently, their more routine application within the TA process for regenerative and cell-based therapies may be an important consideration for future processes. Furthermore, as demonstrated by the exemplar, comparisons of this nature could also provide a more transparent and explicit basis for considering the value of direct price reductions that might be realised via a conventional PAS compared to the provision of additional evidence, in terms of reducing decision uncertainty and its consequences. Such information might provide an important basis for discussions between manufacturers, NICE and other relevant parties in terms of how the existing uncertainties that exist might be appropriately managed ensuring risks and benefits are more appropriately shared.

1.6 Issues arising from the NICE panel meeting

A separate panel and meeting were convened by NICE to discuss the findings of the exemplar appraisal. The panel included clinical experts and current and past NICE committee members and was chaired by Professor Andrew Stevens (current chair NICE TA committee). The objective of the panel meeting was to assess the clinical and cost-effectiveness evidence informing the separate TPPs and to

identify potential issues and challenges for the NICE TA appraisal process and methods. A summary of the clinical and cost-effectiveness evidence was presented to the panel, who were asked to deliberate on the range of scenarios and provide 'hypothetical' decisions and outline the main considerations for these.

The key consideration relating to the clinical effectiveness was how decisions can be made for technologies that look highly promising but for which the evidence base is highly uncertain, at a potentially high, but actually unknown, risk of bias, and extremely immature. As the context for appraisals of regenerative medicines is likely to be in the context of conditional regulatory approval, the panel considered that is would be important to know what research had been mandated by the regulator, and hence what uncertainty could reasonably be expected to be resolved in the near future. There was concern regarding the difficulty of decommissioning services following (what later proved to be) incorrect recommendations.

A key consideration regarding the cost-effectiveness results and implications for the 'hypothetical' decisions was whether the panel considered that existing criteria considered within the TA process in relation to End of Life (EoL) could be applied. The panel accepted that the exemplar met the requirements of the EoL criteria but concluded that other considerations (e.g. innovation) would not be applied in addition. They further noted that the criteria may need to be reconsidered given the large extension of life expectancy possible with regenerative medicines.

The panel raised issues regarding the possible nature and magnitude of any irrecoverable costs that might be incurred by the NHS and the implications for their decisions. The panel acknowledged that the different pricing schemes had important impacts both in terms of the ICER but also in terms of the allocation of any risk between the NHS and manufacturers. The concept of the 'leasing approach' was identified as a potentially important option.

The panel discussed the additional evidence sets that had been generated for each TPP (Scenarios 3 and 4). The panel understood that the scale of consequences was reduced in the more mature evidence sets due to increased precision. The panel acknowledged the challenges and difficulties of generating mature evidence at the point a product is launched. In particular, the panel noted that a comparison of the magnitude of the incremental NHE and the consequences of decision uncertainty provided an important starting point for deliberations in considering the scale of the NHE that could be achieved by immediate approval and that which might be achieved by further research.

Overall, the panel noted that the exploratory approaches provided a clearer and potentially important distinction between the different evidence sets and the impact of alternative pricing schemes.

However, the panel expressed difficulty in reaching consensus decisions for the range of scenarios presented without a formal reference point to establish whether the consequences were sufficiently high to impact on their decisions and/or potential research recommendations.

1.7 Summary conclusions

Our research found that the clinical evidence about regenerative medicines is expected to be associated with much uncertainty. Existing methods are available to adjust for and minimise the risk of bias and uncertainty in data analyses. Whilst there will be a significant level of uncertainty in determining the long term costs and benefits to patients and the NHS, the existing methods available to estimate the implications of this uncertainty are sufficient. The use of risk sharing agreements (Managed Entry Schemes) between the NHS and manufacturers of regenerative medicines should be investigated further.

2 Introduction and aims

The term regenerative medicine refers to a field of research and clinical applications dealing with the process of replacing or regenerating human cells, tissues or organs to restore or establish normal function.[1] Regenerative medicine is not a new field of medicine, since it encompasses bone marrow or organ transplants. However, the development of newer types of regenerative medicine such as cell-based therapies (often using stem cells or progenitor cells to produce tissues), gene therapy, and tissue engineering has raised the possibility that diseases which are currently deemed chronic or fatal may be curable. Most regenerative medicines will be classed by the European Medicines Agency (EMA) as being 'advanced-therapy medicinal products' (ATMPs) which are essentially treatments based on engineered cells or tissues. Although regenerative medicines may offer great potential, the route to this new era of medicine might not be straightforward. Product development and production to commercially viable levels may hold many challenges. Furthermore, efficacy and safety evaluations of regenerative medicines may be difficult when compared with conventional pharmaceutical treatments. For example, while the adverse effects of pharmaceuticals are likely to improve when they are discontinued, some regenerative medicines may cause prolonged toxicities, especially when cells persist long-term; such adverse effects might also not become evident for years.

A House of Lords Science and Technology Committee inquiry into regenerative medicine was set up to pinpoint the UK's strengths in this area, to identify barriers to translation (applying findings from basic research to a clinical setting) and commercialisation (primarily delivering treatments in the healthcare market), and to recommend solutions. The report - published in July 2013 - concluded that although the UK has a great potential resource in the NHS which could make it an attractive place for investment, it is currently underprepared to realise the full potential of regenerative medicine. One of the report's recommendations was that the Department of Health should establish a regenerative medicine expert working group to develop an NHS regenerative medicine delivery readiness strategy and action plan, and report back to the Secretary of State for Health by December 2014.[2] In response to this, the Regenerative Medicine Expert Group (RMEG) was convened and was given the remit to monitor progress on the Government's response to the House of Lords inquiry, and to develop, in partnership with other stakeholders, a strategy for regenerative medicine in the NHS and provide an action plan. One of the major discussion areas for RMEG was that, even where therapies have real potential, this may not be known with a high level of certainty at the time an ATMP first comes to market as the available evidence base is often limited. In its report[3] the RMEG stated that,

"In order for NHS patients to benefit from regenerative medicines, robust and effective product evaluation has to be made to inform commissioning decisions. National Institute for Health and Care Excellence (NICE) guidance is essential in speeding up the adoption and spread of high value

regenerative medicines in healthcare. However, applying the Institute's appraisal methodology, based on cost utility analysis, to products whose true value may not be known for many years can be challenging, due to the inherent uncertainty of estimating long-term benefit from evidence derived from short-term studies."

The assessment of the cost-effectiveness of regenerative medicines may raise particular challenges compared to other types of technologies. Important challenges may include: the potential curative nature and claims of long-term/lifetime benefits; the potentially rapid changes that may arise in product characteristics over time; potential longer-term patient safety issues due to persistence; organisational and scaling issues, as well as the potentially significant up-front costs that may arise. Whether the conceptual differences between regenerative medicines and other types of technologies (e.g. pharmaceuticals and medical devices) require a different approach to the conduct and assessment of cost-effectiveness needs to be investigated.

The RMEG Evaluation and Commissioning Subgroup proposed that NICE commission a "mock technology appraisal" on an exemplar regenerative medicine product and develop an outline plan for such a study. This proposal was reflected in the final report and recommendations of RMEG,[3] which stated further,

"We encourage the Institute to consider the findings from these studies with a view to assessing whether changes to its methods and processes are needed. Evaluation and commissioning, as with all steps of the product development pathway, need to be supported by clear, up-to-date and accessible advice and guidance. "

Through RMEG subgroup discussions and further input from the Cell Therapy Catapult, it was concluded that undertaking a study involving a real commercial product was not feasible for a number of reasons: there would be significant commercial sensitivities; products undergoing regulatory review would be candidates for a real appraisal; and using a product at an earlier stage in clinical development is not helpful as the evidence base would be even less mature and, therefore, it would not have the attributes of an 'exemplar' product. It was therefore proposed to undertake the evaluation of a hypothetical product: CAR T-cell therapies (see section 4.1.1). This decision was made on the basis that CAR T-cell therapies are quite a new product class - none are currently licensed - for which there is emerging evidence of clinical benefit. An evaluation of these therapies might also appropriately exemplify some of the main challenges faced by new regenerative medicines. The Cell Therapy Catapult has knowledge and experience with gene modified T cells and therefore worked with others on the advisory group to develop the basis of the target product profile (TPP).

The objectives of this study were to:

- test the application of NICE appraisal methodology to regenerative medicines, identifying challenges and any areas where methods research and/or adaptation of methodology is appropriate
- identify specific issues related to the appraisal of regenerative medicines using the current NICE appraisal process and decision framework
- develop a framework for those developing regenerative medicines to facilitate understanding of how NICE evaluates clinical and cost effectiveness and to identify the most important evidence areas to develop before cost effectiveness can be reasonably estimated

3 Background

3.1 Issues identified by the EMA as being specific to advanced therapy medicinal products (ATMPs)

Most of the new, innovative regenerative medicines which are evaluated by the EMA are likely to be categorised as being ATMPs. The Committee for Advanced Therapies (CAT) is the EMA committee responsible for assessing the quality, safety and efficacy of ATMPs (and for following scientific developments in the field). The EMA and CAT have issued a range of documents providing guidance regarding the development of ATMPs; one of these (issued in 2008, based on the requirements of an EU regulation) is a guideline for post-authorisation follow up entitled *Guideline on safety and efficacy follow-up - risk management of advanced therapy medicinal products*. Such rules were needed due to the "novelty, complexity, and technical specificity" of ATMPs.[4]

In this guideline the concerns about *risks* relate to:

- Living donors (where applicable)
- Quality characteristics (e.g. origin and characteristics of cells, vectors; quality assurance issues)
- Storage and distribution of product (e.g. stability, preservation, thawing)
- Administration and re-administration procedures (e.g. immune reactions)
- Interaction of the product and the patient (e.g. immunogenicity, malignancy)
- Scaffolds, matrices and biomaterials (e.g. biodegradation)
- Product persistence (e.g. availability of rescue procedures or antidotes)
- Healthcare professionals, caregivers or other close contacts with the product

Concerns about the *efficacy* of ATMPs relate mainly to the uncertainty about how effective they may be in "real life" settings in the long-term. They include:

- Possible temporal changes in the characteristics of the living material in ATMPs may affect efficacy
- Time required for new tissue to be fully functional may be several years (use of surrogate endpoints needed for marketing authorisation, but confirmation with clinical endpoints needed in post-authorisation phase)
- Some ATMPs may be a once in a lifetime treatment: long-term follow up needed to demonstrate sustainability of efficacy
- Efficacy may be highly dependent on the quality of the administration procedure (e.g. patient conditioning, surgery). This may differ between clinical trial and normal healthcare settings

- Cell therapy products with a limited life-time may require an efficacy follow-up system that monitors the dynamics of efficacy (will help determine need and timing of reapplication)

The issues highlighted in the guideline for the *design* of the studies needed to monitor long-term safety and efficacy included careful consideration of:

- Sample size (high potential for drop-outs over many years of follow up)
- Dynamics of the disease and effects of the product (different approaches needed for detecting early versus late complications)
- The use of usual clinical practice for follow-up whenever possible to limit additional procedures and interventions
- Appropriate duration of follow up of living donors (where applicable)
- Feasibility of follow up of close contacts and offspring (where applicable)

Both the safety and efficacy follow up systems are defined as any systematic collection and collation of data that is designed in a way that enables learning about safety and/or efficacy of an ATMP. This may include passive or active surveillance, observational studies, or clinical trials. The guideline stresses that both the efficacy and the safety follow-up systems are not a substitute for the need for adequate data to be available at the time of authorisation to enable proper benefit-risk evaluation.[4]

3.2 Overview of wider regulatory evidence requirement issues and the evolving pathways for approval

The ethics, feasibility and reliability of small randomised controlled trials

Although the randomised controlled trial is the expected level of evidence needed for regulatory assessments, it is recognised that for some indications such expectations are unrealistic. Conducting RCTs in populations with severe or advanced disease may be problematic for a variety of reasons. Such populations may be very small and, consequently, recruitment into an adequately sized trial would require a large number of centres and take a very long time, and great expense. Also, when no alternative treatments exist, patients with life-threatening disease or severe morbidity typically need, and desire, accelerated access to innovative new therapies. Patients with more severe or advanced disease may be more willing to accept the risks of an experimental therapy. In such situations randomisation to a control treatment may be ethically problematic (due to absence of clinical equipoise).

A HTA review of ethical issues in the design and conduct of RCTs described numerous situations where alternative non-randomised designs are morally or practicably preferable. These included: where large differences between treatments are expected; when a disease, if left untreated, is lethal and for which there is no known effective treatment (i.e. unmet need); and when a disease is rare and

recruitment is slow.[5] Additionally, when trial populations are small, it may be difficult to differentiate a true treatment effect from a chance effect. Important chance imbalances in relevant prognostic factors between groups at baseline are more likely in small trials. The HTA review highlighted the problem of underpowered RCTs which were described as 'necessarily unethical' as they were unlikely to produce clear-cut answers. This argument was supported by 15 articles which stipulated the statistical necessity for random errors in measured effects of treatments to be small in comparison with the size of the therapeutic effect sought. Other articles in the review discussed the ethics of stopping RCTs early when there is some evidence of efficacy and the subsequent problems this may cause: reduced statistical precision, clinicians not being persuaded by results, and secondary trial aims being compromised being some of the key problems.

A related HTA review discussed further the ethical issues which may arise when early-phase (e.g. single-arm) trials produce very encouraging results: it *may* be unethical to conduct a further trial if the intervention is apparently effective in a small number of patients. In such a situation the argument for a trial rests on demonstrating a grey area between a reasonable hope that the intervention is effective in a few patients and a rational and justified belief that it is effective for the studied patient population more generally (i.e. the evidence to date has enough external validity).[6]

Possible alternatives to the randomised controlled trial

More recently, a framework for using unfamiliar trial designs when rare diseases are studied has outlined several possible alternative approaches.[7] The framework aims to facilitate research when populations are small. Two of the 'adaptive' designs outlined may be particularly relevant for regenerative medicines, where treatment intentions may be curative. The first is responsive-adaptive randomisation, which maximises allocation to the most effective treatment and minimises the required sample size. Outcomes for previous participants affect the subsequent treatment allocation probabilities. This 'play the winner' rule has the potential to reduce the number of patients who are allocated to less effective treatment and can therefore reduce the ethical concerns associated with randomisation. However, this design is limited to studies which assess rapidly available outcomes (as results from previous patients are needed to influence future allocations). Modified designs have also been outlined to counter the criticism that comparisons may be obtained in which only one patient has received conventional treatment. The second adaptive design which may also be useful for studying regenerative medicines is the internal pilot design; this design eliminates the loss of scarce eligible participants because of participation in a prior pilot study. Once the pilot phase is finished a sample size is recalculated with the study continuing until this number is recruited; patients from the pilot phase are included in the final analysis.

An EMA reflection paper on methodological issues associated with adaptive designs suggested that such designs 'would be best utilised as a tool for planning clinical trials in areas where it is necessary to cope with difficult experimental situations'. Cited examples of such situations included 'small populations or orphan diseases with constraints to the maximum amount of evidence that can be provided' and where there is 'ethical constraints to experimentation'.[8] However, the FDA raised two principal issues with adaptive design methods more broadly:[9]

- whether the adaptation process has led to design, analysis, or conduct flaws that have introduced bias that increases the chance of a false conclusion that the treatment is effective (a Type I error)
- whether the adaptation process has led to positive study results that are difficult to interpret irrespective of having control of Type I error

This draft FDA guidance document also noted that for some of the more recently developed adaptive methods (including adaptive randomisation methods), the magnitude of the risk of bias and the size of the potential bias, and how to eliminate these effects, are not yet well understood.

Although adaptive designs may be useful in some situations it is still likely that single-arm trials will form the basis of many submissions for the regulatory approval of regenerative medicines (due to the nature of the target populations). Nevertheless, a study which reviewed 31 oncology drugs or biologics approved by the FDA (between 1973 and 2006) *without* a randomised trial that incorporated a comparator treatment, supportive care, or placebo arm, concluded that such drugs have a reassuring record of long-term safety and efficacy despite the fact that nearly all the evidence studies were single-arm phase II trials. The median number of patients studied per approval was 79 (range 40 to 413); response rate was the primary endpoint for most drugs, and the median objective response rate was 33%. At the time of publication (2009) all but one of the drugs were still approved; marketing authorisation for gefitinib was rescinded after an RCT showed no survival improvement. Nineteen drugs have additional uses, with formal FDA approvals obtained for 11.[10]

Evolving regulatory pathways

Since the late 1980s and early 1990s, regulators and HTA bodies/payers around the world have produced new approaches to provide patients with timely access to new medicines.[11] These new regulatory pathways can also improve competitiveness; shortened product development times prior to licensing can be very beneficial and more appealing to emerging small and medium sized enterprises.

An overview of the relevant EMA regulatory accelerated access pathways is presented in Table 1. The main mechanism for accelerated access of these pathways is the reduced development time.

Table 1 EMA regulatory access pathways which allow accelerated access to treatments

Designation (year of introduction)	Use	Notes
Conditional marketing authorisation 2005	Seriously debilitating and life-threatening conditions, medicinal product for emergency use, or orphan medicinal products; must address unmet medical need	Authorised for one year with option to renew as long as benefit–risk profile remains positive. The condition is that the manufacturer will initiate or, preferably, continue studies in order to reduce uncertainty about benefits and risks to enable conversion to full authorisation. A Periodic Safety Update Report is required at 6-month intervals.
Approval under exceptional circumstances 1993	Medicines with urgent public health need for which comprehensive data cannot be provided	Justifications for not being able to provide comprehensive data include: rarity of the condition, lack of scientific knowledge (e.g. diagnostic tools), and contrary to medical ethics. Post-authorisation data collection required, which usually includes an identified program of studies. The results which form the basis of an annual reassessment of the benefit–risk profile
Accelerated assessment 2005	Medicines of major interest to the public health, particularly those representing a therapeutic innovation	Review time shortened to 150 days as compared with the standard 210 days. This pathway has very rarely been used.
Parallel scientific advice between EMA and FDA 2009	Important medicinal oncology, vaccine, orphan, paediatric, nanotechnology, advanced therapy, pharmacogenomics, or blood products. Product usually has fast-track designation in the US	Expected advantages are increased dialogue between the two agencies and sponsors from the beginning of the lifecycle of a new product, a deeper understanding of the bases of regulatory decisions, and the opportunity to optimise product development and avoid unnecessary testing replication or unnecessary diverse testing methodologies. Scheduling of parallel scientific advice can be challenging.
Adaptive licensing 2014 pilot	Medicines to treat an unmet medical need for a serious condition, especially where no alternative therapies exist.	Open to interventions in the early stages of development (during or prior to phase II). Multi-stakeholder participation desirable. Enhanced monitoring of drug safety and drug utilisation controls required after initial authorisation.

Adapted from Baird et al 2014[11]

The EMA's most recent development in this area is the adaptive licensing pilot programme which was launched in 2014. The programme utilises the regulatory processes within the existing EU legal framework and is defined as being a prospectively planned adaptive approach to bringing drugs to market. It is more of a staggered iterative system than previous approval pathways. Such a 'life-cycle approach' to acquiring and (re)assessing evidence will consider the basis of decision making in the following stages of a product's life cycle: development, licensing, reimbursement, monitoring/post-license evidence and drug utilisation.[11] Importantly, the approach encompasses both the authorised indication *and* the potential further therapeutic uses of the medicine. The EMA changed the name of the pilot project from 'adaptive licensing' to 'adaptive pathways' to better reflect the idea of a life-span approach.

The pilot project aims to examine whether this kind of approach to medicine development and authorisation will offer advantages in terms of achieving the best balance between the need for timely patient access with the importance of providing adequate, evolving information on benefits and risks. In so doing it is expected to develop thinking in the following areas:[12]

- To encourage developers of medicines to consider all regulatory tools and flexibilities within the existing EU legal framework when planning the life-cycle of the medicine development.
- To explore the extent to which regulatory demands for generation of evidence around efficacy and safety are compatible with demands around evidence generation from other stakeholders (e.g. HTA bodies, payers, patient organisations).
- To investigate in a timely manner the hurdles that exist in realising the most efficient medicine development pathways, including the role and limitations of real-world data.

Ideas for refining and improving this life-span approach are developing at pace. For example, with MAPPs (Medicines Adaptive Pathways to Patients) the development plan across target populations and indications will be agreed up-front with the EMA, which distinguishes the MAPPs process from the conventional indication expansion approach. The MAPPs plan may include a range of studies, such as RCTs, single-arm studies, pragmatic trials and other forms of real world study.[13] A newly formed public-private project called ADAPT SMART (Accelerated Development of Appropriate Patient Therapies: a Sustainable, Multi-stakeholder Approach from Research to Treatment-outcomes), which is funded by the EU Innovative Medicines Initiative, aims to facilitate and accelerate the availability of MAPPs.[14] NICE is one of the 32 international partners which together represent regulators, patients, academia, and industry. The challenge for ADAPT SMART is to develop a MAPPs model that aligns the needs of all stakeholders, including patients, member state payers, regulators, medical practitioners and industry. A major task will be the identification of opportunities and obstacles, and to provide a framework for MAPPs that will overcome the latter and seize the former. ADAPT SMART will address the challenges to the broad implementation of MAPPs by exploring new concepts to align the various stakeholders and create a consensus on what evidence will be required, how multiple sources of available data can be best used to facilitate MAPPs, and which scientific challenges related to MAPPs need to be addressed.[14]

In the UK there is another initiative which may facilitate the pathway to market: the MHRA operates an early access to medicines scheme (EAMS) which was launched in April 2014. This voluntary scheme (which does not replace the normal licensing procedures) is aimed at unlicensed or off-label treatments deemed by the MHRA to be 'promising innovative medicines' (PIMs) for treating life-threatening or seriously debilitating conditions for which there is unmet need. Once a PIM designation is obtained (stage 1 of the process), the MHRA can then provide benefit and risk information (stage 2 – scientific opinion) to doctors who may wish to prescribe the unlicensed medicine under their own responsibility. However, it appears somewhat unclear how the EAMS assessment output may impact on ongoing or forthcoming EMA assessment of the same therapy (e.g. in terms of speeding up processes, or reducing repetition of information). Further uncertainty around EAMS exists regarding

how therapies with this regulatory status can be funded. As EAMS is not accompanied by any funding arrangements, meeting the costs of the therapy is currently the responsibility of the manufacturer; this can act as a barrier to adoption, especially for high cost therapies produced by small enterprises.

Regenerative medicines in the new regulatory environment

The experience gleaned from the EMA adaptive licensing pilot so far appears to be quite limited with respect to regenerative medicines: as of May 2015 one of only three candidate ATMPs had been selected for a 'stage II' proposal.[15] By far the most accommodating regulatory environment for developing regenerative medicines is currently Japan, where, under the new 2014 legislation regenerative medicines can receive accelerated conditional approval after a single clinical study, provided the trial demonstrated the therapy to be safe, with evidence of a probable therapeutic benefit. This approach aims to dramatically accelerate patient access and meaningfully shorten clinical development times, thus promoting investment (since faster, less expensive development, coupled with accelerated commercialisation, would shift the risk-reward ratio favourably from an investment perspective).[16] However, there is concern that this approach may leave Japan with regenerative medicines which are unrecognised by other countries due to efficacy concerns: the lack of an explicit plan for determining efficacy during the conditional approval period points to a strong underlying assumption that regenerative medicines will ultimately prove efficacious; experience from other areas of clinical research suggest that such optimism may be misplaced.[17] The initial demonstration of safety based only on phase I trial data is an additional major concern.

The concern raised about the limited evidence which will likely be presented when a product is submitted for regulatory approval is by no means limited to the Japanese regenerative medicine experience. Since many regenerative medicines will be developed with the initial aim of treating small patient populations where there is unmet need, it is likely that they will be evaluated via a regulatory pathway which offers patients accelerated access to the new treatment. A consequence of this is that many of the studies submitted will be early phase, small single-arm trials. Nevertheless, the Japanese regulations excepted, the newer regulatory pathways being developed across the world do not focus specifically on facilitating the licensing of regenerative medicines. The newer pathways are primarily aimed at addressing unmet need in serious conditions where no alternatives exist, regardless of the *type* of technology. However, much of the focus and expectation for success in this area seems to have been directed at regenerative medicines, possibly because they may evolve over time and may therefore, ultimately, not be restricted and limited by having single modes of action. The submission of evidence which is based on single-arm studies appears to be less to do with regenerative medicines being a 'special case' category of interventions, but rather a consequence of the seriously ill, very

small populations with unmet medical needs which are often the initial target of new regenerative medicines.

4 Technology appraisal methodology issues which may be particularly relevant to regenerative medicines

4.1 Clinical efficacy and safety issues arising from EMA, NICE and FDA assessments of licensed regenerative medicines

4.1.1 Methods

From the regenerative medicine literature and experts in the field we sought to identify regenerative medicines which have been granted marketing authorisation in the EU. In addition to EMA assessment documents, we also sought any NICE or FDA documents. We extracted key details from these reports, with a primary focus on identifying issues which might be unique, or particular, to regenerative medicines.

4.1.2 Results

We identified six regenerative medicines which are (or have been) licensed in the EU: ChondroCelect, MACI, Glybera, Holoclar, Provenge and ReCell. No allogeneic therapies were identified - all were autologous. Summary details are presented in Table 2; more comprehensive details can be found in Appendix 1, Table 43.

ChondroCelect and MACI are both therapies for treating knee cartilage defects. ChondroCelect was the first ATMP to receive marketing authorisation, in 2009. The marketing authorisation for MACI was suspended in September 2014 as an authorised manufacturing site no longer existed. Holoclar is a therapy used for treating corneal lesions resulting from burns to the eye. In 2014 it became the first stem cell-based ATMP to gain regulatory approval (conditional marketing authorisation was granted). Provenge is an active cellular immunotherapy for asymptomatic or minimally symptomatic metastatic hormone-relapsed prostate cancer where chemotherapy is not yet indicated; this therapy purportedly helps the immune system to selectively attack cancer cells (rather than directly attacking tumour cells, as happens with CAR T-cell therapies). EMA marketing authorisation was granted in June 2013 but withdrawn in May 2015 at the request of the manufacturer, for commercial reasons. Glybera is used to treat familial lipoprotein lipase deficiency (a rare genetic disorder) with associated pancreatitis. Its mechanism of action is viral vector delivery of a therapeutic gene to muscle cells. In 2012 it became the first gene therapy to be approved in Europe or the USA. The ReCell spray on skin system is a regenerative medicine device; it harvests a small amount of a patient's skin cells which are then processed to produce a mixed cell population for immediate delivery onto burn wound surfaces. ReCell can be given rapidly as there is no need for proliferation of the harvested skin cells. A CE mark was granted in 2005 (under medical devices Directive 93/42/EEC).

Study designs

Randomised trial evidence formed the basis of the regulatory submissions for four of these six regenerative medicines. This would be expected since, for the four therapies in question (ChondroCelect, MACI, ReCell and Provenge), the disorders being treated were not rare and alternative therapies existed. However, both the EPAR and NICE ERG reports commented on the lack of blinding and the use of cross-over which allowed placebo patients to receive active treatment following disease progression, making interpretation of the post-progression overall survival results difficult. Nevertheless there were no design issues for the other three therapies (ChondroCelect, MACI and ReCell) demonstrating that ATMP/regenerative medicine status in itself may not necessarily be a barrier to submitting randomised trial evidence (as discussed at the end of Section 2).

Holoclar and Glybera were not studied in RCTs. B- both had orphan designations and had indications where there is unmet medical need; randomised trials were therefore not viable. A single group study design was therefore deemed acceptable in both EMA assessments. However, whereas for Holoclar the CAT accepted that the condition (eye burns) would not improve spontaneously (making it more plausible that observed benefits were due to treatment), for Glybera there were concerns that the reduction of pancreatitis events may possibly be due to temporal rarity and inherent variability of events over time (i.e. the resulting apparent benefit may have been due to chance). Perhaps it is for this reason that these two therapies took very different routes to approval. While conditional marketing authorisation was achieved for Holoclar without any prior negative CHMP decisions, Glybera had a much more difficult route to acquiring marketing authorisation. Negative CAT and CHMP opinions on Glybera were issued in June 2011. Following a request for re-examination, the CAT recommended the granting of marketing authorisation under exceptional circumstances in October 2011, but the CHMP did not recommend approval. Glybera was finally granted approval in July 2012 with a more restricted licence (the approval being for patients with lipoprotein lipase deficiency *and* severe or multiple pancreatitis attacks). It appears that EMA concerns about the efficacy of Glybera still remain, prompting Germany's G-BA (which makes reimbursement decisions) to suspend its assessment of Glybera.[18]

The issue this comparison of Glybera with Holoclar raises - the likelihood of cure or improvement without experimental treatment - could be an important consideration for both the design and interpretation of future regenerative medicine trials. It is for conditions where spontaneous cure or improvement is unlikely that so much is expected of regenerative medicines; the extent of the problems perceived to result from single-arm trial evidence may well depend on the 'game changing' possibilities of the therapy being assessed.

Table 2 Regenerative medicines granted EU marketing authorisation: summary data from EMA, FDA, and NICE reports

	Glybera	MACI	ChondroCelect	Holoclar	Provenge	ReCell
Year of EMA MA	2012	2013*	2009	2014	2013**	2005†
Type of RM	Gene therapy	Autologous cells seeded on porcine collagen membrane	Suspension of autologous cells	Autologous tissue-engineered product (includes stem cells)	Autologous active cellular immunotherapy	Stand-alone autologous cell harvesting device (for immediate delivery to wound surface)
Indication	Adults with familial lipoprotein lipase deficiency (LPLD - confirmed by genetic testing), detectable levels of LPL protein, and suffering from at least one pancreatitis episode despite dietary fat restriction	Skeletally mature patients for the repair of symptomatic cartilage defects of the knee	Repair of single symptomatic cartilaginous defects of the femoral condyle of the knee in adults	Corneal lesions, with associated (limbal) stem cell deficiency, due to ocular burns.	Asymptomatic or minimally symptomatic metastatic (non-visceral) hormone-relapsed prostate cancer in men for whom chemotherapy is not yet clinically indicated	Adults or children with: 1) partial thickness burns including scalds caused by hot water where mesh grafting is not required 2) large area burns; full thickness or deep partial thickness burns including where mesh grafting is required
Orphan status?	Yes	No	No	Yes	No	No
Claiming to meet unmet medical need?	Yes	No	No	Yes	No	No
Trial design	3 single-arm studies	One RCT (multi-centre)	One RCT (multi-centre)	3 retrospective case series (multi-centre)	3 RCTs (multi-centre)	3 RCTs (single-centre) & 8 observational studies
Trial size	Combined total n=27	n=144	N=118	Combined total n=148	Main RCT n=512	Main RCT n=82
Length of follow up	12-18 weeks	2 years	5 years	1 year	3 years	6 months
Comparator	Two observational studies (combined n=40) of patients receiving only diet reduction and no active treatment	RCT had a control arm of patients receiving microfracture	RCT had a control arm of patients receiving microfracture	Patients acted as their own controls – outcomes were compared with baseline data.	Placebo group of RCT: one third of the patient's cells were re-infused, but were not activated with the fusion protein	RCT had a control arm of patients receiving split thickness skin grafting
Adverse events	No obvious serious adverse events seemingly related to Glybera	Most were surgery-rather than product-related	Most were surgery-rather than product-related	Out of a total of 11 SAEs, three were judged as related to Holoclar.	Main risks were infusion reactions and (catheter related infections)	None reported
Surrogate outcome?	Yes - levels of fasting	Yes – MRI or histology scoring of structural and	Yes - structural repair	Yes - Corneal epithelial integrity and Absence of	Yes - Time to progression, antigen	No

	Glybera	MACI	ChondroCelect	Holoclar	Provenge	ReCell
	triglycerides	functional repair	(histology)	significant neovascularisation	response	
Real clinical outcome?	Yes - pancreatitis events	Yes - Knee Injury and Osteoarthritis Outcome Score (KOOS)	Yes - Knee Injury and Osteoarthritis Outcome Score (KOOS)	Yes - Visual acuity	Yes - Overall survival	Yes – several wound healing outcomes
Estimate of HRQoL	SF36 for earlier time points	'Lack of good quality of life data' (NICE ERG report)	'Lack of good quality of life data' (NICE ERG report)	Not assessed	Not assessed	Not reported

* subsequently suspended in 2014, ** withdrawn 2015, † granted under medical devices directive, MA marketing authorisation

Persistence and adverse events

The requirement for, and implications of, long-term persistence of the six licensed therapies in treated patients varied. For ChondroCelect, MACI, Holoclar and ReCell the aim is for therapeutic cells to become integrated in recipients for as long as possible and to ultimately produce new cells. Long-term data are needed for evaluations of true therapeutic success in this respect and adverse effects associated with longer-term persistence seem unlikely. Unknown long-term durability was highlighted in the ChondroCelect and MACI EPARs. Although the negative persistence effects of Glybera were thought to be minimal (the risk of cancer by integration of viral vector DNA was thought to be low) the EMA's conclusions on efficacy noted that the proposed single treatment was insufficient to provide a durable and measurable effect on triglycerides, suggesting the therapy did not persist in recipients for long enough. Little information could be found on the implications of the long-term persistence of Provenge within patients. However, prior to infusion into patients, Provenge is associated with a very short shelf life. An overview of the manufacturing and scale-up issues which may be encountered with regenerative medicines can be found in this report's discussion.

The only other adverse events which were noteworthy in terms of informing evaluations of future regenerative medicine studies were immune reactions. For patients receiving Glybera the use of immunosuppression did not result in a reduction of unwanted immunogenicity. Acute infusion reactions were identified as a risk in patients who had received Provenge and the risk of autoimmune reactions in non-prostatic tissues could not be ruled out.

Use of surrogate outcomes

Both surrogate and real clinical outcomes were evaluated for five of the six regenerative medicines; the ReCell studies did not need to use surrogates, with all outcomes having clear clinical importance. The use of surrogate outcomes was most problematic in the assessment of Provenge as the overall survival results were not supported by the progression-free survival nor the time to progression results. Many members of the CHMP felt strongly that - in light of these seemingly contradictory results – the efficacy evidence should be convincing and ideally corroborated by other secondary endpoints, which was not the case. The NICE ERG report also highlighted the lack of consistency between the surrogate outcomes and overall survival. Surrogate outcomes are discussed more broadly in Section 4.3.

Evolving therapies

A key difference between regenerative medicines and conventional medicines is the likelihood that specific treatments may change or evolve over time. The only example of this issue in the reports identified in this section relates to the cartilage cell (chondrocytes) treatments for cartilage defects of

the knee (MACI and ChondroCelect). When both were assessed by NICE they were third generation products. The ERG report noted the "general problem when long-term results are needed but the technology continues to evolve". The implication being that by the time long-term trials results become available, the therapy may well have been superseded by a (apparently superior) next-generation treatment. The extent of this evolution is described in Table 3.

Table 3 The evolution of autologous chondrocyte implantation (ACI) therapies

First generation	ACI-P. Liquid suspension of cultured chondrocyte cells placed in the defect covered with a cap made from periosteum.
Second generation	ACI-C. Liquid suspension of cells placed in the defect and covered with a collagen cap.
Third generation	The cultured cells are seeded on to a membrane or "scaffold" as in MACI (matrix associated chondrocyte implantation).
Characterised chondrocytes	Not all chondrocytes are equally good at producing cartilage. Some are more "chondrogenic" (cartilage-producing) than others. The most useful can be selected and are known as "characterised".
Fourth generation	Newer developments include the implantation not of cells that will form cartilage, but of tissue-engineered cartilage grown from autologous chondrocytes in collagen gel in the laboratory.

Reproduced from the NICE ERG assessment report "Autologous chondrocyte implantation for repairing symptomatic articular cartilage defects of the knee"

4.1.3 Summary

The key issues arising from the reports of licensed regenerative medicines, i.e. the issues which may be beneficial to consider when appraising future regenerative medicines, were:

- The importance of considering the likelihood of cure or improvement *without* experimental treatment when evaluating the results of single-arm studies
- The positive and negative implications of long-term persistence of therapies within patients
- The use of *reliable* surrogate outcomes (i.e. the need for validation of the relationship between surrogates and real clinical outcomes)
- The problems of long-term evaluations when therapies evolve over time
- None of the six regenerative medicines approved for use in the EU to date were allogeneic therapies

4.2 Study biases: an overview of their importance and methods to quantify and adjust for their impact

Regenerative medical technologies will often seek (and receive) EMA/FDA approval with limited or no data from randomised experiments. In such cases estimates of effectiveness will be based upon observational data and single arm experimental studies. Recent examples include Holoclar, which

received EMA authorisation based on retrospective case series (combined n=148), and Glybera which was licensed based on single-arm studies (combined n=27).

The focus of this section is therefore on making comparisons using historical controls and non-randomised evidence more generally, as this is likely to represent the typical way in which single arm studies will be used in any future regenerative medicine submissions where evidence from randomised trials is unavailable. This section will provide an overview of the reliability of using observational data and data from single-arm trials, and current methods used to minimise potential confounding bias. Most manufacturer submissions to NICE are likely to be based on efficacy evidence from randomised trials; this overview is therefore important as it may highlight areas where NICE might consider methods development research is needed to enhance the Technology Appraisal Programme. Specifically this section seeks to address the following three questions.

- To what extent do estimates of effectiveness obtained from non-randomised studies (NRS) agree with those obtained from randomised trials, i.e. the quantification of bias;
- What techniques are available to adjust for confounding bias in NRS and how reliable are they?
- What are the specific challenges of using single-arm studies to estimate treatment effectiveness.

4.2.1 Methods

Pragmatic surveys of the literature were carried out to address these research questions. One review addressed the reliability of obtaining treatment effectiveness estimates in comparative NRS. Two further separate reviews were carried out with respect to the second research question: one focusing on methods to adjust for bias in the evidence synthesis process; and a second on methods of analysing individual patient data from NRS. A final review explored the literature relating specifically to single –arm studies, the second For each review a number of key articles were identified using unstructured searches of Medline and studies known to the team. Based on these key studies snowballing techniques were then applied in which citation searches were carried out and references checked for relevant studies. Citations and references of any additional studies identified were then also checked until no further relevant studies were identified.

Records identified in both the searches of Medline and citation searches were screened by a single reviewer and full texts of those deemed potentially relevant obtained and also screened by a single reviewer.

4.2.2 Results

4.2.2.1 Quantification of bias in observational studies

A total of 14 studies were identified as relevant to the first research question (Quantification of bias in observational studies).[19-32]

All 14 studies relevant to the quantification of bias in observational studies sought to quantify the extent of bias in NRS by comparing the results of randomised controlled trials with NRS. In seven of the studies[22-24, 26, 27, 30]data were sourced from published meta-analyses that included both RCTs and NRSs. Algra et al.,[29] Benson et al.,[19] Dahabreh et al[21] Lonjon et al[31] and Sacks et al.[28] took a different approach and searched for NRSs that compared treatment effects and then carried out a further search to locate relevant RCTs. Beynon et al.,[20] took a similar sampling approach, randomly selecting RCTs from the Cochrane Central Register of Controlled Trials database, and then conducting searches for NRSs that had addressed the same topic.

The method of analysis in the majority of studies involved pooling the evidence from randomised and non-randomised sources separately. The resulting summary effects from the randomised and non-randomised evidence were then compared. Despite these similarities in approach a considerable range of methods were used to compare summary estimates of effect, with multiple outcome measures often being employed. Common outcomes included:

- Assessment of direction of effect;
- Subjective assessment of overlap of confidence intervals and proximity of summary estimates;
- Tests of statistical difference in summary estimates of effect obtained from randomised and non-randomised evidence;
- The calculation of ratios of odds or risk ratios.

The lack of a common method of comparison is problematic as it presents a significant barrier to making comparisons across studies and indicates a lack of consensus around how to measure the degree of concordance between results obtained from randomised and non-randomised studies. Furthermore, the employment of multiple methods of comparison in many studies can be considered a potential source of bias, as no attempt was made to adjust comparisons for multiple testing.

Of the 14 included studies, seven[19, 24, 26, 27, 29-31]concluded that there were no systematic differences in either the size or direction of effect estimates obtained from NRSs compared with those from RCTs. Five studies [20, 21, 23, 28, 32] concluded that effect estimates obtained from NRSs were systematically larger than those obtained from RCTs. This included the largest study by Ioannidis et al.,[23] that contained RCTs and NRSs from 45 topic areas. The authors of the other two studies [22, 25] felt unable to

draw any meaningful conclusions about the comparability of estimates obtained from RCTs and NRSs.

Study design and study quality were investigated in a number of the studies and was discussed in nearly all of the studies included in this review. Study design was identified as a likely factor in determining the reliability of estimates of clinical effectiveness obtained from non-randomised studies.

A number of the studies did not consider (i.e. they excluded) NRSs which used historical controls (Concato et al.,[26] Lonjon et al.[31]Concato[26] justified this exclusion based on previous evidence presented in Sacks et al[28]which reported that 79% of interventions tested were considered effective in trials with historical controls, while only 20% were considered effective in RCTs. Further empirical evidence of the potential for bias in studies using historical controls is also presented in Ioannidis et al.,[23]} Algra et al.,[29] and Golder et al.,[30]who all found that there were fewer discrepancies between the results of RCTs and NRSs when studies with historical controls were excluded. Ioannidis et al. also found that results from prospective NRSs contained fewer discrepancies with effect estimates from randomised studies than did retrospective studies, either with current or historical controls. Investigations into broader measures of quality have also revealed similar results. MacLehose et al.[24] classified NRSs as either high or low quality and observed that comparisons between randomised evidence and high-quality NRSs tended to show much smaller discrepancies than between randomised studies and low-quality NRSs.

4.2.2.2 Adjustment for bias in NRS

A total of 28 studies were identified as relevant to the first research question on the techniques available to adjust for confounding bias in NRS; details of those reviews are presented in Appendix 2).

A key factor in the reliability of estimates of effectiveness based on observational data is the statistical analysis used; a large number of studies have sought to develop and evaluate methods for adjusting and eliminating bias resulting from confounding. A summary of the studies which have looked at methods of adjustment for confounding bias in NRS and how reliable they are is presented in Appendix 3. Overall, it is unclear which methods are most appropriate in certain circumstances and further research is needed. Furthermore, adjusting for bias when comparing single arm trials to historical controls requires individual patient data (IPD); this can be difficult to access though approaches for recreating IPD data have been developed such as the algorithm by Guyot et al.[43] Consequently results generated from NRS will be subject to an unknown degree of uncertainty, even after adjustment for confounding.

4.2.2.3 Challenges of using single-arm trials to estimate effectiveness

A total of 10 articles were identified as relevant to the issue of using single-arm trials to estimate effectiveness. One of these was a a recent review paper, Paulus et al. 2014,[33] that discusses both the opportunities and challenges in using studies without a control group. Single-arm designs have the advantage of requiring fewer patients, all of whom receive the experimental treatment, thereby reducing the cost of trials in terms of patients, funding, and effort. This section discusses the issues of making comparisons using single arm studies and how comparable results from single arm studies and comparative randomised studies.

Making comparisons using single arm studies

Without a direct, concurrent comparator in single group studies, both explicit and implicit comparisons are frequently made Paulus et al. 2014.[33] Implicit comparisons are made when the expected outcomes in the absence of the intervention of interest are believed to be well known, and the expected effect size from the intervention is large. Explicit comparisons are made when the investigators compare the single group of subjects before and after an intervention, or when the investigators choose to incorporate a historical comparator in the analysis (e.g., historical data from the research institution or from an external cohort or existing database). Each of these alternative study designs has particular challenges and advantages. The particular challenges are discussed below and summarised in Table 4.

Table 4 Summary of challenges to interpretation of single group study designs

Study design	Challenge
Implicit comparison	Disease evolution may be variable and unpredictable
Before and after comparison	Changes in factors other than the intervention of interest across the periods compared are common Disease status fluctuates over time and natural recovery may occur Subject selection may be related to disease severity and prognosis
Comparison to historical control	Changes in factors other than the intervention of interest across the periods compared are common Information may be unavailable or not reported on the variability of effect estimates in historical control groups Inadequate reporting of data sources for historical response rates may occur

Adapted from Paulus et al. 2014[33]

Implicit comparisons

Implicit comparison is acceptable when the natural history of the disease is known with (near) certainty, the study participants are representative of the broader patient population in terms of disease severity and prognosis (in the absence of treatment), and the outcomes in untreated patients are well known, with a large observed effect in the study group.[33] Examples can be seen in the recent

technology appraisals of new drugs for hepatitis C by NICE, where because of the objective outcome and large treatment benefit, regulatory approval had been granted based on short-term single arm trials. However, even for diseases with an apparently uniform prognosis, there may be subtle yet clinically relevant differences between patients who are enrolled in the single–arm trial and those who do not qualify, and also between those in the trial and the historical control. Careful review of the study population and eligibility criteria is needed to make an assessment concerning external validity.[33]

When considering clinical effectiveness based on single-arm trials the comparison is often made implicitly: a survey found that roughly half of Phase II studies did not cite the source of their historical response rates.[34] This is never sufficient for the purposes of a cost-effectiveness analysis, where it is essential to have some reasonable estimate of the treatment's effectiveness ***relative to a control***. This requirement has the implication that such implicit comparisons are likely to be rarely of relevance to submission to NICE which will by necessity alwayscontain an economic component.

Before and after designs

Studies which use before and after designs (sometimes referred to as pre-post designs) assess the difference in response before and after the administration of an intervention in a single group of patients. Patients therefore serve as their own controls. For before and after designs to provide unbiased estimates of effectiveness it is necessary to eliminate all alternative explanations for observed treatment effects. It is therefore necessary to eliminate the possibility of improvement due to adjunctive therapies administered concurrently, or carryover effects from therapies administered before the intervention of interest should be considered. Furthermore, natural recovery presents another potential explanation for an observed before-after improvement in a health outcome in a single group comparison. Drawing valid and meaningful inferences about treatment effect using single group observational studies is therefore problematic when evaluating conditions that are fluctuating or intermittent and limits their applicability. Further to the above, before and after designs can be subject to the effects of regression to the mean which can simulate improvements in disease outcomes, but result from the elective sampling of patients at a peak severity in the natural history of disease that have a tendency to return to average severity levels over time regardless of interventions administered.[35] Before and after designs are therefore most appropriate for chronic conditions where disease status is stable over time or where natural history of the disease is certain such that any variation in disease status/progression is likely to due to the intervention. Before and after designs consequently are most commonly used for the evaluation of surgical interventions and other irreversible interventions. Before and after studies can also be useful when a disease is rare (as fewer patients need be recruited) or where there are ethical issues mean regarding using a control

group would be inappropriate such as in end of life care and childhood diseases. In these cases, however, the weaknesses highlight above are likely to remain, but can be mitigated by the inability to carry out comparative studies.

Historical controls

Comparative estimates of effectiveness can be generated from single arm studies by comparing results with data obtained from the study with historical data from the research institution or from an external cohort or existing database not drawn from the same institution or population. The interpretation of single group studies with historical controls is however complicated by specific challenges to the validity historical comparisons resulting from differences between patients selected as historical controls and those recruited to the single arm studies. Differences between the patient populations of a single treated group versus historical controls can arise for a variety of reasons including differences among accrual sites or over time in patient characteristics (e.g., age, performance status, or other prognostic factors). For example, more recently diagnosed patients may have milder manifestations of a condition due to improved (and therefore commonly increased) diagnostic sensitivity. Treatment effects may also be attributable to secular trends in clinical care (e.g., changes in diagnostic methods, classification criteria, or outcome ascertainment).

There are many additional reasons why patients in a single-arm Phase II study may not be comparable to those in some hypothetical historical group.[36] Phase II trials involving new agents are typically undertaken in large academic medical centres, where the patient population may vary in many ways from those in a subsequent phase III trial (e.g. they may be more mobile, more heavily pre-treated, have better socio-economic status, or better supportive care). For new agents there is a natural enthusiasm amongst the investigators for the new agent and a desire for it to "look good". This enthusiasm may manifest itself in various ways, such as setting the historical response rate at a low value[34]or only enrolling patients who look in some sense 'promising'. These aspects cause problems in an uncontrolled phase II study, but not in a randomized Phase II study.

On the other hand, if historical data are available from previous randomized phase III trials in such a case the historical estimate of the response rate for the standard treatment may be more accurate than the estimate obtained from the control arm of a randomized phase II trial, which is based on a smaller sample size.[36]

To address the problem of reliable historical benchmarks for single-arm phase II trials, efforts have been made to amass historical databases and derive historical control data for future trials in specific disease sites. Examples include stage IV melanoma advanced pancreatic cancer[37] and advanced non–small cell lung cancer.[38] The availability of this kind of data is extremely important for the better

evaluations and analysis of data from single-arm trials, and is essential to generate the estimates of relative effectiveness needed in economic models for the assessment of cost-effectiveness.

Comparability of results from single arm studies and randomised designs

There is a growing body of literature on whether Phase II trials should be single-arm or randomised (which is now the more common approach), with the focus on which design is most efficiently associated with success in Phase III randomised control trials; particularly in the context of cancer drug development. From one perspective this appears to not be directly relevant to the issue of the product development of regenerative medicines, where the issue is not which design best helps companies decide which drugs to take on to a Phase III trial, but rather how companies and regulators can manage development where the long established expectations for pivotal evidence are unlikely to be met. This body of literature, however, includes a number of studies that have sought to evaluate the reliability of estimates of effectiveness from single arm studies and their relative performance to randomised trials.

A simulation study[39] investigated the difference between randomised phase II trials and single-arm phase II trials under realistic parameters (alpha=beta=0.01, historical control success rate =20%, target success rate =40%). The study found that both designs produced similar results when there was no variation in historical control success rate but that even a modest variation in historical control success inflated the false-positive rate in single-arm trials. Furthermore, increasing the size of the single-arm trial inflated the false positive rate. Another simulation study[40] aimed to quantify the impact of a policy of all single-arm phase II trials versus randomised phase II trials on the number of phase III trials conducted using active agents. The parameters modelled in this study included between-institution variability in the standard-care response rate, in the treatment effect, and in the estimate of historical control rate; the presence of historical bias (over or under estimation of the response rate in the historical controls due changing care), and the proportion of phase II trials conducted using active agents. The study found that single-arm trials resulted in a higher percentage of Phase III when there was a minimal standard care (i.e. high unmet need), or when there was positive historical bias. Randomised phase II trials performed better in the presence of negative historical bias, high variability and were more consistent across variation in historical bias. These results reflect those of Tang.[39] Similar findings were reported when a Bayesian approach was used to compare single-arm and randomized studies, based on a binary response variable, in terms of their abilities of reaching the correct decision about the new treatment.[41] The study found that the accuracy of the estimate of the success rate for the standard agent, obtained from historical data, plays a crucial role: when the response rate for the standard agent is correctly estimated, the single-arm studies are

preferred. As the magnitude of the misspecification increases or as the total number of patients accrued get larger, two-arm studies tend to be preferred.

A very recent publication investigated the superiority of randomised phase II trials over single-arm phase II trials to predict success at Phase III for oncology drugs.[42] In this study, published phase III trials testing systemic cancer therapy were identified through a Medline search. Statistical analysis was performed using the Generalized Estimating Equation method correlating phase II features with phase III outcome. The results found that of 189 eligible phase III trials the primary outcome was positive in 79 (41.8%) (success) and these were supported by 336 phase II trials, including 66 randomised phase II trials; positive phase II outcome, randomised or not, correlated with positive phase III outcome ($p = 0.03$). Randomised phase II trials were not superior to single-arm phase II trials at predicting phase III study success. The authors concluded that given the added resources required to conduct randomised phase II trials, further research into phase II trial design is required.

In summary these studies confirm that results from single-arm trials can only be considered as reliable indicators of treatment benefit when the disease natural history is very well known, the patient population is homogenous, and the control (standard care) treatment has little impact on outcomes. It is interesting that increasing the size of single-arm trials is not helpful.

4.2.2.4 Effect estimates from single- versus multi-centre trials

Single-centre trials may produce significantly larger effect estimates than multi-centre trials. Although no publications were found examining this effect in non-randomised studies, there are relevant publications for RCTs. Over-estimation of treatment effect in single-centre RCTs has been discussed and quantified in critical care medicine;[44, 45] a relative overestimation of 36% was found in a study which compared 41 single-centre studies (median n=40) with 41 multi-centre studies (median n=223).[44] Trial- or review-specific examples of this effect have also been reported in neonatology.[46]

Possible reasons for the larger effects may be that single-centre studies:

- are more prone to bias than multi-centre studies[44]
- recruit fewer patients than multi-centre studies - smaller studies tend to report larger effects
- may have treatment effect magnitudes which are affected by the high levels of centre expertise
- may recruit populations which are unduly homogenous

These factors may limit the reliability or the external validity (generalisability) of single-arm trial results.

4.2.3 Relevance to future regenerative medicine submissions

While RCTs continue to be the dominant method for evaluating treatment effectiveness, a large number of studies have been conducted devoted to establishing the relatability of evidence from NRS. This sizable literature demonstrates both the value and challenges of using observational data. While the evidence is mixed regarding the reliability of observational data for evaluating treatment effectiveness, the existing studies do seem to indicate that in some cases at least confounding is a potential issue and will impact on treatment effectiveness estimates. Further, the current evidence suggests that retrospective studies and in particular historical control studies are more likely to result in biased estimates of effect. As observed in Section 4.1.2, many recent regenerative medicine submissions have been based upon data from single arm studies, which have been compared to historical controls. The findings of this review therefore suggest that a degree of caution is necessary in interpreting estimates from these comparisons, as bias in estimates of effectiveness from historical comparisons will add additional uncertainty not accounted for in the confidence/credible intervals presented. A key factor in the reliability of estimates of effectiveness based on observational data is the statistical analysis used and a large number of studies have similarly sought to develop and evaluate methods for adjusting and eliminating bias resulting from confounding. Despite this, it is unclear which methods are most appropriate in certain circumstances and further research is needed. Consequently results generated from NRS will be subject to an unknown degree of uncertainty, even after adjustment for confounding. Single-arm trials are only reliable indicators of treatment benefit when the natural history is very well known, the patient population homogenous and the control treatment has little impact on outcomes. It is interesting that increasing the size of single-arm trials is not always helpful.

If regenerative medicines continue to be targeted at tightly defined conditions, with a narrow population to minimise heterogeneity, where otherwise patients have little or no chance of recovery/improvement, the use of NRS and in particular single-arm studies may be adequate. To complement the data from such trials robust accurate evidence of the outcomes achieved with standard care must be provided. Where appropriate methods to adjust for confounding should be employed, the selection of the method so used must be explicit and based on sound reasoning. Confidence in estimates of effect may also increase by utilising multiple methods of adjustment, though care should be taken to ensure that methods are appropriate to the decision problem in hand. However, many regenerative medicines may require highly skilled and specialised facilities for optimum delivery. Consequently, the evidence on their efficacy and safety may be derived from only

small, single-centre studies which (more often than not) might over-estimate effect estimates or which might lack the external validity needed to support more widespread uptake of the intervention.

In terms of NICE methods and processes, methods research may be considered to inform guidance both for manufacturers (e.g. minimum reporting requirements on analysis methods for comparing single-arm trial data with historical control data) and for ERGs (e.g. checklists for appraising how historical control data were identified and analysed by manufacturers).

4.3 Review of the use of surrogate endpoints as primary outcome measures in definitive effectiveness trials of new therapeutic agents

4.3.1 Introduction

As discussed in Section 4.1 it can be anticipated that almost all of the pivotal trials of regenerative medicines submitted for assessment for marketing authorisation will utilise a surrogate or intermediate outcome (or endpoint). A surrogate may be either a laboratory or physiological measure of the patients' experience which could be used to predict or provide an early measure of therapeutic effect. This section presents an overview of surrogate outcome measures and their use in clinical research and highlights issues pertinent to the development and appraisal of regenerative medicines.

4.3.2 Methods

In order to describe the use of surrogate endpoints as the primary outcome measure in trials of new therapeutic agents a review of the most relevant and up to date literature was performed. The review was not systematic, but more designed as a pragmatic rapid review to assimilate current information and opinion on the use and suitability of surrogates in therapeutic trials. The review began with a search of key guidelines on the use of surrogate endpoints produced by the FDA, NICE DSU (University of Sheffield), EUnetHTA and survey results produced by the NIHR HTA Programme of HTA reports on the cost-effective use of surrogate outcomes. Citation and reference searches followed which produced a library of relevant peer reviewed publications and statistical reports on evidence for the use of surrogate endpoints in medicine. All relevant studies identified are presented in Appendix 4, Table 46.

4.3.3 Definition and examples of surrogate outcomes

Ideally, it is expected that the relative effectiveness of drugs and treatments be based on final clinical endpoints.[47] That is an outcome that the patient, the clinician and other stakeholders hope to avoid such as morbidity, impaired quality of life and/or death[48]. Randomised controlled trials with large sample sizes and extended follow up periods are often required to capture the statistical significance of a treatment or intervention's impact on a patient relevant outcome.[47] However, the requirements of RCTs are often impractical when considered alongside pressures of time for product to go to market

and in particular the urgent need for new treatments for patients with chronic but life-threatening diseases. The principal rationale for the use of a surrogate outcome is a more rapid assimilation of data without the need for large and lengthy trials in patients where mortality is high or treatment options are few.[49]

For example, overall survival (OS) is considered the gold standard to measure benefit in many clinical trials as OS provides a precise, statistically and clinically meaningful endpoint. However mature OS data is difficult to achieve due to the length of time needed and the number of deaths required for appropriate statistical analyses. Furthermore, overall survival as a measure of therapeutic success becomes less useful as the course and duration of diseases such as cancer move from being acute to more chronic; longitudinal effects of chronic disease such as comorbidities and additional ongoing treatments add further limitations to OS and an outcome.[50, 51] As a solution, there has recently been a steady move (by regulatory bodies) away from OS as a clinical endpoint measure and towards more short-term surrogate measures.

A generally accepted definition of a surrogate has followed that of Temple (1995):[52] "a biomarker or physical sign which can substitute for a clinically meaningful outcome measuring how a patient feels, functions and/or survives". However, chronic disease programmes and patient reported outcomes have meant that a broader definition is now needed to better fit the HTA perspective[53, 54]. Although the term 'intermediate endpoint' is sometimes used synonymously with surrogate endpoint,[55] it is often used to refer to more patient-relevant outcomes than those typically thought of as surrogates. However for the purposes of this report the term surrogate outcome will be used in its broadest sense.

Examples of approved drugs based on the use of validated surrogate endpoints include anti-hypertensives'and blood pressure in stroke research; cholesterol lowering agents and serum cholesterol and treatments for glaucoma and intra-ocular pressure;[56] CD4 count for AIDs or death in HIV infection[57]; and bone density for bone fracture in osteoporosis.[49] However, occasionally such approvals have to be revised when long-term data become available. The drug gefitinib was approved in the US in 2003 for patients with non-small cell lung cancer based on tumour response rate, a surrogate endpoint. When, in 2005, the results from later studies showed no significant benefit on survival, the FDA withdrew approval for its use in new patients. So, although surrogate endpoints offer the potential of real benefit - in providing patients with faster access to treatments, and saving trialists time and resources - they may also have important drawbacks. Most notably (as the gefitinib example demonstrates) there may be uncertainty about the relationship between surrogate and real clinical endpoints; this may result in treatment efficacies being over-estimated. A meta-epidemiological study which compared 84 trials which used surrogate outcomes with 101 trials which

used patient-relevant outcomes showed that trials reporting surrogate endpoints had larger treatment effects: on average, trials using surrogate outcomes reported treatment effects that were 28% to 48% higher than those of trials using final patient relevant outcomes, and this result was consistent across sensitivity and secondary analyses. The study characteristics of trials using surrogate outcomes and those using patient relevant outcomes were well balanced, except for median sample size (371 *v* 741) and single centre status (23% *v* 9%). Their risks of bias did not differ. This finding illustrates the importance of surrogate endpoints being appropriately validated, and of quantifying the association of treatment effect between the surrogate and patient relevant final outcomes and its uncertainty.[58]

4.3.4 Validation

Surrogate outcomes can be unreliable without sufficient validation, for example, two major antiarrhythmic drugs encanaide and flecanaide reduced arrhythmia but caused a more than threefold increase in overall mortality,[59] and cardiac inotropes improve short-term cardiac haemodynamic function but can increase mortality.[60] Such examples may fuel uncertainty about the validity of surrogates. The results of a questionnaire study of 74 stakeholders in the drug development of cardio-renal disease indicated that although the use of surrogates is not opposed, most are not considered valid. Out of the four surrogate outcomes suggested as an endpoint for trials - blood pressure, HbA1c, albuminuria or CRP - only blood pressure was considered moderately accurate. Questionnaire responders from industry valued the accuracy of surrogates consistently higher than academic and regulatory responders.[61]

4.3.4.1 General principles of validation

For a surrogate to be a reliable outcome measure it is generally accepted that the measure must be on the 'causal pathway' from the intervention to the clinical outcome.[49] The possible reasons for treatment or trial failure associated with surrogate endpoints have been discussed by Fleming[62] and more recently by Taylor:[49]

- The surrogate is not on the causal pathway of the disease process,
- Of several causal pathways of disease, the intervention affects only the pathway mediated by the surrogate.
- The surrogate is not in the pathway of the interventions effect or is insensitive to its effect.
- The intervention has mechanisms of action independent of the disease process (and so its effect will not be captured by a surrogate outcome)

There are a number of guidelines proposed for assessing the validity of surrogate endpoints[47, 49, 60, 62] and further work has also been published on scoring schemas for the value of surrogates.[63]

As a result of a review Elston and Taylor[48] recommend that before a surrogate outcome is accepted, a systematic review of the evidence for the validation of the surrogate/ final outcome relationship should be conducted. Furthermore, the evidence on surrogate validation should be presented according to an explicit hierarchy, such as:

- Level 1: evidence demonstrating treatment effects on the surrogate correspond to effects on the patient-related outcome (from clinical trials)
- Level 2: evidence demonstrating a consistent association between surrogate outcome and final patient-related outcome (from epidemiological/observational studies)
- Level 3: evidence of biological plausibility of relationship between surrogate and final patient-related outcome (from pathophysiologic studies and/or understanding of the disease process)

Methods for the statistical validation of surrogates as outcome measures have also developed.[59, 64, 65]

4.3.4.2 Validation of specific surrogate outcomes

Surrogate outcomes in oncology

A recently published systematic review of trial-level meta-analyses of randomised trials quantifying the association between surrogate and final outcomes in cancer, included 36 studies.[66] The review found that all validation studies used only a subset of available trials and that the evidence supporting the use of surrogate outcomes in cancer trials is limited. The results are summarised in Table 5.

Table 5 Summary of results from systematic review of trial-level meta-analyses of randomised trials quantifying the association between surrogate and final outcomes in cancer

Surrogate and clinical outcome	Number of studies	Range of correlation coefficients	Level of correlation (low, medium, or High)
Pathologic complete response for Event free survival.	2	0.17 – 0.28	Low
pCR for overall survival	2	0.30 - 0.49	Low
Response rate for overall survival	11	0.32 – 0.68	Low- medium
Locoregional control for overall survival	2	0.52 – 0.84	Medium - High
Event free survival for overall survival	3	0.79 – 0.86	High
Disease free survival for overall survival	7	0.62 – 0.98	High
Progression free survival for overall survival	30	0.29 – 0.99	Low - High
TTP for overall survival	3	0.54 – 0.69	Medium

The results of the review indicate that little research effort has been invested in validating tumour response as a surrogate for clinical outcomes; the available evidence suggests that better tumour level surrogate outcomes are required. The clinical outcome surrogates (intermediate outcomes) for overall

survival, particularly PFS, have been better studied and appear to perform better. However, the range of results for PFS indicates that the validation of a surrogate in one disease and setting cannot be assumed for other diseases and settings.

Progression-free survival (PFS) or time to progression (TTP)

The suitability of progression-free survival (PFS) or time to progression (TTP) as appropriate surrogate measures in advanced or metastatic cancer research has been reviewed.[54] The review identified 19 papers covering eight different tumour types. Data sets included the relationship between the measures within aggregated trial data, and the effect on individuals within individual patient data (IPD). The studies employed a variety of different datasets and statistical techniques, but the lack of standardisation across the studies made it very difficult for the review to identify any consistent relationship between the surrogate and overall outcome measure.

In a recent review of current statistical approaches to surrogate endpoint validation based on meta-analysis in various advanced-tumour settings, the suitability of PFS and time-to-progression was assessed using three validation frameworks: Elston and Taylor's framework, the German Institute of Quality and Efficiency in Health Care's (IQWiG) framework and the Biomarker-Surrogacy Evaluation Schema (BSES3).[67] Findings suggest that the strength of the association between the two surrogates and OS was generally low. The level of evidence (observation-level versus treatment-level) available varied considerably by cancer type and by evaluation tools, and was not always consistent even within one specific cancer type. This study emphasizes the challenges of surrogate-endpoint validation and the importance of building consensus on the development of evaluation frameworks.

A recently published study analysed the degree of difference in treatment effects between surrogate endpoints and OS in RCTs of pharmacologic therapies in advanced colorectal cancer.[68] Univariate and multivariate random-effects meta-analyses were used to estimate pooled summary treatment effects. The ratio of hazard ratios (HRs)/odds ratios (ORs) and difference in medians were used to quantify the degree of difference in treatment effects on the surrogate end points and OS. The study found a larger treatment effect for the surrogates than for OS. The authors suggested that previous surrogacy relationships observed between PFS and TTP vs. OS in selected settings may not apply across other classes or lines of therapy.[68]

Minimal residual disease (MRD)

MRD is a surrogate outcome that has been accepted by a regulatory agency, the FDA. With current intensive treatments, many acute leukaemia patients will enter a morphologic complete remission (CR). This is typically defined as patients having fewer than 5% blasts (abnormal, immature cells) in the bone marrow. If no further therapy is given after entering CR most patients will relapse,

demonstrating that microscopy-based evaluations are incapable of detecting all tumour cells. However, diagnostic techniques can now quantify and monitor minimal amounts of residual disease (MRD) - invisible to the trained eye - in patients in CR. The ability to quantitatively measure the amount of MRD at various times after achieving CR can guide subsequent treatment.[69]Studies have shown that MRD before stem cell transplantation is a strong independent predictor of subsequent relapse in children with high risk, or very high risk, ALL.[70, 71]

Threshold levels for MRD may vary depending on the population being considered. For children receiving first line chemotherapy for ALL, leukaemia cell concentrations of 0.01% (1 in 10,000) have been described as optimal for identifying higher-risk patients for potential intervention.[72] For children with ALL who have had a previous relapse the best MRD threshold for predicting disease free survival at 10 years has been reported as 0.001%.[73] The FDA have concluded that the evidence base to indicate that early MRD status is the strongest predictor of long-term EFS in ALL is unequivocal. They added that the magnitude of the importance of its critical role in risk stratification for treatment decisions has raised the consideration of its potential as a surrogate endpoint for clinical trials of investigational therapeutic interventions.[74] However, results from the UKALL R3 trial, which compared different chemotherapy treatments for children in first relapse, showed that the longer term outcome of having MRD negative status in patients who have already had one relapse may well vary according to *how* the status was achieved.[75] There is therefore some uncertainty in how MRD negativity correlates to long-term outcomes in relapsed populations.

4.3.5 Current issues for HTA and cost effectiveness models

Regulatory bodies find it acceptable for trials to be shorter and to have fewer participants and to use surrogate outcomes when populations are rare and there is a high unmet clinical need. However a commitment to ongoing research is mandatory to receive longer term approval; if research is not continued or if continued but efforts to validate the surrogate fail the approval will be withdrawn.[56] From a regulatory and HTA perspective the absence of data on clinical endpoints might be acceptable when a clinical endpoint is difficult or impossible to study. The EUnetHTA summarised their findings into eight recommendations for endpoints used in relative effectiveness assessment of pharmaceuticals.[47]

- Efficacy assessments of pharmaceuticals should be based whenever possible on final patient-relevant clinical endpoints (e.g. morbidity, overall mortality)
- Biomarkers and intermediate endpoints will be considered as surrogate endpoints if they can reliably substitute for a clinical endpoint and predict its clinical benefit

- Surrogate endpoints should be adequately validated and must have been demonstrated based on biological plausibility and empirical evidence
- Validation of a surrogate is normally undertaken in a specific population and for a specific drug intervention. Demonstration of surrogate validation both within and across drug classes should be thoroughly justified
- The availability of a sufficiently large safety database is particularly important and evidence on safety outcomes should always be reported
- The absence of data on clinical endpoints might be acceptable when a clinical endpoint is difficult or impossible to study (very rare or delayed) or target population is too small to obtain meaningful results on relevant clinical endpoints even after very long follow-up (very slowly progressive and/or rare diseases). However, these exceptions need to be carefully argued and agreed in advance
- Re-assessment requirements for further data should be clearly defined when an assessment has been previously made based on surrogate endpoints for the first assessment

Similarly, Elston and Taylor (2009)[48] recommend an HTA or CEM based on a surrogate outcome should only be undertaken where it is not possible to base the assessment of clinical effectiveness and cost-effectiveness on final patient-related outcomes (i.e., mortality, important clinical events, and health-related quality-of-life). In such cases a systematic review of the evidence for the validation of the surrogate/ final outcome relationship should be performed and the evidence on surrogate validation should be presented according to an explicit hierarchy.

Given the difficulty in validating surrogate outcomes, which conflicts with a need to use such outcomes in clinical research Ciani and Taylor[53] comment on the need to recognise the need for pragmatic high level evidence, preferably from meta-analyses and regression modelling using both surrogate and final outcomes for HTA. The potential of this is demonstrated by a study conducted to illustrate the potential to reduce uncertainty around the clinical outcome by estimating it from a multivariate meta-analysis.[76] Bayesian multivariate meta-analysis was used to synthesise data on correlated outcomes in rheumatoid arthritis. Estimates of Health Assessment Questionnaire (HAQ) were mapped onto the health-related quality-of-life measure EuroQol five-dimensional (EQ-5D) questionnaire, and the effect was compared with mapping HAQ obtained from the univariate approach. The results showed that use of multivariate meta-analysis can lead to reduced uncertainty around the effectiveness parameter. By allowing all the relevant data to be incorporated in estimating clinical effectiveness outcomes - including data from surrogate outcomes - multivariate meta-analysis can improve the estimation of health utilities through mapping methods.

In their review of HTA and CEM Taylor et al 2009[49] found that only one of the 4 reports undertook a systematic review to specifically seek the evidence base for the association between surrogate and final outcomes. Furthermore, this was the only report to provide level 1 surrogate–final outcome validation evidence (i.e. RCT data) showing a strong association between the change in surrogate outcome (BPAR) and the change in final outcome (graft survival) at an individual patient level. The outcome of the review was to make recommendations for the evaluation of surrogate endpoints in HTA (these are listed in Appendix 4, Table 46).

Elston and Taylor's HTA publication has been key to providing insight into the use of surrogates within the HTA and cost effectiveness models framework and presents the range of approaches, including hazard ratio calculation, transition probabilities within a model of natural history CEM and predictive risk equations, used by researchers to quantify the relationship between surrogate and clinical endpoints.[48]

In addition to calls for validation of commonly used surrogate outcomes there is a need for novel, more appropriate, more valid outcomes. An editorial in the Journal of Clinical Oncology[77] commented on the large number of novel antitumor agents currently being tested in ever smaller groups of patients with increasingly specific tumour characteristics. Cancer types will continue to be divided up into many sub-entities that differ from each other in terms of genetic makeup, natural course, and sensitivity to systemic treatments. This, together with the limited number of patients who are available for clinical studies, means that a new approach to oncology research is needed. The editorial called for more intensive efforts at the pre-clinical stage to better understand the mode of action of potential new agents, and for this information to be used to select more precisely the target population and appropriate and valid surrogate outcomes. By so doing it should be possible to achieve a higher success rate in phase III studies, with smaller numbers of patients needed.

4.3.6 Summary

- Studies looking at surrogates for overall survival demonstrate how difficult it is to validate even commonly used surrogates
- On average it seems that trials using surrogate outcomes report larger treatment effects (28% to 48%) than those of trials using final patient-relevant outcomes
- However, a desire to get regenerative medicines to market quickly means that manufacturer submissions are likely to be supported by short-term trials reporting primary outcomes which are surrogates
- Regulatory agencies may accept evidence based on surrogate outcomes, for example the FDA accepts that MRD is the strongest predictor of long term event-free survival in acute

lymphoblastic leukaemia, although there is considerable uncertainty about its value in relapsed populations

- The choice of surrogate outcomes must be researched, explicit and justified. Ideally a systematic review of the evidence for the validation of the surrogate/ final outcome relationship should be performed and the evidence on surrogate validation should be presented according to an explicit hierarchy
- Analyses, at whatever stage of development and maturity of data, should include all available outcome data in order to minimise uncertainty.

4.4 Scoping exercise of potential cost-effectiveness issues

The assessment of the cost-effectiveness of regenerative medicines and cell therapies may raise additional challenges compared to other types of technologies. A focused scoping review was undertaken to help to identify potential conceptual differences between regenerative medicines/cell therapy products and other more conventional technologies. The objective of the scoping review was to identify possible characteristics which could make any assessment of cost-effectiveness, uncertainty and the value of further evidence different from other technologies. These characteristics also provided a basis for subsequent exploratory work to assess the appropriateness of existing decision frameworks for health technologies. A related objective was to identify areas where additional methodological development may be required.

The scoping review was based on completed and ongoing NICE Technology Appraisals (TAs) for regenerative medicines and broader literature which has attempted to identify potential challenges.

4.4.1 Previous regenerative medicine evaluations evaluated within the NICE TA process

Methods

A review of previous NICE TAs of regenerative medicines and cell therapy products was conducted. The primary aim of the review was to identify any common themes and potential analytic challenges relating to the assessment of their cost-effectiveness.

Results

NICE has previously evaluated two regenerative medicines within the existing TA process: autologous chondrocyte implantation (ACI) for the treatment of cartilage defects in the knee joints[78-80] and sipuleucel-T for treating asymptomatic or minimally symptomatic metastatic hormone-relapsed prostate cancer.[81]

Autologous chondrocyte implantation (ACI)

ACI has now been appraised on 3 separate occasions by NICE: originally in 2000 (TA16),[79] and as separate re-reviews in 2005 (TA89)[80] and in 2015 as part of an ongoing review.[78] The original guidance from TA16 has since been replaced by TA89 and documentation from the initial appraisal has been removed from NICE's website. Hence, our review focused on the separate re-reviews. However, it was reported in the Final Appraisal Determination (FAD) for TA89 that when the original guidance was produced in 2000, data from completed RCTs for ACI were not available.

For the re-review in 2005,[80] 4 controlled trials were subsequently considered; 2 comparing ACI with microplasty (n=40 and n=100) and 2 comparing ACI with microfracture (n=80 and n=66). Follow-up across the trials varied between 1 to 2 years. Three publications relating to a Swedish longer-term case series for ACI were also identified, describing outcomes for up to 11 years after surgery.

In reviewing the various documents for TA89 there appears to be no specific reference made to any distinct challenges of evaluating ACI based on its classification as a regenerative medicine or any specific discussion related to possible innovation. However, the lack of medium to long-term outcomes associated with ACI and their durability was highlighted as a key limitation in the FAD. The committee also noted concerns that the comparative trial evidence had a follow up of only 1 to 2 years and longer-term case series data appeared to show similar benefits for most treatment modalities.

While uncertainties surrounding long-term outcomes are clearly not unique to regenerative medicines, the Assessment Group (AG) concluded in TA89 that there was insufficient evidence for ACI to produce a robust cost per QALY estimate for ACI. Instead the AG undertook "*illustrative modelling*" of the cost effectiveness of ACI in three "*increasingly speculative*" stages, incorporating alternative assumptions relating to the short term (2-years), medium term (10-years) and long term (up to 50-years). The conventional NICE reference case for cost-effectiveness was applied, although deterministic approaches (i.e. point estimates were assumed for input parameters) were applied. A discount rate of 1.5% for health benefits and 6% for costs were applied in line with the recommended rates at the time of the appraisal.[82]

The cost-effectiveness results were sensitive to the time horizon and the assumptions employed within these. In both the short-term and medium-term analyses, ACI was reported to be dominated by the current standard of care (microfracture/mosaicplasty). In the long-term analyses, the possible avoidance of knee replacements were taken into account and the ICER of ACI versus microfracture was reported to be between £3,200-£3,650 per QALY.

ACI was subsequently not recommended for routine use in TA189, being given an only in research (OIR) recommendation. Hence the use of ACI in the NHS was restricted to use within "*studies that are designed to produce good-quality information about the results of the procedure. These results should include measuring any improvement in patients' quality of life, and the benefits and risks of ACI over a long period of time*".

While the re-review[78] is still ongoing, the initial appraisal consultation document (ACD) was issued in March 2015 and is now in the process of consultation. The rapid evolvement of ACI over time was highlighted in the ACD and the branded MACI product being appraised was now classified as a third-generation ACI.

The AG undertook a 'review of reviews' comparing the effectiveness of ACI (any generation) with microfracture. 12 systematic reviews were identified. Studies within the review were reported by the AG to be heterogenous with follow-up between 6.5 months to 7.5 years. The AG considered the results of the reviews to be inconclusive on the effectiveness of ACI compared with microfracture

NICE received separate submissions for ChondroCelect (SOBi), MACI (Aastrom), and OsCell (Robert Jones and Agnes Hunt [RJAH] Orthopaedic Hospital). While both ChondroCelect and MACI had EMA approval, the submission by OsCell was based on a product provided via the Hospital Exemption License that allows provision for OsCell to supply chondrocytes for use in ACI under the professional responsibility of a medical practitioner. There was a marked difference in list prices (excluding VAT) between the products: £18,301 for ChondroCelect; £16,226 for MACI and £4,135 for OsCell. However, costs were also noted to vary in different setting due to negotiated procurement discounts.

The manufacturer submission supporting ChondroCelect provided evidence of clinical effectiveness from 4 new sources not considered in TA89: a randomised controlled trial (n=118) with up to 5-years follow-up; a 'compassionate use' case series (n=370); an ongoing registry based cohort study (n=308) with up to 3-years follow-up; and data from a Belgian reimbursement scheme (254 procedures undertaken over a 3-year period).

The submission supporting MACI described new clinical evidence from 2 RCTs (n=144 and n=60, both with up to 2-years follow-up) and a subsequent ongoing extension study (up to 3-years additional follow-up; with interim data for 1st year reported).

The OsCell submission reported interim (up to 5-years follow-up) clinical effectiveness evidence from a UK RCT (ACTIVE trial: n=390 including first, second and third generation ACI) and a separate cohort study (n=366) with up to 3-years follow-up.

Separate cost-effectiveness analyses were presented by the manufacturers of ChondroCelect, OsCell and the AG. A discount rate of 3.5% for both health benefits and costs were applied in line with the recommended rates at the time of the appraisal.[83] The base-case ICERs for ACI compared with microfracture in the companies and Assessment Group models were approximately £6,000-£7,000 and £16,000 per QALY gained respectively.

In the ACD, the Committee concluded that while there was more clinical-effectiveness data than at the time of the previous NICE technology appraisal guidance, the evidence base for the technology is still emerging and no comparative clinical effectiveness data had been reported beyond 5 years. Innovation was formally considered and the Committee agreed that ACI, albeit not new, is technically innovative. However, they concluded that:

"*in the context of a technology appraisal, innovation needs to be judged by the benefit for patients, and that with the current uncertainties in the clinical effectiveness, it was not possible to conclude that these technologies can be considered innovative*" (ACD: Section 5.2.1).[78]

In relation to the cost-effectiveness evidence, the Committee considered that OsCell had underestimated its cell costs, and that the true cost may approach that of MACI and ChondroCelect. The Committee concluded that, although the cost to the NHS of providing the cells for ACI was somewhat uncertain, the cost estimate used within the AG and the Chondrocelect model were reasonable for the purposes of decision making.

The Committee concluded that a lifetime horizon was preferable because it captured all of the costs and consequences of treatment, but the lack of long-term data with which to populate a model generated large uncertainties. The Committee concluded that there was no ICER available that included the assumptions they considered plausible. Further they were not persuaded that ACI was proven to be a cost-effective treatment; neither did they consider that the available data robustly supported that ACI was better than other treatments.

The Committee therefore issued a provisional OIR recommendation because the clinical effectiveness and cost-effectiveness of ACI continues to remain uncertain. Hence, ACI was not recommended for routine use in the NHS unless it is part of existing or new clinical studies. It was stated that *"these*

studies should generate robust outcome data and include both interventional and observational studies".

Sipuleucel-T (Provenge®)

NICE issued guidance for sipuleucel-T (Provenge®) in February 2015.[81] The appraisal was subsequently withdrawn in May 2015 following the withdrawal of the marketing authorisation for sipuleucel-T. However, prior to this NICE had conducted a full appraisal of the technology, rejecting it due to the cost-effectiveness estimates exceeding the threshold considered to represent value for money to the NHS.[83]

Clinical effectiveness evidence was based on 3 phase III double-blind multicentre RCTs conducted in USA and Canada that compared sipuleucel-T with placebo (n=512, n=127 and n=98). The primary endpoint for the pivotal (IMPACT) trial was overall survival with a median follow-up time of 34 months. The main secondary endpoint was time to disease progression. The risk of death was reported to be statistically significantly lower in the sipuleucel-T group than in the placebo group (hazard ratio [HR] 0.78, 95% CI 0.61-0.98). The trial also demonstrated that patients randomised to sipuleucel-T survived for longer (median 25.8 months) than patients randomised to placebo (median 21.7 months), with a difference of 4.1 months.

Subgroup analysis suggested important clinical differences based on baseline PSA concentration, with a difference of 13 months (HR 0.51, 95% CI 0.31 to 0.85) in median survival for the quartile of patients with the lowest baseline PSA concentration compared to 2.8 months (HR 0.84, 95% CI 0.55 to 1.29) for the quartile with the highest PSA concentration. The company suggested that sipuleucel-T has a delayed onset of action because it is an immunotherapy, so giving it early in the course of disease progression (as indicated by a low PSA) could provide patients with more time to benefit from sipuleucel-T. However, the ERG cautioned that the subgroup of patients in the IMPACT trial with lowest quartile baseline PSA had been identified in a post-hoc analysis, with no clinical significance attached to the specific PSA concentration in this group.

A conventional Markov (partitioned-survival) model was submitted by the manufacturer to inform cost-effectiveness with a lifetime time horizon (10-years). Parametric survival analyses were used to extrapolate the trial data to the lifetime horizon. Conventional rates of discount (3.5% costs and benefits) were applied.

In the company's base-case analysis, the ICER for sipuleucel-T compared with BSC was £124,875 per QALY gained. In the subgroup with the lowest quartile of baseline PSA concentration, the ICER for sipuleucel-T compared with BSC was £48,672 per QALY gained. The company also conducted

sensitivity analyses with an alternative comparator (abiraterone rather than BSC) and applied assumed discounts to the price of abiraterone of 30% or more; these analyses resulted in ICERs for sipuleucel-T compared with abiraterone of at least £511,663 per QALY gained.

The ERG noted uncertainty surrounding the extrapolation of survival data and chose an alternative survival distribution for overall survival in its exploratory analyses, alongside other proposed amendments. In the ERG's base case, the ICER for sipuleucel-T compared with BSC was £111,417 per QALY gained. The ERG's analysis for the low-PSA subgroup resulted in an ICER of £61,381 per QALY gained for sipuleucel-T compared with BSC.

In considering the cost data and assumptions within the manufacturer's submission, it noted that the acquisition cost of sipuleucel-T included the costs of leukapheresis, patient tests associated with leukapheresis, transportation of white blood cells, and manufacture and transportation of sipuleucel-T. However, given the complex administration of sipuleucel-T and the lack of experience in the UK of using the treatment, the Committee was unsure whether the NHS would incur additional costs of using sipuleucel-T that were not included in the economic model. The Committee also considered that there may be patient travel costs associated with sipuleucel-T treatment, due to its provision within specialist centres, which had not been included in the model. These issues were considered to add uncertainty to the estimates of cost-effectiveness.

In considering the clinical relevance of the subgroups, the Committee heard that the clinical experts were unable to identify a single PSA value that was currently used for guiding treatment decisions. The Committee considered that registry data could have been used to assess whether outcomes after treatment with sipuleucel-T in clinical practice were similar to those in the IMPACT trial for patients with low baseline PSA concentration, but that they were not presented with this information by the manufacturer. The company reported that such a registry had been established (PROCEED) but that data were considered too immature to inform overall survival.

In relation to potential innovation, the Committee reported that it was aware that sipuleucel-T is an autologous cellular immunotherapy and is the first treatment for this indication that is not cytotoxic or based on hormone therapy. However, they concluded no evidence had been presented for "*demonstrable and distinctive benefits that had not been captured in the reference-case measure of QALYs*".

The Committee concluded that there were areas of considerable uncertainty in the results generated by the model, and noted that all the ICERs estimated by the company and the ERG fell substantially above the range normally considered cost effective (£20,000 to £30,000 per QALY gained).[81]

Issues and common themes

The existing NICE TAs raise a number of potential issues and several common themes emerge. The innovative nature of ACI (most recent ACD only) and sipuleucel-T were acknowledged by both committees. However, these considerations appear to relate more to an appreciation of the technical nature of the innovation as opposed to any specific attributes of the innovation which might lead to a distinct benefit that may not be appropriately reflected in the reference case measure of QALYs. Importantly, no evidence was presented in either appraisal which led the committee to consider that these specific attributes of innovation were relevant.

The high levels of uncertainty surrounding cost-effectiveness results were highlighted in both appraisals. In the most recent appraisal of ACI, this led the committee to conclude that there was no ICER available that included the assumptions they considered plausible, neither were they persuaded that ACI was 'proven' to be a cost-effective treatment. The committee appraising sipuleucel-T concluded that despite the considerable uncertainty in the results generated by the model, all the ICERs estimated by the company and the ERG appeared to be substantially above the range normally considered cost effective. The difference in the committee's subsequent recommendations (i.e. OIR for ACI and reject for sipuleucel-T) suggests that the committees may have reached different conclusions on the potential for both products to be cost-effective despite the inherent uncertainties.

The rapidly changing nature of regenerative medicines and challenges raised by this is evident across the series of appraisals for ACI. Over the 15 year period that the separate appraisals have been undertaken, the initial first-generation ACI products (ACI-C) have been superseded by second and third-generation products. This has raised potential challenges in relation to quantifying the long-term uncertainties as newer generations emerge. That is, during the time over which longer-term evidence has emerged newer generations of ACI have also arrived. The generalisability of the longer term evidence to the newer generations has raised additional issues and challenges. For example, the AG in the most recent ACI appraisal excluded longer term evidence available from the first generation of ACI on the basis that these products had now been superseded by newer generations. This approach effectively assumes that existing evidence cannot be generalised across different generations of products. If such a position were routinely taken, this may pose a potential challenge to manufacturers in terms of providing data which is considered sufficiently robust within a time frame that permits sufficient commercial return to warrant their R&D expenditure. The extent to which evidence can be generalised or transferred between generations remains an important consideration.

Similar uncertainties arise for more conventional technologies in relation to the constant evolution of knowledge over time and subsequent challenges for HTA and cost-effectiveness assessments. The

challenge of determining when evidence is sufficiently 'robust' within a technology's overall life cycle to undertake an HTA/cost-effectiveness assessment are summarised by what has been termed 'Buxton's law' (i.e. "*it is always too early until, unfortunately, it's suddenly too late*").[84] These challenges have led to an increased appreciation of the importance of employing an iterative approach to cost-effectiveness assessments, such that as new evidence emerges, progressively more certain estimates are derived and earlier policy decisions can be subsequently reconsidered.[85, 86] However, as highlighted by the ACI appraisal, more specific challenges may arise for appraising newer generations of products in relation to the extent to which evidence is considered generalisable or transferable across different product generations. Furthermore, the potential high up-front costs and the scale of any irrecoverable costs, as discussed in more detail in later sections, may be important additional considerations within these iterative assessments.

Additional uncertainties were also identified across both appraisals in relation to the costs that would be incurred by the NHS. Within the ACI appraisal, uncertainties were identified surrounding the acquisition costs of the technologies themselves (i.e. due to local price negotiations and concerns regarding the proposed cost of the product provided under hospital exemption), as well as to the appropriate cost or tariff to apply to other elements of the overall procedure. The complexity of provision and lack of experience in the UK of using the product was also identified as an issue within the appraisal of sipuleucel-T. Uncertainties arising from this, alongside the proposed provision within specialist centres and possible impact on travel costs for patients, led the committee to conclude that additional uncertainties existed surrounding whether all relevant costs had been appropriately included within the model.

Importantly, the RCTs which informed the basis of regulatory submissions for ChondroCelect, MACI and sipuleucel-T were also central to the subsequent submissions to NICE and the economic models developed to support these. Follow-up ranged from between 2 to 5 years for ChondroCelect and MACI and additional evidence was also submitted from ongoing extension studies and other registries. In the case of sipuleucel-T, the pivotal (IMPACT) trial was powered on overall survival with a median follow-up time of 34 months. Consequently, neither appraisal provides an indication of any additional challenges which may be raised for regenerative medicines or cell-therapies which have received regulatory approval based on uncontrolled studies or employing surrogate outcomes. However, it seems reasonable to conclude that the uncertainties expressed in relation to cost-effectiveness within the existing appraisals are likely to be magnified in this eventuality.

4.4.2 Broader consideration of potential conceptual differences and possible methodological challenges for cost-effectiveness analyses

A separate review of known references and key citations of these was undertaken to identify other potential conceptual differences between regenerative medicine and cell-therapies and more conventional medicines to identify potential methodological challenges for cost-effectiveness assessments.

During the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) conference held in November 2014, a workshop was held to discuss potential HTA and reimbursement challenges for regenerative medicines and curative treatments. As part of the workshop, Towse[87] argued that many of the challenges for curative therapies appear similar to those for disease modifying therapies for chronic diseases, sharing several associated problems; most notably: (i) short term trials using surrogate outcomes that may not produce relevant clinical outcomes; (ii) outcomes that may not be sustained over time and (iii) safety problems which may emerge over time.

In considering how these uncertainties might be formally incorporated within policy decisions regarding reimbursement, Towse[87] acknowledged that value of information (VOI) approaches provide a potential analytic framework. This framework formally evaluates the potential trade-off between the net health benefits to current patients from early access to the technology and those to future patients from withholding access to the technology until additional research has been conducted. The framework can then be used to help guide more appropriate policy choices between: (i) adopt and reimburse now, (ii) delay adoption/reimbursement and undertake further research (i.e. only in research), (iii) adopt/reimburse now and undertake further research (akin to coverage with evidence development [CED] or approval with research [AWR]).[88] Towse also acknowledged the importance of risk sharing approaches and particularly how these could enhance the value of CED/AWR.

As discussed in Section 4.1, assessments of the clinical effectiveness and cost-effectiveness of regenerative medicines and cell-based therapies, particularly early in their life-cycle, may be less extensive and lower in quantity than evidence for more conventional pharmaceuticals. Under these circumstances it may become even more critical to consider conditional reimbursement and possible risk sharing agreements between the manufacturer and the payer.

Towse[87] concluded that the main reimbursement challenges for regenerative medicines relate more to their financing as opposed to the methodology of HTA and cost-effectiveness. In particular, concern was raised by Towse regarding whether health care systems could cope with the potential high upfront costs of a curative treatment that appeared cost-effective using conventional thresholds; thereby presenting a potential barrier to adoption.

In the same workshop, Faulkner[89] explored differences between regenerative medicines/cell therapies and conventional biologics. Specific reference was made to the more limited understanding by physicians and payers, leading to potentially greater requirements being made for longer-term data collection to more robustly demonstrate value. The multiple procedural steps required for some therapies were identified as another potential challenge, as these may be subjected to different regulatory and reimbursement process. Similar concerns were raised by another speaker, Husereau[90], highlighting that regenerative medicines can be considered as both a biologic and a device/procedure, with many countries also having different reimbursement procedures often based on cost minimisation for the funding of procedures due to fixed HRG pricing. In common with the arguments made by Towse, both Husereau and Faulkner also highlighted the challenge of applying a single financing approach for reimbursement.

The potential to enable a disease cure or prolonged therapeutic effect was also identified as relevant characteristic by both Faulker and Husereau. However, Husereau identified a specific challenge surrounding the classification of a cure and its distinction from a prolonged therapeutic effect. To be considered curative, a therapy needed to demonstrate '*no chance of re-entering suboptimal health state from same disease'*. In reality it seems unlikely that a new therapy can be definitively classified as curative prospectively since many of the required elements cannot be demonstrated until a full lifetime of a cohort of treated patient has passed.

Husereau also raised the question regarding whether there is anything specific regarding curative therapies relative to standard treatments that could be perceived as providing additional benefits to patients beyond the current QALY framework. This was further considered in a subsequent publication.[91] Husereau reported that while there was limited direct empirical evidence to address this specific question, important insights could be generated from the large literature exploring valuation issues for treatment (which he noted was often labelled as 'cure' within these studies) versus prevention. Husereau reported a potential disconnect between existing literature reporting individual and societal preferences. That is, when given a choice between prevention and treatment, individuals appear to state a preference for prevention. However, similar preferences are not apparent when (societal) willingness to pay is considered. This disconnect was attributed to separate psychological factors including time, certainty of individual decisions and the valuation of identifiable versus statistical lives. Husereau concluded that if a similar disconnect exists for curative therapies relative to standard treatments this could lead to considerable public debate and that further research was required.

Importantly, several of these issues and challenges highlighted in the workshop do not appear unique to regenerative medicines and cell-based therapies and similar challenges are often present when appraising more conventional biologics, companion diagnostics and devices more generally. However, it appears reasonable to conclude that these issues will be more commonly faced within evaluations concerning regenerative medicines and cell therapies. Furthermore, since many regenerative medicines and cell therapies will be considered as both a biologic and a device/procedure, manufacturers may have to address the specific regulatory and reimbursement challenges faced by both pharmaceutical and device manufacturers internationally.

Due to the personalised nature of regenerative medicines, the manufacturing and production process is typically more complex than that for traditional drug therapies. Current pharmaceutical manufacture is largely based around drugs being prepared, tested and manufactured in bulk at a consistent quality in advance of need, using automated processes. In contrast, many regenerative medicines require significant personalisation at the point of need. The complexity of the process, and the high level of personalisation required, may result in significantly higher marginal costs of production compared to conventional pharmaceuticals or biologics. Inevitably, these additional complexities are likely to lead to higher upfront costs to healthcare systems.

This complexity and personalisation is likely to be coupled with a requirement for the provision of additional healthcare services within the overall process. The additional demands raise issues around the impact on the wider healthcare setting, both at a marginal cost and wider infrastructure level. These demands may differ according to the extent of the services provided by the manufacturer and those requiring separate funding by the health care system. In the sipuleucel-T appraisal, the lack of experience in the UK of using the treatment raised additional uncertainties surrounding whether the NHS would incur additional costs that were not reflected in any of the scenarios evaluated, raising an additional source of uncertainty.[81]

Since provision of regenerative medicines and cell-based therapies often entails multiple procedural steps (e.g. cell extraction, processing and administration), and may be undertaken alongside additional procedures (e.g. leukapheresis in the case of sipuleucel-T and arthrotomy for ACI), additional uncertainties are likely to be raised concerning the generalisability and transferability of evidence between different settings. That is, separating out the specific effect of the regenerative medicine or cell therapy from the effect of broader health provision, which itself may be subject to significant variation across different health care systems, represents an important challenge to HTA and cost-effectiveness assessments.

Many regenerative medicines and cell therapies also appear likely to share similar 'unique' characteristics which have been reported for medical devices.[92] For example, particular parts of the procedural process may change significantly over time, experiencing incremental or step changes as new processes and infrastructure develops. Additionally, the requirement for highly specialised infrastructure and staff indicates the potential for a learning curve over time both for manufacturers and health care providers. While increased automation methods and the 'scaling out' of services may subsequently reduce the need for highly specialised staff (and lower the marginal costs of production), the infrastructure requirements and implications for possible learning curve effects are likely to be an important consideration when assessing the cost-effectiveness of regenerative medicines and cell therapies.

In addition to issues related to uncertainty, issues of irrecoverable costs may pose an additional challenge. Irrecoverable costs are those costs which once committed cannot be recouped if changes occur at a later time, most commonly thought of as investment costs associated with the capital expenditure on equipment, new facilities, or training and learning costs. These are likely to be most significant when the introduction of a new regenerative medicine or cell therapy imposes additional infrastructure requirements on the health system. Within economic evaluations, these costs are commonly annuitised and allocated as 'per-patient' costs by spreading the cost over the number of patients likely to be treated during the lifetime of the equipment. However, if reimbursement decisions about the technology change before the end of the lifetime of the equipment (e.g. approval is withdrawn), then these costs may not be recovered and hence need to be explicitly considered.

The risks of these more conventional types of irrecoverable costs to the health system may be more limited if the manufacturer provides the necessary infrastructure and associated training. However, irrecoverable costs also potentially exist at the patient level. Regenerative and cell therapies are developed, by design, to have a significant (if not permanent) period of effect, during which they may be neither removable nor reversible. The irreversibility of these therapies implies any uncertainty associated with the long term efficacy and adverse event profile has a greater potential significance than for treatments that can easily be reversed or switched.

4.4.2.1 Potential approaches to addressing HTA challenges

Many of the issues associated with regenerative medicines will inevitably impact on the level of uncertainty associated with the cost-effectiveness of the technology when introduced into clinical practice. Even where products have significant potential to confer important clinical advances over current therapies, this may not be known with a high level of certainty at the time a regenerative medicine is licensed. Inevitably a new technology's cost-effectiveness may be more difficult to

determine in these circumstances and schemes that allow the development of further evidence or entail a risk-sharing component may be required.

Managed entry agreements (MEAs) are an increasingly common policy response to dealing with uncertainty in the evidence base of new health technologies entering the market. MEAs are also commonly referred to as performance-based risk sharing agreements (PBRSAs) and patient access schemes (PASs). A taxonomy provided by Walker *et al.*[93] makes an important distinction between MEAs based on reductions in the effective price (e.g. akin to majority of PAS schemes implemented currently alongside NICE guidance) and those associated with evidence generation (e.g. OIR, CED). Both approaches have the aim of reducing risk and decision uncertainty to the health care system, albeit via separate mechanisms. Walker et al concluded that reimbursement decisions and the possible use of MEA should be based not only on the expected value of a technology but also the value of further research, the anticipated effect of coverage on further research, and the costs associated with reversing the decision (i.e. irrecoverable costs).

Similar conclusions are reflected in NICE's current methods guide for technology:

"*When the evidence of clinical effectiveness or impact of a technology on other health outcomes is either absent, weak or uncertain, the Appraisal Committee may recommend that the technology is used only in the context of research or while the technology is recommended as an option, research is also conducted. Before issuing such recommendations the Committee will consider the following factors:*

- *the need for and potential value to the NHS of additional evidence that can inform the development of NICE guidance and clinical practice on the use of the technology*
- *the uncertainty in the analysis and what could be gained by reconsidering the decision in the light of research findings whether the research is feasible in circumstances when the Appraisal Committee recommends the intervention for NHS use outside the context of research*
- *irrecoverable costs incurred from introducing the technology the likely net benefits for all NHS patients of use only in a research setting during the time that the recommended research is being conducted.*

In considering these factors the Committee will balance the potential net benefits to current NHS patients of a recommendation not restricted to research with the potential net benefits to both current and future NHS patients of being able to produce guidance and base clinical practice on a more secure evidence base".[83]

How to determine when efficiency is sufficiently weak or uncertain, such that MEAs are appropriate policy responses remains a key methodological issue that has important implications both for policy making and research investments made by the regenerative medicine industry. A more formal framework been recently been proposed which has established the key principles of the assessments needed.[88] There is also ongoing work being undertaken by NICE's Decision Support Unit (DSU) related to the methods for assessing MEAs within the TA programme.

Concerns surrounding the potential high-upfront costs of regenerative medicines and affordability to health care systems have also received particular attention in the literature, leading several authors to conclude that alternative financing approaches may also have to be considered. For example, Towse[87] outlined 3 alternative financing routes: (i) pay for performance; (ii) amortisation and (iii) innovative financing schemes. The potential issues and challenges related to alternative financing approaches in the context of regenerative medicine and cell-based therapies are discussed in more detail below.

Fixed price schemes

The simplest and most common approach to reimbursement is the payment of a fixed price at the time of treatment, potentially subject to discounts agreed via a PAS. This approach has the benefit of being relatively manageable and low cost to implement. Furthermore, subject to uncertainties concerning the eligible patient population size and subsequent uptake rates, the budget impact is also largely predictable. However, if the therapy is expensive and/or the patient population is large, the total budget impact to the funder may raise issues concerning system 'affordability' (i.e. the ability to displace sufficient activities elsewhere in the health care system to provide the additional funds necessary to provide the new treatment). This may have implications for subsequent implementation.

While 'affordability' is not explicitly considered by NICE, the Committee is requested to take *"account of how the incremental cost effectiveness of the technology being appraised relates to other interventions or technologies currently or potentially applied in the NHS".* When significant uncertainty exist surrounding future outcomes, a fixed price scheme exposes the funder to the risk of overpayment for outcomes which may not be realised. A fixed pricing mechanism may be potentially optimal in situations in which: there is little uncertainty about the long term outcomes; there are high costs of patient follow up; and/or the resulting budget impact is likely to represent a marginal change to overall NHS spend.

Amortisation

Amortisation has been raised as an alternative financing approach for curative treatments, particularly for chronic diseases.[87, 94] Gottlieb and Carino[94] identified that most health care finance systems are not currently structured to be able to pay to rapidly cure everyone of a chronic disease using a treatment

that may be priced much higher than the existing chronic therapies. Payment models were therefore advocated that could more easily spread the potentially high upfront costs of a curative treatment and be more closely aligned to the time period over which health and economic benefits are realised. Such financing schemes are common for capital equipment payment for medical devices and other capital equipment. However, for a regenerative medicine or cell therapy, instead of spreading out payment over the lifetime of the capital equipment, amortisation would spread payment over the expected duration of benefits.

Gottlieb and Carino[94] highlighted several issues and challenges relating to operationalising such as scheme. These included: the potential need for another financial intermediary to act as a third party to the transactions; the need to alter accounting standards; and potential conflict with the manufacturer's desire to secure immediate revenue to maximise return on their investment. The magnitude of initial R&D costs versus the ongoing marginal costs of production might also influence the interest rate the funder would have to offer to sufficiently incentive both existing and future manufacturers.

Pay for performance

Although the use of an amortisation approach to financing might address the constraints imposed within current financial structures, this approach does not reduce the risk to health care systems of uncertain future health benefits which may not be realised in routine clinical practice.

In contrast, a pay for performance type mechanism ensures that total price paid is more directly related to the performance of the therapy in clinical practice. This mechanism requires agreement between the funder and manufacturer on the measure of performance (e.g. response, relapse or mortality), the program of data collection and analyses required to monitor performance and the payment mechanism itself (e.g. fixed price at time of treatment with rebate, retrospective reimbursement for treatment 'successes', or amortised payments directly linked to performance over time).

As with amortisation, the potential to spread repayments over the longer-term reduces the short-term budget impact. This financing approach also potentially addresses the uncertainties surrounding the potential health benefits and the risk of overpayment (i.e. where the opportunity costs are subsequently revealed to be greater than the acquisition cost) by the funder. Inevitably, such a mechanism is likely to be both more complex and expensive to implement than a simple PAS or amortisation approach. However, there are examples of existing PAS schemes within the NICE TA programme which already incorporate performance assessments and discontinuation rules are commonly applied within NICE appraisals to 'optimise' cost-effectiveness and reduce decision uncertainty.

A pay for performance mechanism is potentially optimal if there is large uncertainty about long term outcomes, a relatively low cost of patient follow up and monitoring of the outcome(s) of interest (relative to the level of uncertainty), and a large total budget impact that, as with amortisation, can be spread over time. The potential challenges concern how performance in clinical practice would be monitored and evaluated and whether a simple assessment of continued treatment 'success' is feasible or not.

A recent paper by Edlin et al.,[95] proposed a leasing approach for innovative technologies as an alternative payment strategy combining elements of amortisation with pay-for-performance approaches. The advantages of this approach is that it addresses both the funding constraints caused by existing finance structures whilst also ensuring that the risks associated with uncertain future health benefits are more appropriately shared between the funder and the manufacturer.

Edlin proposed that having established the price at which the technology is expected to be cost-effective, the 'lease' payment due for each period of health delivered could be established by calculating a stream of payments over the expected lifetime of the technology that has the same expected net present value as the agreed price. The subsequent leasing scheme would work by paying the manufacturer for each period of time that health is delivered at the individual patient level. That is, if the observed effectiveness in clinical practice was equal to the expected effectiveness, the manufacturer would receive the full value of the technology over the agreed period. However, if observed effectiveness was less than expected, payment would stop and the risk to the health system of over-payment would be limited. Furthermore, manufacturers would be rewarded for technologies which resulted in better health outcomes than expected by receiving additional payment over extended periods of time.

Using trastuzumab (Herceptin) in early breast cancer as an exemplar, and linking the lease to relapse-free survival, Edlin demonstrated that the scheme not only reduced the total budgetary impact but also resulted in a more appropriate share of risk between the manufacturer and the funder, while simultaneously reducing the value of further research. Edlin concluded that such a scheme could help promote the rapid adoption of innovative technologies into routine clinical practice.

Innovative financing

Several authors have argued for even more innovative approaches to pricing to be considered, seeking inspiration from the wider financial world. For example, innovative licencing and the issuing of bonds, by which third party payers cover the costs of treatment, benefiting from the respective interest rate paid by the healthcare funder.[91] Such mechanisms have has some success in the provision of vaccination programs in developing nations (through the International Finance Facility for

Immunisation scheme), appealing to investors seeking ethical investments. An alternative mechanism, considerably closer to a pay for performance mechanisms is the Health Impact Fund, where manufacturers distribute innovations at cost but are rewarded with performance based bonuses.

Possible implications for NICE methods and processes

In considering the potential characteristics of regenerative medicines and cell-based therapies and associated challenges for HTA, NICE will need to consider whether changes to their current processes and methods are required or not. Importantly, some of the potential challenges highlighted are already considered within the existing Methods Guide.[83] For example, the Committee is already requested to recognise that the evidence base will necessarily be weaker for some technologies, such as those used to treat patients with very rare diseases. If considered appropriate, this could be extended to include regenerative and cell therapies. Similarly, although the magnitude of budget impact is stated not to determine the Appraisal Committee's decision, the existing method guide indicates that the Committee may require more robust evidence on the effectiveness and cost effectiveness of technologies that are expected to have a large impact on NHS resources. However, a potential conflict may arise between the certainty required for interventions with large budget impact and subsequent deliberations regarding the acceptability of 'weaker' evidence.

NICE's existing processes also make separate provision for specific disease and technology characteristics which may be relevant to many regenerative medicines and cell therapies. NICE's current end-of-life criteria (EoL) allows the Committee, when considering the overall health benefits, to explore a QALY weighting that is different from that of the reference case, assuming all the stated criteria are met. The methods guide also states that this approach can also be used in other circumstances when instructed by the NICE board. Further research may be warranted to determine whether a similar weighting approach might be appropriate for regenerative medicines and cell therapies. However, there remains an issue regarding whether such a weight should be based on product specific characteristics or patient specific characteristics (i.e. not confined to product type).

Within the provisions and regulations of the Health and Social Care Act 2012 relating to NICE, due regard is also required concerning "*the desirability of promoting innovation in providing health services or social care in England*" (Section 6.1.3 Methods Guide)[83]. This is currently incorporated within the Committee's deliberative process for situations in which the most plausible ICER exceeds £20,000 per QALY gained. In this situation, the innovative nature of the technology, specifically if adds "*demonstrable and distinctive benefits of a substantial nature which may not have been adequately captured in the reference case QALY measure*". Importantly, neither of the previous regenerative medicines appraised by NICE to date were considered to demonstrate such benefits.[78, 81]

The NICE methods guide also permits separate provision to be made via the specific discount rate it applies. Within NICE reference case for cost-effectiveness analysis, the same annual discount rate is required to be used for both costs and benefits (currently 3.5%). However, the use of a non-reference-case discount is permitted using rates of 1.5% for both costs and health effects in cases when treatment restores people who would otherwise die or have a very severely impaired life to full or near full health, and when this is sustained over a very long period (normally at least 30 years). Hence, certain regenerative medicines and cell therapies may be considered to meet these criteria. However, uncertainties remain regarding how the likelihood of achieving these long-term health benefits will be considered by the Committee, particularly in the context of the uncertainties outlined in this section. Furthermore, the stipulation that the Committee will also need to be satisfied that the introduction of the technology does not commit the NHS to significant irrecoverable costs, raises additional issues about whether this criteria will be met.

Issues of discounting have been widely considered in the economic literature in relation to preventative treatments and particularly vaccination programmes. The appropriateness of employing different discount rates and/or different rates over time is an area which requires further consideration, particularly for potentially curative regenerative medicines and cell based therapies. Westra et al.[96] explored the impact of employing alternative discount rate approaches for HPV vaccination based on different time varying methods: a stepwise approach (a constant rate is applied for a set period and lowered in subsequent periods), a hyperbolic approach (the discount rate declines over time), and a time-shifted approach. A recent review by Jit et al.[97], also noted that the UK Treasury currently recommends stepwise discounting to all public sector bodies, but at a very slowly declining rate (3.5% for the first 30 years, declining to 3.0% from year 31 and with further declines from year 76). While the use of discounting seeks to incorporate social preferences rather than alleviate uncertainty, further consideration could be given to their application to regenerative medicines and cell therapies.

Another approach commonly employed to better characterise the uncertainties surrounding longer term benefits and to inform the Committee's deliberations relate to the time horizon of the analysis and the methods of extrapolation. Within NICE's current methods guide, alternative scenarios are requested to be routinely considered to compare the implications of different methods for extrapolation of the results.[83] Several other country-specific guidelines for cost-effectiveness analyses request presentation of alternative scenarios based on different time horizons and comparing within-trial results with extrapolated results.

Alteration of the time horizon of the evaluation, away from the lifetime analysis recommended by NICE where appropriate, acts to reduce the uncertainty by excluding the impact of costs and

outcomes that occur in the long term, where uncertainty is likely to be greatest. Whilst such analyses are potentially informative in terms of understanding how influential particular assumptions are over the period of extrapolation, restricting the horizon risks omitting important costs and outcomes related to a particular technology and simply shifts the risk associated with particular uncertainties from the health system to the manufacturer.

Similarly, separating the within-trial results from the extrapolated results (or considering alternative scenarios for extrapolation) has been argued to allow separation of the uncertainty due to downstream consequences separately from other sources of uncertainty. Mortimer[98] suggested that this approach could enable decision-makers to assign a weight to the results of the extrapolation to take account of various uncertainties. However, it was also acknowledged that such a comparison is not always explicitly made and that implicit comparisons were often problematic since the relationship between the within-trial and extrapolation period may not be predictable. Issues of predictability may be a particular challenge for regenerative medicines and cell therapies. While it is commonly argued that a within-trial analysis is conservative with respect to cost-effectiveness estimates, the author identified situations where this may not be true e.g. long term adverse effects offsetting any initial gains or where increased survival is associated with additional costs related to the disease and/or other unrelated diseases. Mortimer also highlighted that the relativity in results between within-trial analysis and the results of extrapolation was made even more problematic when uncertainty due to future technological change is introduced. Several key factors were highlighted as affecting the relativity of results between within-trial and extrapolated result including the timing of potential technological advance, the proportion of patients who could benefit when the new technology becomes available and the effectiveness of the new technology.

Conclusions

The review has identified a number of common themes and potential challenges in relation to HTA and assessments of cost-effectiveness for regenerative medicines and cell therapies. Some of the challenges identified do not appear unique to these types of therapies and are also faced by manufacturers of more conventional pharmaceuticals, biologics and devices. However, it seems likely that these challenges may be faced more routinely for regenerative medicines and cell therapies.

There already exists provision within NICE's methods guide to accommodate some of these aspects, although potential challenges may arise in ensuring these are consistently applied between committees and understood by manufacturers. NICE will also need to consider whether further amendments to their processes and methods are required. Broader consideration will also need to be given to approaches which may extend beyond NICE's existing remit e.g. alternative funding approaches.

Consequently other bodies and manufacturers themselves may also have an important role in identifying more innovative approaches to seeking reimbursement which recognise the inherent uncertainties and lead to a more efficient sharing of associated risk.

5 Exemplar technology appraisal of a regenerative medicine

5.1 Selection of exemplar

Following Regenerative Medicine Expert Group (RMEG) subgroup discussions and further input from the Cell Therapy Catapult, it was decided that undertaking an exemplar appraisal involving a real commercial product was not feasible for a number of reasons: there would be significant commercial sensitivities; products undergoing regulatory review would be candidates for a real appraisal; and using a product at an earlier stage in clinical development would not be helpful as the evidence base would be even less mature and, therefore, would not have the attributes of an 'exemplar' product. It was therefore proposed to undertake the evaluation of a hypothetical product.

As a result of both RMEG subgroup discussion and technical meeting discussion, the type of regenerative medicine chosen as the hypothetical product was CAR (chimeric antigen receptor) T-cell therapy specific to the antigen CD19. The chosen indication was relapsed or refractory acute lymphoblastic leukaemia. This specific combination was selected based on the existence of relatively mature data sets – in the context that none of the currently available CAR T-cell products are licensed.

5.1.1 About CAR T-cell therapies

CAR-T therapies have been under development for around 20 years. The specific CAR T-cell therapies considered in this appraisal consist of autologous (i.e. the treated individual's) T-cells which are genetically modified to redirect the target of the T-cell receptors. These receptors target specific proteins found on the surface of leukaemia cells, in this case the protein CD19, which is present on B-cell leukaemias as well as on healthy B cells, but it is not found on hematopoietic stem cells (which are situated in the bone marrow) nor on other tissues.[99] The activated T-cells can then attack and destroy the leukaemia B-cells. Persistence of a given CAR T-cell therapy within the body is linked to the properties of the T-cell from which the cells were derived as well as the immune environment into which they are infused. CAR T-cell therapies have already begun to evolve, with 2nd generation therapies currently being trialled in phase II studies. Research efforts at developing future generations are focused on addressing the key challenges of T-cell target specificity, persistence and ability to exert the desired anti-tumour effects as well as identifying new target antigens.[100] CAR T-cell therapies have recently emerged as regenerative medicines with promising potential to treat haematologic cancers. In July 2014 the FDA granted 'breakthrough therapy' status to the CAR T-cell therapy CTL019 (manufactured by Novartis) for the treatment of adult and paediatric relapsed or refractory acute lymphoblastic leukaemia.[101]

Although CAR T-cell therapies may offer relapsed or refractory B-ALL patients a 'bridge' to stem cell transplantation, or possibly even a cure for B-ALL, it is likely that patients will need to be

monitored for some key adverse effects which are often reported. These include cytokine release syndrome (CRS) and B-cell aplasia. Cytokine release syndrome occurs as a result of cytokines being released from the successfully-targeted cancer cells and can result in various symptoms such as fever, headache, nausea, and a rash. The severity of CRS appears proportional to the tumour burden. Although CRS is an adverse effect of CAR T-cell therapy, there may be a correlation between the development of CRS and response to therapy; patients who do not develop CRS may be less likely to benefit from CAR T-cells, while those who develop CRS often respond to the therapy. While there may be some correlation between developing CRS and efficacy, there does not appear to be a strong correlation between the *degree* of CRS and response to therapy.[102]

An absence of B cells - referred to as B cell aplasia - is an expected adverse effect of successful CAR T-cell therapies which eliminate normal mature and precursor B cells. As long as CAR T-cells persist, B cell aplasia continues (which provides what appears to be a highly accurate pharmacodynamic marker of CAR function).[102] B-cell aplasia is a manageable disorder; patients may be treated with intravenous immunoglobulin (IVIG) though this is an expensive treatment. Persistent B-cell aplasia could result in an increased risk of infection even with replacement therapy.[103]

5.1.2 Overview of disease

B-cell acute lymphoblastic leukaemia (B-ALL) is a sub-type of acute lymphocytic (lymphoblastic) leukaemia (ALL). ALL is a cancer that starts from the immature lymphocytes in the bone marrow, invading the blood fairly quickly, and then can spread to other parts of the body, including the lymph nodes, liver, spleen, central nervous system (brain and spinal cord), and testicles (in males). The term "acute" means that the leukaemia can progress quickly, and if not treated, would probably be fatal within a few months. The American Cancer Society's estimates for acute lymphocytic leukaemia (ALL) in the United States for 2015 (including both children and adults) are that there are about 6,250 new cases of ALL (3,100 in males and 3,150 in females) and around 1,450 deaths from ALL (800 in males and 650 in females).[104] The risk for developing ALL is highest in children younger than 5 years of age. While most cases of ALL occur in children, most deaths from ALL (about 4 out of 5) occur in adults.[104]

UK statistics present a similar picture. Statistics for the incidence of ALL in the UK (2009-2011) are provided by Cancer Research UK, based on data sourced from Office of National statistics, ISD Scotland, Welsh cancer Intelligence and Surveillence Unit and the Northern Ireland Cancer Registry (Figure 1).[105] Across all ages in 2011 there were 654 new cases reported in the UK (males 377, females 277), with crude incidence rates of 1.2 for males and 0.9 for females (per 100,000). Incidence is strongly related to age, but ALL is unusual as it does not follow the pattern of increasing incidence

with age seen for most cancers, instead the highest incidence rates are in children, teenagers and young adults. In the UK between 2009 and 2011, an average of 65% of cases were diagnosed in people aged under 25, and only 6% were diagnosed in those aged 75 and over. Age-specific incidence rates are highest in infants aged 0-4 and drop sharply through childhood, adolescence and young adulthood, reaching their lowest point at age 35-39, and increasing slightly thereafter. Incidence rates are similar between males and females in all age groups except age 15-19, when age-specific rates are significantly higher in males (male:female ratio of around 22:10). Averaged across all patients aged <30years the mean number of cases of ALL per year is 462.

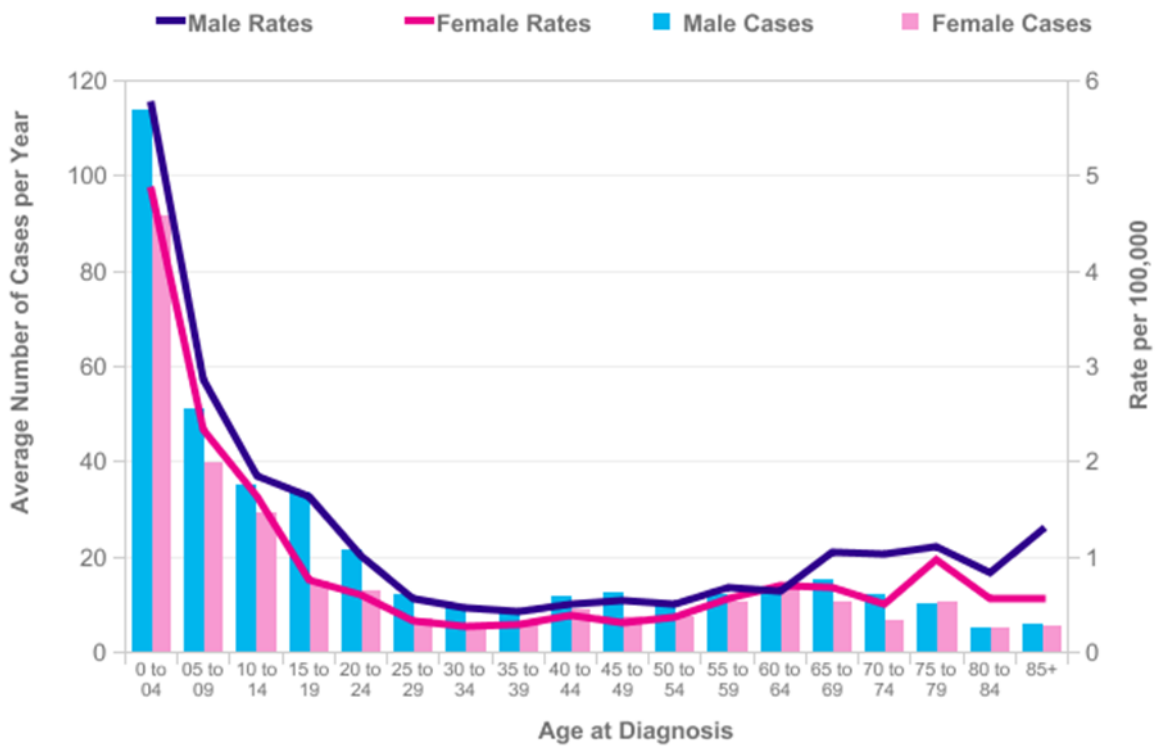


Figure 1 Average number of cases of ALL per year in UK and Age-Specific Incidence Rates (2009-2011)

There are subtypes of ALL based on the type of lymphocyte (B cell or T cell) and how mature these leukemia cells are.[104]B-cell ALL comprises about 80% to 85% of ALL cases. There are several subtypes of B-cell ALL: Early precursor B (early pre-B) ALL (also called *pro-B ALL*); Common ALL; Pre-B ALL; and Mature B-cell ALL (also called *Burkitt leukaemia*). This last type is rare, accounting for only about 2% to 3% of childhood ALL; it is essentially the same as Burkitt lymphoma and is treated differently from most leukaemias. T-cell ALL comprises about 15% to 20% of cases of ALL. This type of leukaemia affects males more than females and generally affects older children more than does B-cell ALL. It often causes an enlarged thymus (a small organ in front of the

windpipe), which can sometimes cause breathing problems. It may also spread to the cerebrospinal fluid (the fluid that surrounds the brain and spinal cord) early in the course of the disease.

Based on an estimate of 82.5% of ALL being B-cell ALL, there will be 540 cases of B-cell ALL in the UK per year, of which 381 will be in those aged under 30 years. Approximately 20% of these will be refractory to treatment or relapse,[106] giving an estimate for young people with relapsed/refractory B-cell ALL in the UK to be 76.

5.1.3 Overview of current practice

Whilst the management of ALL in adults and children is similar, the prognosis is different, with that in adults (aged over 30 years) being much poorer than in the younger age group. Hence they are generally considered as two distinct clinical groups.

With stepwise improvements of risk-adapted chemotherapy and supportive care over the past five decades, current overall cure rates of newly diagnosed ALL are approaching 90% in the developed world in children and around 50% in adults.[107] Treatment involves induction with combination chemotherapy for the attainment of complete remission (CR) (both clinical and haematological) followed by post-remission maintenance therapy with or without HSCT (which enhances relapse prevention particularly in patients younger than 35 years). However, due to the morbidity and mortality risks associated with transplant, HSCT is usually reserved for high-risk patients.

US National Comprehensive Cancer Network Guidelines (NCCN) recommendations for first-line treatment are based on risk stratification and age, as follows:[108]

- Philadelphia chromosome–positive (Ph+) ALL (in AYA): Chemotherapy and tyrosine kinase inhibitor (TKI), followed by allogeneic stem cell transplantation (SCT) if an appropriate donor is available; if transplantation is not feasible, continue multi-agent chemotherapy and a TKI
- Ph+ ALL (Adults < 65 y): Chemotherapy and tyrosine kinase inhibitor (TKI); consider allogeneic SCT if an appropriate donor is available and the patient has good performance status and no or limited comorbidities; if transplantation is not feasible, continue multi-agent chemotherapy and a TKI
- Ph+ ALL (Adult ≥ 65 y or with substantial comorbidities): TKI and corticosteroids or TKI and chemotherapy (evaluate end-organ reserve, end-organ dysfunction, and performance status)
- Ph- ALL (AYA): Paediatric-style multi-agent chemotherapy
- Ph- ALL (Adults < 65 y): Multi-agent chemotherapy
- Ph- ALL (Adults ≥ 65 or with substantial comorbidities): Multi-agent chemotherapy or corticosteroids (evaluate end-organ reserve, end-organ dysfunction)

However little progress has been made in the treatment of relapsed ALL. Following initial induction and maintenance therapy most adults will relapse and long-term leukaemia-free survival is achieved in only 20-30% and following relapse, response rates to further chemotherapy are low at around 20-30% and long-term OS rates of 3-24% have been reported[109] From a UK study of 608 adult patients, OS at 5 years in newly diagnosed patients was 38% (95% CI: 36%-41%) but only 7% (95% CI: 4%-9% after relapse.[110]

Relapse is less common in paediatric ALL but accounts for the highest proportion of cancer deaths in children. [111]Studies of Nordic and Austrian data found that of children with ALL 25% had a first relapse, 8% had a second[112, 113] and 2% a third relapse.[113] Around 50% of relapsed ALL in children does not respond to salvage therapy and for these patients survival rates are below 10%.[113] In children, age and white blood cell count at primary diagnosis of ALL are the most important prognostic factors for relapse: age < 1 year or $\geq$10 years is associated with the worst prognosis. In addition, site of relapse and duration of first remission are the major criteria for the classification of patients after first relapse.[111]

Therapy after relapsed ALL comprises re-induction chemotherapy followed by consolidation chemotherapy and/ or HSCT. Time to relapse (length of first remission), site of relapse and ALL-immunophenotype are established factors that are prognostic at first relapse and can be used to determine further treatment.[111] B-cell ALL has a better prognosis than T-cell. Various regimens have been investigated and re-induction remission rates of 71% to 95% have been reported; the higher rates are generally associated with later first relapse. Patients who are refractory to re-induction therapy or who have a further relapse have a poor prognosis, with survival rates below 10%.[113] Failure to achieve complete remission (CR) to late re-induction chemotherapy is associated with previous failures to achieve CR or short remission. The proportion of patients achieving CR has been shown to reduce with subsequent relapses: in a study of 225 patients with ALL (195 B-cell ALL) mean CR rates (SE) were 83% (4%) for early first marrow relapse, 93% (3%) for late first marrow relapse, 44% (5%) for second marrow relapse, and 27% (6%) for third marrow relapse. Five-year DFS rates in CR2 and CR3 were 27% (4%) and 15% (7%) respectively.[106] Although some therapies with curative intent are capable of inducing a second remission in patients refractory to previous therapy, these are often associated with high treatment-related morbidity, -mortality and minimal survival.[114] Such patients are eligible for innovative therapies in Phase 1 or 2 trials. Therapies for relapsed B-ALL which have been licensed by the EMA or FDA are discussed in section 5.3. In particular clofarabine, a purine nucleoside anti-metabolite, (which affects DNA elongation, synthesis and repair) was granted EMA marketing authorisation in 2006 for use in children and young adults with a 2nd or greater relapse (or refractory patients). The pivotal trial of clofarabine (n=61 with 2nd or greater relapse) reported an

overall remission rate of 20% (12/61 patients) with 16% (10/61) of patients going on to receive HSCT. Clofarabine had been studied only in single arm trials, marketing authorisation was granted 'under exceptional circumstances'.

As discussed in Section 4.3, the use of complete remission as an outcome is not specific at predicting which patients might subsequently relapse. In recent years evaluation of response to therapy in B-ALL patients has become more precise with the development of methods to detect minimal residual disease (MRD). Although the FDA have concluded that the evidence base to indicate that early MRD status is the strongest predictor of long term EFS in ALL is unequivocal, there is some uncertainty in how MRD- correlates to long-term outcomes in relapsed populations. The use of MRD as a surrogate endpoint is discussed further in section 4.3.

Haematopoietic stem cell transplantation (HSCT)

Allogeneic HSCT is a potentially curative treatment option and a number of studies have demonstrated improved outcomes compared with chemotherapy. However there are difficulties in the interpretation of the findings of many such trials such as patient selection bias, specific nature of HSCT, source of donor cells, and the adverse effects associated with various specific transplant therapies.[111]

The risks associated with transplant include graft rejection, delayed immune reconstitution, graft versus host disease, and vulnerability to infections. In addition there is significant toxicity associated with the chemo-radiotherapy conditioning required before transplant.[111] The adverse effects of HSCT are not limited to those occurring in the short-term. One study investigated long-term survival and late deaths among 1,458 ALL patients who were disease-free two years after allogeneic HSCT; the median follow up was around 80 months.[115] Of the 167 deaths, new cancers accounted for around 10% of the primary causes of death, while graft versus host disease accounted for 23%.

One study examined MRD in 157 patients with ALL in morphologic remission undergoing allogeneic HSCT following a myeloablative conditioning regimen, (12 patients were post 2 or more relapses). The 3-year OS for those who were MRD-negative pre-transplant was 68% compared to 40% in the MRD-positive group, while the probabilities of relapse were 16% vs 33% for the two groups respectively. Twenty-four of the 153 patients had T-ALL, and the trend towards increased relapse in those with MRD was similar to that seen with B-ALL (HR 7.07). Post-transplant among those with any sample positive for MRD, the risk of subsequent relapse was higher (HR 3.21) as was the risk for overall mortality (HR 2.54).[116] Similar findings were reported in a study that included a slightly larger sample of post-2nd or later relapse (n=18).[70] Based on these and other studies there is the suggestion of benefits through MRD-directed therapies, though controlled trials are needed to define their value.[69]

Immunotherapies are showing promise and being investigated. These therapies are targeted at specific surface antigens expressed on the target cells: naked and un-conjugated antibodies; immunoconjugates and immunotoxins; bi-specific T cell engaging (BiTE) therapy (blinotumobab); and CAR T-cells. The latter two target CD19.[109]

5.1.4 Decision problem

The key aspects of the decision problem to be addressed were identified and agreed at a meeting between the Academic Group and a sub-group of the project advisory group (topic experts). These were based on discussion of the phase I/II CAR T cell trials in B-ALL which had been identified by literature searches performed by Cell Therapy Catapult. The agreed components of the decision problem were:

Intervention: CD19 CAR T-cell therapies

Indication: Patients with B-cell acute lymphoblastic leukaemia (B-ALL) who have relapsed (with no further planned curative chemotherapy or haematopoietic stem cell transplant (HSCT)) or who are refractory to standard chemotherapy. As described in section 5.1.3 the treatment pathways and prognosis for patients under- and over- aged 30 years are very different. Consequently, the indication considered in this assessment has focussed on those aged under 30 years.

Subgroups: Sources of heterogeneity such as relapsed/refractory status, previous HSCT, CAR design, dose, conditioning chemotherapy, tumour burden at the time of therapy, or age of the patients may be explored.

Comparators: Best supportive care (e.g. salvage chemotherapy)

Efficacy outcomes: Response criteria such as complete response/remission (CR), partial response/remission (PR), and minimal residual disease negative (MRD); overall survival (OS); progression and/or event-free survival; persistence of CAR T-cells; health-related quality of life; rates of HSCT

Adverse event outcomes: Cytokine release syndrome (CRS), B-cell aplasia, febrile neutropenia, neurologic effects.

5.2 Review of evidence of clinical effectiveness: CAR T-cell therapies

5.2.1 Methods

Initially, studies of CD19 CAR T-cells in B-ALL were identified by staff at Cell Therapy Catapult (who comprise part of the project Advisory Group, being experts in cell-based regenerative medicines). PubMed was searched using the search terms "CD19", "chimeric antigen receptor", "CAR", and "CD19 CAR" with a cut-off date of 21st October 2014. The clinicaltrials.gov trials registry, and relevant published reviews of CAR T-cell therapies were also searched.

In order to identify any further relevant clinical trials we performed update searches in Medline and Embase to May 2015. One reviewer performed an initial screen of the abstracts; those deemed potentially relevant were then screened by a second reviewer. Google searches to identify further data on already identified trials were also undertaken. Our clinical advisor was also contacted regarding any relevant 2015 conferences where new data may have been recently presented. To further inform the study design details of the hypothetical data sets, the clinicaltrials.gov trials registry was searched for ongoing trials that had commercial involvement: the focus was on trials designed with the likely aim of acquiring marketing authorisation.

5.2.2 Overview of studies

Three published papers were identified from the Cell Therapy Catapult searches.[117-119] No further studies with results were identified from the Medline and Embase update searches. Two conference abstracts[120, 121] (relating to two of the published studies) and one conference video[122] (relating to one of the published studies) with more up to date data were identified from the Google searches.

Of the planned, or (other) ongoing CAR T-cell studies identified on clinicaltrials.gov, seven had commercial involvement: three were phase II trials, all with estimated enrolments of 67 patients; one was a phase I/II trial with an estimated enrolment of 80 patients and three were phase I trials (Table 7). All were single-arm studies. Two of the phase II trials were multi-centre studies (i.e. they listed more than one centre for recruitment in the Contacts and Locations field) and three had a primary outcome which assessed response or remission; the time frames stated for these outcomes ranged from 9 weeks to one year. Only one trial reported the collection of longer-term survival data, with a stated time frame of 5 years for overall survival, event-free survival, and relapse-free survival.

5.2.3 Efficacy Results

Details of the three trials are presented in Table 2. Two of the studies were phase I trials and one was categorised as a phase I/IIA trial (on clinicaltrials.gov); safety was the primary outcome in all trials (one study also had maximum tolerated dose as a co-primary outcome). All trials had recruited fewer than 40 patients, though two were ongoing.[118, 119]

As can be seen from Table 6, notable clinical heterogeneity was evident both within and across trials. One study was of children and adults and one studied children and *young* adults. The remaining study was only of adults[119] and so ultimately was not of further use for the assessment. Most patients had had a prior relapse following remission; a small proportion of patients were refractory to previous treatments. In two studies the CAR T-cell treatments were mostly used to enable patients to receive HSCT (i.e. used as a bridging therapy).[117, 119] The remaining study appeared to recruit a more difficult to treat population, with most patients having two or more relapses *and* previous HSCT; here the treatment intention may possibly have been curative.[118] CAR T cell design also varied, with either the CD28 or the 4-1BB co-stimulatory domains being used; a difference which might explain the more prolonged persistence of circulating CAR T-cells seen in one of the studies.[102] Persistence of CAR T-cell therapies in the body can result in benefit and risk, depending on the duration of persistence.

Two surrogate endpoints were reported in all 3 trials: complete remission (CR) and minimal residual disease (MRD). Rates of CR ranged from around 70% to 90%. As would be expected, the rates of achieving a status of MRD were lower, ranging from around 60% to 80%, although only one trial stated the MRD threshold used, which was 0.01% (i.e. 1 cancer cell in 10,000 normal cells).[117] All 3 trials reported overall survival (OS) data. In one trial the probability of OS at 9.7 months was 52%.[117] The other 2 trials reported probabilities of OS at 6 months which were 58%[121] and 78% respectively.[118]

5.2.4 Adverse effects

The key adverse events noted in the trials were cytokine release syndrome, B-cell aplasia, febrile neutropenia, and various neurological effects. In 2 studies most patients had mild to moderate cytokine release syndrome (CRS) although a greater incidence of severe CRS was evident in the trial in adults.[119] Affected patients were treated with steroids or tocilizumab. Two of the 3 studies reported incidence of B-cell aplasia. In one study prolonged B-cell aplasia did not occur[117] and in the other B-cell aplasia occurred in all patients who had a response, and persisted for up to one year after CAR T-cells were no longer detectable.[118] Significant proportions of patients had febrile neutropenia or neurological adverse effects such as hallucinations or altered mental status (Table 6).

Table 6 Data available for the three published CAR T-cell trials in patients with B-ALL

Study details	Study (authors, year, product and sponsor)					
	Lee et al 2014[117] CD19-CAR T cells, NIH funded	**Maude et al 2014[118] CTL019, Novartis**	**Grupp abstract[120]**	**Grupp video[122]**	**Davila et al 2014[119] 19-28z CAR T Cells, Juno Therapeutics**	**Park 2015 abstract[121]**
Design						
ct.gov identifier	NCT01593696	NCT01626495 (children, young adults) NCT01029366 (adults)			NCT01044069	
Primary outcome(s)	SAEs, Maximum tolerated dose	SAEs			SAEs	
Study design	Phase I feasibility/dose escalation	Phase I/IIA			Phase I	
No. of centres	1	2 (1 adult, 1 children)			1	
Planned duration of follow up	5 years (from ct.gov)	Unclear (but appears to be 1 year for most outcomes)			2 years (from ct.gov)	
Sample size	21	30	30	39	16	33 (32 evaluable)
Duration of follow up (so far)	Median 10 months	6 months (median follow up 7 months)	6 months	6 months	NR	Median follow up 5.1 months
Population						
Relapsed/refractory	14 relapsed, 7 refractory	27 relapsed, 3 refractory			All 16 relapsed	14 pts had ≥ 3 prior lines of therapy
No of prior relapses	1:6 pts; ≥2: 8 pts	1:5 pts; ≥2: 22 pts			Appears to be 1	
Age	14 children, 7 adults; range 5 to 27 yrs	25 children, median 11 yrs (range 5 to 22); 5 adults, median 47 yrs (range 26 to 60)	All children or young adults, median age 10yrs (range 5 to 22)	All children or young adults	Adults; median 50 yrs (range 18 to >60)	Median 54 yrs (range 22 to 74)
No. of B-ALL patients	20/21*	29/30*			16/16	
MRD threshold	<0.01% by flow cytometry	Unclear, but measured by flow cytometry			Unclear, but measured by flow cytometry	
MRD- at baseline	0	5/25 children	5/25		2/15	
Previous HSCT?	8	18/25 children			4/16	11
Intervention						
CAR T cell dose (per kg)	Mostly 1 $x10^6$ (15/21) or 3 $x10^6$ (4/21)	1 to 10 $x10^7$, or 5 to 50 $x10^8$ (if over 50kg)	10^7 to 10^8		3 $x10^6$	1 to 3$x10^6$
Conditioning regimen	Fludarabine and cyclophosphamide	15 pts had Fludarabine and cyclophosphamide			Cyclophosphamide	
Type of dosing regimen	Single dose; 3 pts had 2nd infusion	Given over 1-3 days			Split dose (days 1 and 2)	
Efficacy outcomes						
Complete response/remission	Day 28: 14/21 (66.7%, 95% CI 43.0 to 85.4)	1 month: 27/30 (90%)	27/30 (90%)	Day 28: 36/39 (92%) Probability at 6	Complete remission: 10/16 (63%) Time point unclear (median time to CR/CRi	29/32 (91%)

Study details	Study (authors, year, product and sponsor)					
	Lee et al 2014[117] CD19-CAR T cells, NIH funded	**Maude et al 2014[118] CTL019, Novartis**	**Grupp abstract[120]**	**Grupp video[122]**	**Davila et al 2014[119] 19-28z CAR T Cells, Juno Therapeutics**	**Park 2015 abstract[121]**
	70% (95% CI 45.7 to 88.1) in 20 patients with B-ALL			months: 76% (95% CI 61 to 94)	24.5 days)	
Partial response/remission	Not reported	Not reported			Not reported	
MRD negative	Day 28: 12/20 (60%, 95% CI 36.1 to 80.9) in 20 patients with B-ALL	22/30 (at perhaps 1 month – time point unclear)	23/30 (77%)		12/16 (75%) Time point unclear	23 of the 28 MRD evaluable pts (82%)
Overall survival	Probability at 9.7 months 51.6% (median 10 months)	Probability at 6 months 78% (95% CI 65 to 95)	Probability at 6 months: 78% (95% CI 63 to 95)		NR	6 months: 58% (95% CI: 36 to 74)
Progression-free survival	78.8% in the 12 pts achieving MRD- status (beginning at 4.8 months)	NR			NR	
Event-free survival	Not reported	Probability at 6 months 67% (95% CI 51 to 88)	Probability at 6 months: 63% (95% CI 47 to 84)	Probability at 6 months: 70%	NR	
Persistence of CAR T cells	18/21 had detectable CAR T cells with peak expansion occurring around day 14. No CAR T cells detected after day 68 in any patient.	Probability at 6 months: 68% (95% CI 50 to 92)	6 months: 68% (95% CI 50-92)		Peak of CAR T cells within 1-2 weeks, with decreases to low or undetectable levels by 2 to 3 months.	
HSCT	10	3 children withdrew from study following treatment with CTL019 to have HSCT		3	7 out of 10 eligible pts. Of the remaining 6: 3 were contraindicated, 2 declined and 1 was being evaluated	11
QoL	NR	NR	NR	NR	NR	NR
Safety outcomes						
Cytokine Release Syndrome (CRS)	Grade 3: 3 Grade 4: 3	All patients had CRS; mild to moderate in 22/30 pts. Severe in 8 patients (needed ICU treatment). 9 patients received tocilizumab. All patients recovered fully	All responding pts developed grade 1-4		7/16 had severe CRS (definition of severe provided). Patients received steroids and/or tocilizumab.	7 had severe CRS
B-cell aplasia	Prolonged B cell aplasia did not occur	Occurred in all patients who had a response and persisted for	6 months: 73% (95% CI 57 to 94)		NR	NR

Study details	Study (authors, year, product and sponsor)					
	Lee et al 2014[117] CD19-CAR T cells, NIH funded	**Maude et al 2014[118] CTL019, Novartis**	**Grupp abstract[120]**	**Grupp video[122]**	**Davila et al 2014[119] 19-28z CAR T Cells, Juno Therapeutics**	**Park 2015 abstract[121]**
		up to one year after CTL019 cells were no longer detectable. These patients received immunoglobulin replacement.				
Febrile neutropenia	Grade 3: 7	22/30 required hospitalisation			Grade 3: 11	
Neurologic effects	Hallucinations: 5 Dysphasia: 1	13 had neurologic effects: 6 had delayed encephalopathy			Grade 3 altered mental status: 5 Grade 3 altered mental status: 1	
Notes						
IPD data available in published trial report	Age, sex, no of relapses, % marrow blasts, previous treatment, complete response, MRD, HSCT	Lymphodepleting chemo, complete remission, severe CRS, persistence, B-cell aplasia			Age, salvage chemo, MRD, HSCT, CRS	
CR definition	Complete response: <5% marrow blasts, absence of circulating blasts, and no extra medullary sites of disease with absolute neutrophil count 1000 per μL or more and platelets 100 000 per μL or more.	Complete remission: morphologic assessment of the bone marrow as M1 (<5% leukemic blasts) with no evidence of extra medullary disease.			Complete remission: restoration of normal hematopoiesis with a neutrophil count > 1,000 x 106/L, a platelet count > 100,000 x 106/L, and haemoglobin > 10 g/dL. Blasts should be < 5% in a post-treatment bone marrow differential. No clinical evidence of leukemia for a minimum of four weeks.	
Other notes	All toxicities fully reversible 15 years follow up for delayed AEs (FDA guidance) * 1 patient had non-Hodgkin lymphoma	Reasons for not having HSCT after CTL019: lack of suitable donor, prior HSCT, family choice. * 1 patient had T-cell ALL			CAR T cells given as bridge to HSCT 2 patients who achieved CR after CAR T cells had a CR prior to CAR T cell infusion	

Table 7 Ongoing commercial CAR T-cell trials registered on clinicaltrials.gov

Clinicaltrials.gov identifier	Study name	Phase / design	Estimated enrolment	Primary outcome(s) and time frame	Secondary Outcomes and time frame
NCT02228096	Study of Efficacy and Safety of CTL019 in Pediatric ALL Patients (Relapsed/refractory)	II Single-arm, open label, multi-centre	67	Overall Response Rate (ORR), which includes Complete Remission (CR) and Complete Remission with Incomplete Blood Count Recovery (CRi), as determined by assessments of peripheral blood, bone marrow, CNS symptoms, physical exam (PE) and CSF (1 year)	Adverse events and laboratory abnormalities (type, frequency and severity) (1 year)
NCT02167360	Study of Efficacy and Safety of CTL019 in Adult ALL Patients (Relapsed/refractory)	II	67	Safety (1 year)	None reported
NCT02435849	Determine Efficacy and Safety of CTL019 in Pediatric Patients With Relapsed and Refractory B-cell ALL	II Single-arm, open label, multi-centre	67	Overall remission rate (ORR): includes Complete Remission (CR) and CR with incomplete blood count recovery (CRi) as determined by independent review committee (IRC) assessment. (6 months)	Percentage of patients who achieve CR or CRi at Month 6 without SCT between CTL019 infusion and Month 6 response assessment Percentage of patients who achieve CR or CRi and proceed to SCT while in remission before Month 6 response assessment Duration of remission (DOR) (60 months) Percentage of patients who achieve CR or CRi with minimal residual disease negative bone marrow (60 months) Relapse-free survival (60 months) Event-free survival (60 months) Overall survival (60 months) In vivo cellular PK profile (levels, persistence, trafficking) of CTL019 cells (60 months) Prevalence/incidence of immunogenicity (60 months) to CTL019
NCT01840566	High Dose Therapy and Autologous Stem Cell Transplantation Followed by Infusion of Chimeric Antigen Receptor (CAR) Modified T-Cells	I Single-arm, open label	18	Maximum tolerated dose Safety (2 years)	2 year progression-free (PFS) Overall survival (2 years)

Clinicaltrials.gov identifier	Study name	Phase / design	Estimated enrolment	Primary outcome(s) and time frame	Secondary Outcomes and time frame
	Directed Against CD19+ B-Cells for Relapsed and Refractory Aggressive B Cell Non-Hodgkin Lymphoma				
NCT01430390	In Vitro Expanded Allogeneic Epstein-Barr Virus Specific Cytotoxic T-Lymphocytes (EBV-CTLs) Genetically Targeted to the B-Cell Specific Antigen CD19 Positive Residual Or Relapsed Acute lymphoblastic leukemia After Allogeneic Hematopoietic Progenitor Cell Transplantation	I Single-arm, open label	26	Safety (3 years) Persistence of escalating doses (3 years)	To assess the effects of the adoptively transferred CD19 specific T-cells on the progression of leukemia. (3 years) To quantitate the number of chimeric antigen receptor (CAR) positive T-cells in the blood at defined intervals post infusion in order to determine their survival and proliferation in the host (3 years) To assess long-term status of treated patients (15 years)
NCT01683279 (aka PLAT -01)	A Pediatric Trial of Genetically Modified Autologous T Cells Directed Against CD19 for Relapsed CD19+ Acute Lymphoblastic Leukemia	I Single-arm, open label	18	Number of Participant with Adverse Events (42 days)	Persistence of the CD19 CAR+ T cells (42 days) Determine if there is anti-leukemic activity of the CD19 CAR+ T cells (42 days)
NCT02028455 (aka PLAT -02)	A Pediatric and Young Adult Trial of Genetically Modified T Cells Directed Against CD19 for Relapsed/Refractory CD19+ Leukemia	I/II Single-arm, open label	80	Safety (30 days) MRD negative complete remission (63 days) Releasable cell product generated (28 days)	Persistence of the CD19 CAR+ T cells (63 days) Number of participants with recrudescence or development of acute GVHD (63 days) Number of participants who have T cells ablated with cetuximab (3 years)

5.2.5 Summary issues for the target product profile and hypothetical data sets

- B-ALL population is narrowly defined with extremely poor prognosis and limited alternative therapy options. This is likely to be typical of regenerative medicines
- Therapy potentially offers a 'cure'
- Potentially serious adverse effects
- Limited data available (single-arm studies)
- Appropriate comparator and control data need to be identified /generated.

5.3 Review of licensed treatments for relapsed/refractory B-ALL

A pragmatic review of the other treatments for relapsed/refractory B-ALL which have been licensed by the EMA or FDA was undertaken. This was done to further inform decisions to help construct the CAR T-cell therapy hypothetical data sets, and to help to put them in context. Three treatments were quickly identified from the B-ALL literature and EMA/FDA websites: Evoltra, Blincyto and Marqibo. This number of treatments was deemed sufficient for the purposes of this exercise.

Evoltra (clofarabine) - known as Clolar in the U.S. - was granted EMA marketing authorisation under exceptional circumstances in 2006. It is a purine nucleoside anti-metabolite (which affects DNA elongation, synthesis and repair). Blincyto (blinatumomab) and Marqibo (vincristine sulphate liposome injection) were both licensed by the FDA under the accelerated approval programme (in 2014 and 2012, respectively); in this programme, drugs for serious conditions that fill an unmet medical need may be approved based on a surrogate endpoint. Blincyto, which has also been granted 'breakthrough therapy' designation, is a monoclonal antibody designed to specifically attach to CD19 proteins on leukaemia cells. Marqibo is a targeted delivery of vinicristine which involves encapsulation of vinicristine in nanoparticle liposomes.

Marqibo and Blincyto are licensed for use in adults, and Evoltra (clofarabine) for use in children and young adults. All three treatments have orphan product designation, all claim to meet unmet medical need, and the submissions for all were primarily based on data from phase II single-arm trials. However, whereas Marqibo and Evoltra are licensed for patients with a 2nd or greater relapse (or refractory patients), the approval for Blincyto is broader, covering patients with "Philadelphia Chromosome negative relapsed or refractory B-precursor ALL"; over half the patients in the pivotal Blincyto study had had only one relapse. A consequence of these different populations can be seen in the pivotal trial sample sizes, with smaller studies for Marqibo (n=65) and Evoltra (n=61) and a larger study for Blincyto (n=189).

Primary outcomes for all three trials were based on remission status (complete and/or overall remission). All studies also reported overall survival. For the 2nd or greater relapse populations, treatment with Evoltra resulted in an overall remission rate of 20% (12/61 patients) with 16% (10/61) of patients going on to receive HSCT; treatment with Marqibo resulted in a complete remission rate of 15% (10/65 patients) - although the figure was 12% based on the FDA's assessment - with 18% (12/65) receiving HSCT. However, most of these patients did not achieve complete remission with Marqibo. As would be expected, a higher rate of complete remission was seen in the Blincyto trial (42%), when compared with the Marqibo and Evoltra trials, since most patients were at 1st relapse. Further results and other assessment details are presented in Appendix 4.

The EMA review of Evoltra stated that, given the efficacy seen early on in the clinical programme, studies using a placebo comparator were considered clinically unethical. Active comparator studies were also not deemed to be appropriate as there were no other recognised therapeutic options available: "The indication is encountered so rarely that the applicant cannot reasonably be expected to provide comprehensive data on clinical efficacy and safety." Marketing authorisation was therefore granted 'under exceptional circumstances'.[123] The AWMSG (All Wales Medicines Strategy Group) recommended Evoltra only if the intended use was as a bridge to HSCT (and that it should not be used with palliative intent).[124] The FDA approval of Marqibo seemed less straightforward; committee members consistently stated that the proposed phase 3 trial was critical in assessing the benefit of Marqibo. Some members indicated that the trial should be completed before approval, while several indicated that accelerated approval may be appropriate, but with the expectation that this approval would be withdrawn if the phase 3 trial failed to confirm clinical benefit. The post-approval study was a multi-centre phase III randomised trial comparing standard vinicristine with Marqibo in older adults with newly diagnosed untreated Philadelphia negative ALL; the proposed sample size was 348.

For Blincyto, a confirmatory phase III RCT was required which will compare Blincyto with standard care chemotherapy in relapsed/refractory adults; this was ongoing at the time of submission to the FDA. A 2:1 randomisation ratio was used, with more patients receiving blinatumomab. Around 400 patients were expected to be enrolled and the primary endpoint is overall survival.

5.3.1 Single arm B-ALL trials: identifying appropriate control data

As discussed in section 4.2, although the results of single-arm trials can be compared with historical control data, the results of such comparisons can only be considered as reliable indicators of treatment benefit when the disease natural history is very well known, the patient population is homogenous and the standard care treatment has little impact on outcomes. For the Evoltra, Blincyto and Marqibo different approaches were used to devise a control data set.

For Blincyto a weighted analysis of patient-level data from 694 retrospective controls (1990-2014) was performed.[125] This study, which was of relapsed/refractory adults treated with standard of care therapy, utilised databases in several EU countries and the US. The manufacturer of Blincyto (Amgen) also conducted a model-based meta-analysis of clinical study data to project the effect of Blincyto relative to existing therapies. For their ongoing trial in children, Amgen cited both a key paper on prognostic factors in B-ALL,[106] and the complete remission rates seen with Clolar (clofarabine), stating that their primary efficacy endpoint would be met if the CR+CRh* (CRh*= CR with incomplete haematological recovery) rate was at least 22.5% (suggesting efficacy similar to, or greater than that for clofarabine).[126] For the Marqibo FDA submission literature searches were performed to identify response rates in relevant patients. The study reporting the best historical comparison data still had some key differences with the Marqibo trial population, most notably in terms of line of treatment, eligibility for transplant, and site of adjudication.[127] Comparisons were made with the closest matched subgroup of patients in the historical study: patients who received third-line single-agent treatment.[128] Clofarabine was assessed in 2006 so the aforementioned O'Brien study was not available. Instead data were obtained from German and Dutch cancer registries; simple comparisons of median survival results were presented to the EMA.[123]

5.3.2 Summary

- Studies which form the basis of regulatory submissions of treatments for patients with 2nd or greater relapsed/refractory B-ALL will be small (around 65 patients), phase II, single-arm trials
- Primary endpoints will be surrogate endpoints such as complete remission
- Confirmatory randomised trials may be appropriate and viable in related larger populations, where other treatment options exist.
- For very small patient populations it is likely to be difficult to identify published prognostic studies which have suitable historical control data. Other strategies may therefore be needed, such as seeking access to national patient databases (in order to perform new studies)

6 The target product profile and hypothetical data sets

6.1 Summary of issues to consider to inform creation of the Target Product Profiles (TPPs) and hypothetical data sets

The innovative nature of regenerative medicines, together with the indications that many of them will be expected to target initially (populations with high levels of unmet medical need), means there is a collective desire to expedite their approval and appraisal. This ambition may run counter to the need for additional vigilance relating to robust evidence and long-term outcomes. Regulatory bodies must therefore endeavour to balance urgency of patient need with the requirement for robust evidence on efficacy and safety. This can be managed through a combination of regulatory approval based on limited - although promising - data, combined with post-approval requirements for continued data collection. From the perspective of NICE appraisals, this means that the evidence base available at the time of product approval may be highly uncertain; the cost of this uncertainty has to be a key part of the decision making process.

The reviews have identified several broad issues relevant to uncertainty around the clinical evidence for the creation of the TPPs and hypothetical data sets for the exemplar:

- It is not universally the case that regenerative medicines (or ATMPs) will be tested using non-randomised study designs. Rather, submitted pivotal studies may well in fact be randomised, notably when levels of unmet need are low and diseases/conditions are not rare; in such cases the maturity of data (which would be available at the time of a NICE appraisal) has been up to five years duration.
- When single-arm trials, or case series, *do* form the basis of a regulatory submission, a key consideration when judging uncertainty should be the likelihood of cure or improvement *without* experimental treatment. However, it may be very difficult to identify published prognostic studies which have suitable historical control data. Other strategies for obtaining historical data may well be needed, such as seeking access to national patient databases.
- Where single-arm trial data are compared with historical data and appropriate methods to adjust for confounding are employed, the selection of the method used must be explicit and based on sound reasoning; despite advances in statistical techniques clear challenges remain in generating accurate unbiased estimates of effect from non-randomised data.
- Results from single-arm trials can only be considered as reliable indicators of treatment benefit when the disease natural history is very well known, the patient population is homogenous, and the control (standard care) treatment has little impact on outcomes.
- Although more mature evidence, such as confirmatory RCTs, may sometimes be viable in the specific population, it might also only be expected in larger, similar populations (for example, B-

ALL patients in *first* relapse). This raises the possibility of incorporating indirectly relevant but more reliable (and possibly more mature) data into the analysis, to reduce uncertainty.

- The high technology status of regenerative medicines may imply greater potential for variation in response across both individuals and centres. This is likely to have implications in terms of the generalisability of efficacy and safety estimates obtained from small single-centre (probably expert centre), single-arm studies; in the absence of larger or more varied trials, this might only be addressed by access to individual patient data so that potential predictors of response or effect modifiers may be investigated.
- Another key issue is that pivotal trials in regulatory submissions are likely to report primary endpoints which are surrogates for real clinical endpoints. On average, trials using surrogates report larger treatment effects than trials using final patient-relevant outcomes. This has implications for effect estimate uncertainty, especially when *only* surrogate endpoints are reported; the choice of surrogate outcomes used should be researched, explicit and justified. Nevertheless, to maximise the use of all available data, and to reduce overall uncertainty, multivariate meta-analysis methods to analyse data should be considered - whatever the maturity of the evidence base.
- Related to the issue of surrogates as primary outcomes is that of duration of follow-up: use of intermediate shorter-term outcomes avoids the need for long follow-up. The consequence of this is that even where overall survival data are recorded, these data are immature at the point of regulatory approval.
- Regenerative medicines are by their nature innovative products and may be subject to continuing development, with new generations having improved efficacy. This may pose problems when evaluating long-term efficacy and safety; for example, to what extent can the long-term safety data from a first-generation product be used to inform long-term safety of a related newly-licensed second-generation product? This may mean that as well as bioavailability-type studies, key trials conducted earlier in the development process may have to be replicated or adjustments be made in the analyses of trial data to account for their indirectness.

For the specific purpose of deciding what to include in the exemplar hypothetical data sets, the best information to begin with comes from the published and ongoing trials for CAR T-cells, together with the EMA/FDA licensed non-regenerative medicines for relapsed B-ALL. These indicate that a minimum data set would comprise of a small (around 65 patients), phase II, single-arm trial, with surrogate endpoints such as complete remission as the primary endpoints. Minimum residual disease (MRD), the surrogate endpoint which is the strongest predictor of long term event-free survival in ALL, is also likely to be reported, although there is considerable uncertainty about its value in relapsed populations.

Historical control data must be identified which should reflect the treatment B-ALL patients would receive in the absence of CAR T-cell therapies being available. This is necessary to utilise the hypothetical trial evidence within the economic analyses. A key challenge for constructing the historical control group will therefore be identifying the population included within the single-arm studies and selecting an appropriate control group: any selected control group is unlikely to exactly match the tiny population included in the single-arm studies so comparisons will therefore be subject to confounding. To mitigate the effects of any such bias a second challenge will be to identify and apply the most appropriate methods to adjust for confounding.

The small sample sizes available from the trials of CAR T-cell in relapsed/refractory B-cell ALL imply that estimates of effect are likely to be inexact and imprecise and this should be considered when creating the more mature data sets.

More mature data sets would be expected to have larger (tending towards appropriately powered) sample sizes to reduce the width of the confidence interval around any effect estimate. It should be noted that this would not influence the magnitude of any potential bias and may lead to increased confidence in an incorrect estimate of effect. Increasing the sample size may however, also allow for a wider range of statistical methods to mitigate the effects of confounding and therefore have an indirect effect on reducing any bias in the effect estimate.

An RCT could be included in a more mature data set or the availability of data from an RCT in B-ALL patients in *first* relapse could be proposed. This latter possibility raises methodological questions of how the results of confirmatory RCTs in an indirect population might be used to re-evaluate the uncertainty of the direct evidence base.

6.2 Background to developing the TPPs and evidence sets

Data from the 3 published trials for CAR T-cell therapies were discussed at a meeting of the Project Advisory Group on 24th June 2015. Based on these discussions it was decided that for the purposes of

the exemplar, the population would comprise children and young adults who had experienced two or more relapses or were refractory to treatment (with older adults excluded). It was further decided that the exemplar would explore both the therapeutic goals of the CAR T-cell therapy encompassing bridging and remission/curative intents.

Two target product profiles (TPP) were subsequently developed to be considered as part of the exemplar appraisal:

- CAR T-cell therapy used as a "*bridge to HSCT*", where the primary goal of treatment is to induce short-term remission of disease in order to maximise the opportunity for successful HSCT, and
- CAR T-cell therapy used with "*curative intent*", where the primary goal of CAR T-cell treatment is long-term remission/cure of disease (with or without HSCT)

These two approaches to treatment with CAR T-cell therapy imply two potentially different contexts in which therapy may be appraised. Consequently, there are separate implications arising from the different applications that require their consideration as two distinct scenarios.

In the "*bridge to HSCT*" scenario, the survival benefits of treatment are determined primarily by the subsequent receipt of HSCT and the associated benefits which stem from this. As such, the health benefits of CAR T-cell therapy are closely linked to the HSCT status of the cohort in the immediate period following CAR T-cell therapy. From a regulatory and reimbursement standpoint, the primary determinant of treatment efficacy is likely to include short-term endpoints such as remission, and potentially MRD status. These data may also be supported by data on the outcomes of HSCT after CAR T-cell. Marketing approval may therefore be achieved through demonstrating clinical benefit in terms of remission, potentially MRD status and subsequent rates of HSCT.

In the *"curative intent"* scenario, the survival benefit of treatment is considered to be as a direct result of CAR T-cell therapy itself. In this context, there is no separate surrogate treatment or process (i.e. HSCT) which determines the long-term benefits of therapy. From a regulatory standpoint, the primary determinant of the efficacy of treatment in this scenario is likely to include longer-term clinical endpoints such as event-free survival (EFS) and OS and increased levels of data maturity may be required.

New technologies are submitted to licensing agencies to seek regulatory approval and are subject to NICE appraisal at various stages of development of the supporting evidence base. To explore the impact of different levels of precision and maturity in the evidence base, 3 hypothetical data sets were constructed for each TPP:

- **The minimum set**: the minimum data considered potentially sufficient for CAR T-cell therapy to be granted conditional regulatory approval.
- **The intermediate set**: a variant of the minimum set, where the efficacy and safety of CAR T-cell therapy has been assessed over a longer follow-up period.
- **The mature set**: a variant of the intermediate set where the efficacy and safety of CAR T-cell therapy has been assessed in a larger clinical study but with a similar follow-up period as the intermediate set.

In developing the TPPs, it was not our intention to directly compare the separate scenarios or to use these to infer differences between the alternative CAR T-cell therapies currently being developed. Neither were the different evidence sets intended to be prescriptive regarding the sufficiency of evidence for the purposes of regulatory or reimbursement processes. Instead, the hypothetical TPPs were developed to provide an exploration of potential issues and challenges associated with varying levels of precision and maturity in the underlying evidence base and the potential impact that these might have on subsequent assessments of cost-effectiveness and associated decision uncertainty.

In total, six evidence sets were developed spanning the separate TPPs (3 sets for "bridge to HSCT", and 3 sets for "curative intent"). Each of the 3 evidence sets includes hypothetical efficacy and safety data for CAR T-cell therapy and for a historical control. The efficacy and safety estimates in the "*bridge to HSCT*" and "*curative intent*" TPPs, were derived from data from Lee et al[117] and Maude et al[118] respectively, reflecting the clinical heterogeneity and the potentially different treatment intentions reported in Section 5.2.

6.3 Defining a historical control

The lack of control data within existing CAR T-cell studies necessitates the selection of a historical control from existing published literature to inform the TPPs and economic model. As discussed in Section 4.2, the use of a historical control introduces potential bias as observed or unobserved confounders other than the treatments may impact the outcomes of interest. As such a direct comparison of the CAR-T cell results and a historic control may be subject to bias.

Observable sources of confounding can be potentially adjusted for in a number of ways depending on if the available data is being considered at a single study level or through the synthesis of evidence from a number of studies (as discussed in Appendix 2). A key source of potential observable confounding relates to differences in patient characteristics which are known to be related to subsequent prognosis. To identify prognostic factors which might provide a basis for adjusting a historic control to account for potential prognostic imbalance, a search was conducted to identify previously published multivariate prognostic models reported for patients with ALL.

The search identified 12 potentially relevant studies. However, five of these did not reported sufficient detail on the results to be considered further.[113, 114, 127, 129, 130] A summary of the patient characteristics in the remaining 7 studies are reported in Appendix 7. None of the prognostic models specifically focussed on the population of interest. In addition, there appeared little consistency across the prognostic factors selected for inclusion in the multivariate analyses, with no single factor considered across all models and only few of which could be applied to the patient characteristics reported within existing CAR T-cell studies. Hence, while the formal adjustment for potential bias is desirable, the lack of access to individual patient data meant this was not considered feasible within the exemplar.

A separate search was subsequently conducted to identify possible historical control studies that might be more generalisable to the population of interest (i.e. based on age and prior history of relapse) and which might minimise the potential for bias in the absence of a formal adjustment for confounding. This search identified two studies considered to be potentially generalisable to the population considered within the exemplar evaluation.[114, 131] Jeha et al.[131] reported on a phase II open label study of clofarabine in paediatric patients with refractory or relapsed ALL. Von Stackelberg et al[114] conducted a retrospective analysis of outcomes in children and adolescents with ALL who had not responded to salvage therapy and evaluated the overall of survival of the patient population given different treatment modalities (curative, palliative and no therapy).

For consistency across TPPs and evidence sets, the same historical control and data was considered in both scenarios. Clofarabine and the study by Jeha et al were subsequently selected to act as the control treatment and source of historical control data, for the following reasons:

- Clofarabine is considered a standard of care chemotherapy for B-cell relapsed refractory ALL, alongside other chemotherapies such as fludarabine, cytarabine, granulocyte colony-stimulating factor, and idarubicin (FLAG-IDA).
- It is the only EMA licensed treatment available for ALL in paediatric patients who have relapsed or are refractory after receiving at least two prior regimens and where there is no other treatment option anticipated to result in a durable response.
- Although clofarabine has not been appraised by NICE for this indication, clofarabine is currently funded through the cancer drugs fund (CDF). At the time of writing, clofarabine was the only treatment for relapsed and refractory B-cell ALL approved on the CDF.
- The phase II study population was also considered to be broadly consistent with populations enrolled to Maude et al and Lee et al.

6.4 Developing the evidence sets and target product profiles

The first step in developing the hypothetical evidence was to define the sample size and maturity of evidence (i.e. follow-up) for the minimum, intermediate and mature evidence sets. The second step involved generating the efficacy and safety data conditional on the sample and maturity levels specified for each of the evidence sets. This involved the synthesis of data reported in the existing CAR T-cell, historical control studies and simulation modelling.

6.5 Defining the sample size and maturity of evidence in the evidence sets studies

Current evidence on the efficacy of CAR T-cell therapies is limited to early phase I/II studies with patient numbers reported between 16 and 30.[117, 118, 132] It is anticipated that larger clinical studies will be needed to meet the minimum requirements for positive regulatory approval. This evidence is expected to come from a series of phase II/III studies. Several previous pharmaceutical treatments for ALL have been granted regulatory approval based on efficacy and safety data from single arm phase II studies with sample sizes ranging from 61 to 189. This appears to be reflected in the design of planned and ongoing CAR T-cell studies in ALL. According to the ClinicalTrials.Gov trials registry, there are currently 3 registered phase II trials investigating the efficacy and safety of CAR T-cell therapies. The planned sample size of these trials is 67 patients (see Table 7). There is also one phase I/II trial with an estimated enrolment of 80 patients.

The expected minimum data requirement for regulatory approval with CAR T-cell therapy was therefore set in the region of 60-80 patients. This sample size was used in both the minimum and intermediate evidence sets. The mature evidence set was assumed to be based on trial evidence derived from a larger sample of patients than the minimum and intermediate evidence sets. This evidence set was designed to reflect a scenario where the evidence base for CAR T-cell therapy could include data from a more conventional RCT (or alternatively a larger uncontrolled study) with sufficient duration of follow-up to determine the longer-term efficacy for key clinical endpoints including OS. In practice, the sample size and maturity for such a study would be determined by a number of factors, including conventional statistical power calculations, likely accrual rates, the competitive landscape, and overall study costs. In the time available, it was not feasible to formally consider these elements in estimating the anticipated sample size for this study. Instead, the sample size was based on the planned enrolment size of an ongoing phase III trial of blinatumomab in adult ALL, identified from previous hand-searching of Clinical Trials.Gov. In this study, the planned enrolment was for 400 patients to be randomised to blinatumomab or standard of care at a ratio of 2:1. As such, of the total 400 randomised patients, 133 would be randomised to the control arm and 267 would be patients randomised to blinatumomab. For the mature set, the study sample size was therefore set in the region of 120-140 patients per treated group (240-280 in total).

In using the specific sample size for the mature evidence set, we recognise that there are differences between the patients recruited into the blinatumomab trial and the specific population being considered here both in terms of the age and prior history of relapse. However, for the purposes of a hypothetical exemplar this was considered to provide a reasonable basis for investigating the potential impact of increased precision.

The intermediate and mature evidence sets were also assumed to be based on trial evidence with longer efficacy follow-up than the minimum set. For the minimum set, trial follow-up was based on a similar duration reported within existing CAR T-cell studies, with a median follow-up of approximately 10 months. For the intermediate and mature sets, trial follow-up was based on the maximum planned study duration for all phase II CAR T-cell trials registered on the Clinical Trials.Gov registry. Across these studies, the longest planned follow-up period was 5-years.

A summary of the targeted sample size and levels of evidence maturity considered across each of the evidence sets is provided in Table 8.

Table 8: Summary of the sample size and maturity of trial evidence assumed in the 3 evidence sets

	Minimum	**Intermediate**	**Mature**
Sample size	60-80	60-80	120-140
Study follow-up	10 months (median)	60 months (maximum)	60 months (maximum)

6.6 Estimating the efficacy of CAR T-cell and comparator treatments in the evidence sets

For all dichotomous outcomes, including response, remission and use of HSCT, parameter estimates were extracted directly from the existing CAR-T cell and clofarabine publications. The effect of increased sample size on the variance parameter for each dichotomous outcome was modelled using a beta distribution. As these outcomes tend to be measured during the first few months of a study, it is expected that longer follow-up would not directly impact these parameter estimates.

For the overall survival endpoint, parameter estimates were derived by digitising the Kaplan-Meier (KM) curves reported in the main study publications and using the algorithm by Guyot et al[43] to impute the patient-level time to event, and event type (censored or event) data. These data were then analysed using conventional semi-parametric survival modelling techniques using the statistical programming platform R.[133] This included assessments of landmark survival probabilities at 6-, 12- and 60-months, and derivation of the hazard ratio for CAR T-cell versus standard of care therapies

and restricted mean survival times (i/e/ in the event of non-proportional hazards, formally tested using the Schoenfeld residual test).

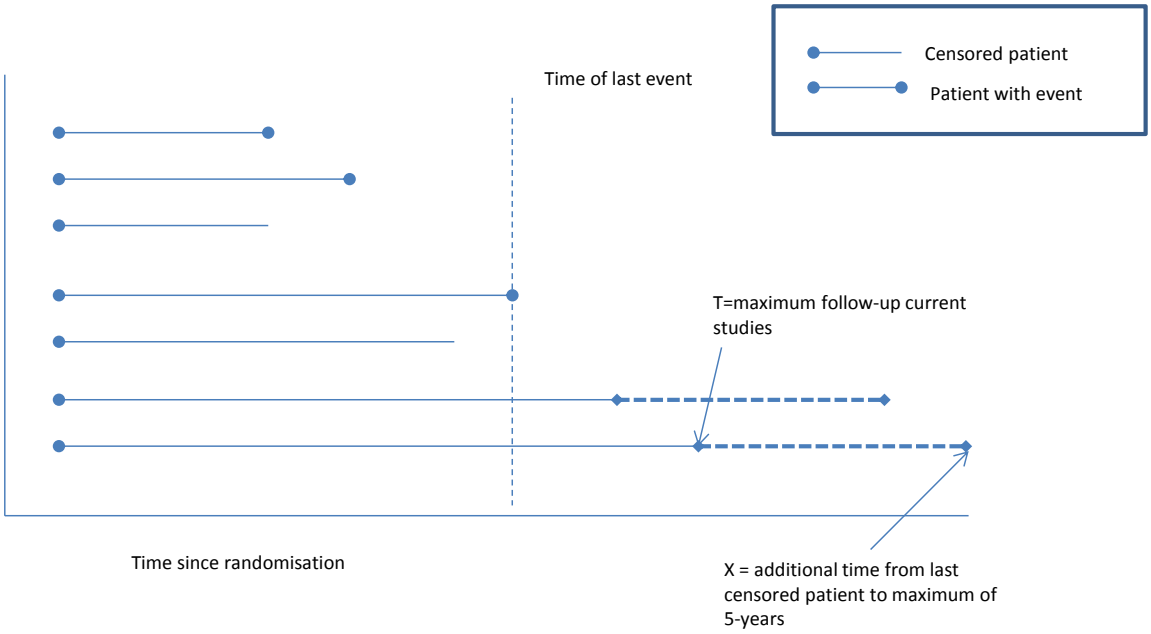
Using the Maude and Lee studies, it was possible to generate samples of between 21 and 30 patients treated with CAR T-cell therapy. However, the expected minimum sample size for regulatory approval may be in the region of 60-80 patients. Therefore, it was necessary to increase the size of the imputed data sets from between 21 and 30 to between 60 and 80. This was achieved by replicating each of the imputed data sets until the total sample size was between 60-80 patients for the minimum/intermediate evidence sets and 120-140 for the mature evidence set. By creating the pooled sets in this way, the mean survival probabilities and KM plots for overall survival remained consistent across evidence sets, whereas the variance around those estimates were allowed to vary in line with the sample size.

For the mature and intermediate data sets, it was also necessary to simulate an increase in the duration of study follow-up to account for a scenario where CAR T-cell studies had longer follow-up, up to a maximum of 5-years. This adjustment was made by adding survival time to the imputed patient records that were censored after the last recorded event in each study. These patients were subsequently assumed to be re-censored at their new survival time. The same approach was applied in both the CAR T-cell and clofarabine arms. By extending the study time of patients who were censored after the last recorded event, the following assumptions are made:

- Patients who were alive and censored at the end of the studies were likely to be ‘cured’ of ALL, such that the patient would also be alive and censored at the end of the longer follow-up period
- Patients who were censored prior to the last event were assumed to have been lost to follow-up, such that additional trial follow-up would not lead to further information on the timing of death (or re-censoring) in that individual patient.

This approach ensured that the KM curves remained consistent across the evidence sets. The net result of this adjustment is that the more mature evidence sets contain more information on the long-term survival of those alive and censored at the end of the current studies. An illustration of this approach is provided in Figure 2.

Figure 2: Illustration of adjustment to data set to account for additional study maturity



It is important to highlight that in practice, additional follow-up times in trials would likely result in changes to the KM curves and estimates of survival benefit. To predict these changes would require access to individual patient level data, so as to elicit the characteristics of patients who are censored at shorter follow-up and to then predict the unobserved event time for the censored patients, conditional on their characteristics. With the imputed data, it is not possible to identify the characteristics of those who are censored, and there is insufficient data to develop a prediction equation for the unobserved event times.

6.7 Finalised TPPs

6.7.1 Bridge to HSCT TPPs

Data on the evidence sets assumed for the clinical efficacy of CAR T-cell therapy as a bridge to HSCT is reported in Table 9 (overall survival), and

Table 10 (dichotomous endpoints, adverse events). The associated KM plots are reported in Figure 3.

Minimum data set

In terms of overall survival, CAR T-cell therapy was assumed to be associated with improved probabilities of survival at months 6 (66.8%) and 12 (51.8%) compared to standard of care therapy (32.0% at 6-months and 20.7% at 12-months). Treatment with CAR T-cell therapy was associated with a statistically significant improvement in the time to death, with a hazard ratio (HR) of 0.33 (95% confidence interval: 0.205 – 0.539).

In a restricted mean survival time analysis, treatment with CAR T-cell therapy was associated with a mean extension to life expectancy of 5.38 months (95% confidence interval: 3.18 – 7.60 months), versus standard of care therapy. The median follow-up in the minimum set was 11.3 months.

Intermediate data set

Given the consistency in the assumptions and the KM data assumed across the evidence sets, similar expected results were observed in the intermediate set as reported in the minimum set. However, evidence was now also assumed to be reported on the survival benefits up to 5-years, with 5-year landmark survival probabilities of 51.8% and 20.7% for CAR T and standard of care therapy assumed respectively. Treatment with CAR T-cell therapy was associated with a statistically significant improvement in the time to death, with a HR of 0.31 (95% CI: 0.19 – 0.50).

With increased data maturity compared to the minimum set, there was a greater trend towards non-proportional hazards. In the restricted mean survival time analysis, treatment CAR T-cell therapy was associated with a mean improvement in life expectancy of 22.06 months (95% CI: 12.87 – 31.25 months), versus standard of care therapy. The median follow-up in the intermediate data set was 53.6 months.

Mature data set

With increased precision and data maturity compared to the minimum set, treatment with CAR T-cell therapy was associated with a statistically significant improvement in the time to death, with a hazard ratio of 0.31 (95% CI: 0.22 – 0.43). In the restricted mean survival time analysis, treatment CAR T-cell therapy was associated with a mean improved in life expectancy of 22.06 months (95% CI: 15.56– 28.17 months), versus standard of care therapy. The increased sample size is reflected in a more precise estimate of the mean life expectancy in this evidence set compared to the intermediate set, evidenced by the tighter confidence intervals reported. The median follow-up in the mature data set was 53.6 months.

6.7.2 Curative intent TPPs

Data on the evidence sets assumed for the clinical efficacy of CAR T-cell therapy as a curative treatment option is reported in Table 11 (overall survival), and Table 12 (dichotomous endpoints, adverse events). The associated KM plots are reported in Figure 3.

Minimum data set

CAR T-cell therapy was assumed to be associated with improved probabilities of survival at months 6 (78.5%) and 12 (72.5%) compared to standard of care therapy (32.0% and 20.7% at 6 and 12-months).

Treatment with CAR T-cell therapy was associated with a statistically significant improvement in the time to death, with a hazard ratio (HR) of 0.20 (95% confidence interval: 0.11 – 0.37).

In a restricted mean survival time analysis, treatment with CAR T-cell therapy was associated with a mean extension to life expectancy of 10.47 months (95% CI: 7.59 – 13.34 months), versus standard of care therapy. The median follow-up in the minimum set was 11.3 months.

Intermediate data set

5-year landmark survival probabilities of 72.5% and 20.7% for CAR T and standard of care therapy were assumed respectively. Treatment with CAR T-cell therapy was associated with a statistically significant improvement in the time to death, with a HR of 0.18 (95% CI: 0.10 – 0.33).

In contrast to the 'bridge to HSCT' there was no apparent trend towards non-proportional hazards over time. In the restricted mean survival time analysis, CAR T-cell therapy was associated with a mean improvement in life expectancy of 32.94 months (95% CI: 24.38 – 41.43 months), versus standard of care therapy. The median follow-up in the intermediate data set was 53.6 months.

Mature data set

Treatment with CAR T-cell therapy was associated with a statistically significant improvement and more precise estimate hazard ratio compared to the minimum and intermediate evidence sets; HR=0.307 (95% CI: 0.12 – 0.27). In the restricted mean survival time analysis, treatment CAR T-cell therapy was associated with a similar mean to the intermediate set but with increased precision. The estimate of the improvement in life expectancy was 32.94 months (95% CI: 26.87– 38.93 months), versus standard of care therapy. The median follow-up in the mature data set was 53.6 months.

Table 9: Overall survival endpoint for <u>bridge to HSCT</u> datasets

Endpoint		**Minimum dataset**			**Intermediate dataset**			**Mature dataset**		
		Mean estimate and uncertainty		Comparative efficacy / all patients	Mean estimate and uncertainty		Comparative efficacy	Mean estimate and uncertainty		Comparative efficacy
		CAR T	SOC		CAR T	SOC		CAR T	SOC	
Sample size		63	61	124	63	61	124	126	122	248
Median time to censoring (follow-up; OS end-point)		-	-	11.3 (9.9 -14.1)	-	-	53.6 (14.1 – 54.2)	-	-	53.6 (52.9 – 54.2)
Overall survival	Landmark survival probability at 6-months	66.8% (52.4% - 77.8%)	32.0% (20.6% - 43.9%)	-	66.8% (52.4% - 77.8%)	32.0% (20.6% - 43.9%)	-	66.8% (57.0% - 74.9%)	32.0% (23.8% - 40.4%)	-
	Landmark survival probability at 12-months	51.6% (36.1% - 65.0%)	20.7% (11.4% - 32.1%)	-	51.6% (36.1% - 65.0%)	20.7% (11.4% - 32.1%)	-	51.6% (40.8% - 61.3%)	20.7% (13.8% - 28.7%)	-
	Landmark survival probability at 60-months	NA	NA	NA	51.6% (36.1% - 65.0%)	20.7% (11.4% - 32.1%)	-	51.6% (40.8% - 61.3%)	20.7% (13.8% - 28.7%)	-
	Hazard ratio, 95% confidence interval	-	-	**0.331 (0.203 – 0.539)**	-	-	**0.309 (0.190 – 0.503)**	-	-	**0.307 (0.218 – 0.434)**
	Test for proportionality (p-value< 0.05 indicates non-proportional hazards)	-	-	0.101	-	-	0.0368	-	-	0.0211
	Restricted mean survival time analysis (months)	11.37	5.98	5.388 (3.175 – 7.601)	33.02	10.96	22.06 (12.868 – 31.254)	33.02	10.96	22.06 (15.56 – 28.57)

Notes: NA – not applicable

Figure 3: Kaplan-Meier plots of time from randomisation to death for CAR T and standard of care therapy in the scenarios of "bridge to HSCT" (A, C) and "curative intent" (B, D), and with current trial follow-up (A,B) and mature trial follow-up (C,D)

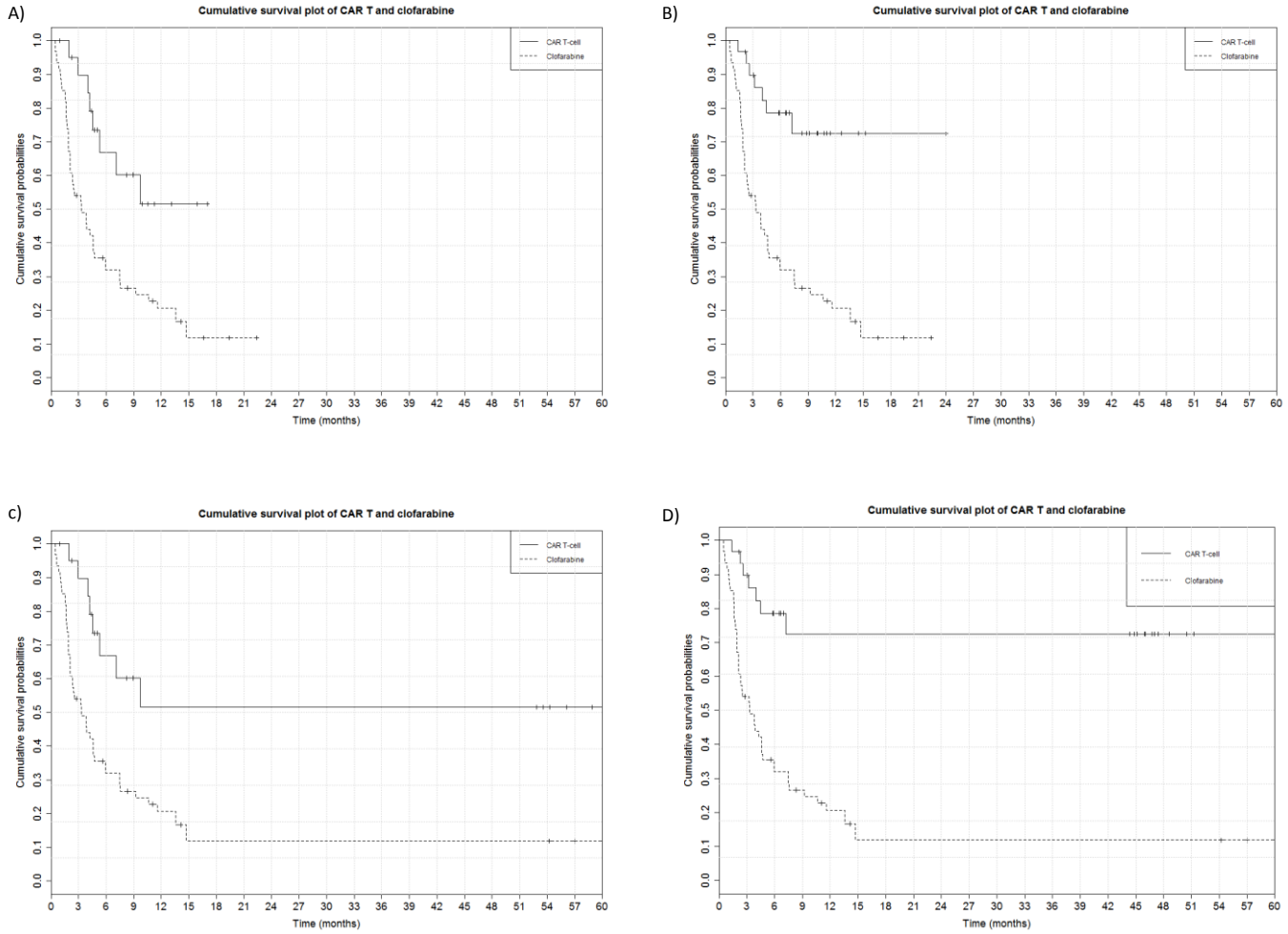


Table 10: Other efficacy and safety endpoints for bridge to HSCT datasets

Endpoint	Minimum and intermediate datasets			Mature dataset		
	Mean estimate and uncertainty		Odds ratio / all patients	Mean estimate and uncertainty		Comparative efficacy
	CAR T	SOC		CAR T	SOC	
Sample size	63	61	124	126	122	248
Complete remission	64.9% (52.6% ,76.5%)	12.5% (6.1% ,21.5%)	15.6 (5.5 ,34.1)	65.7% (57.7% 73.6%)	12.0% (6.9% ,18.1%)	15.4 (7.5 ,28.9)
MRD status	54.7% (42.8% ,67.%)	2.7% (.4% ,7.5%)	88.5 (14.3 ,394.2)	56.0% (47.6% 64.7%)	2.3% (0.5% ,5.4%)	82.7 (21.2 ,243.9)
Probability of HSCT	47.3% (35.6% ,59.6%)	16.4% (8.5% ,26.6%)	5.2 (2.2 ,11.2)	47.5% (39.1% 56.8%)	15.6% (10.1% ,22.3%)	5.2 (2.9 ,9.3)
Cytokine release syndrome	28.9% (19.0% ,40.7%)	1.5% (0.0% ,5.3%)	-	28.2% (20.8% 36.1%)	0.9% (.% ,3.3%)	-
Encephalopathy	6.3% (2.0% ,13.3%)	1.6% (0.0% ,5.8%)		5.7% (2.4% 10.7%)	0.8% (0% ,2.9%)	
Hypotension	22.7% (13.7% ,34.2%)	19.2% (10.7% ,29.6%)		22.5% (15.8% ,30.4%)	18.6% (12.4% ,25.8%)	
Febrile neutropenia	33.8% (22.8% ,45.0%)	48.6% (36.7% ,61.0%)		33.2% (24.8% ,41.6%)	49.1% (40.5% ,58.2%)	
neutropenia/ neutrophil count decreased	87.7% (78.8% ,94.6%)	16.2% (8.0% ,25.9%)		88.4% (82.2% ,93.2%)	15.5% (9.8% ,22.1%)	
Anemia	67.4% (55.9% ,78.0%)	1.6% (0.0% ,5.8%)		67.8% (59.8% ,75.4%)	0.8% (0% ,2.9%)	
Thrombocytopenia/ platelet count decreased	52.9% (41.2% ,64.7%)	1.6% (0.1% ,5.7%)		52.7% (44.3% ,61.7%)	0.8% (0% ,2.9%)	
Leukopenia/ white cell decreased	88.% (79.% ,94.5%)	1.6% (0.0% ,6.2%)		88.4% (82.% ,93.4%)	0.8% (0% ,2.7%)	

Hypokalemia	46.9% (34.6% ,58.9%)	1.8% (0.0% ,6.1%)		47.3% (39.3% ,55.5%)	0.8% (0% ,3.0%)	
Hypophosphatemia	42.% (30.6% ,54.1%)	1.6% (0.0% ,5.9%)		42.1% (33.7% ,50.7%)	0.8% (0% ,3.0%)	

Table 11: Overall survival endpoint for curative intent data sets

Endpoint		Minimum dataset			Intermediate dataset			Mature dataset		
		Mean estimate and uncertainty		Comparative efficacy / total patients	Mean estimate and uncertainty		Comparative efficacy / total patients	Mean estimate and uncertainty		Comparative efficacy / total patients
		CAR T	SOC		CAR T	SOC		CAR T	SOC	
Sample size		60	61	121	60	61	121	120	122	242
Median time to censoring (follow-up for OS)		-	-	10.03 (8.33 – 11.41)	-	-	45.2 (14.1 – 46.8)			45.2 (44.3-46.0)
Overall survival	Probability at 6-months	0.785 (0.653 - 0.872)	0.320 (0.2061 - 0.439)	-	0.785 (0.653 - 0.872)	0.320 (0.2061 - 0.439)	-	0.785 (0.697 - 0.851)	0.320 (0.2379 - 0.404)	-
	Probability at 12-months	0.725 (0.573 - 0.831)	0.207 (0.1135 - 0.321)	-	0.725 (0.573 - 0.831)	0.207 (0.1135 - 0.321)	-	0.725 (0.622 - 0.804)	0.207 (0.1382 - 0.287)	-
	Probability at 60-months	NA	NA	NA	0.725 (0.573 - 0.831)	0.207 (0.1135 - 0.321)	-	0.725 (0.622 - 0.804)	0.207 (0.1382 - 0.287)	-
	Hazard ratio, 95% confidence interval	-	-	0.204 (0.113 – 0.370)	-	-	0.180 (0.099 – 0.327)	-	-	0.179 (0.117 – 0.272)
	Test for proportionality (p-value < 0.05 indicates non-proportional hazards)	-	-	0.699	-	-	0.784	-	-	0.678
	Restricted mean survival time analysis (months)	17.04	6.57	10.47 (7.59 – 13.34)	43.86	10.96	32.94 (24.38– 41.43)	43.86	10.96	32.94 (26.87 – 38.93)

Notes: NA – not applicable

Table 12: Other efficacy and safety endpoints for curative intent datasets

Endpoint	Minimum and intermediate datasets			Mature dataset		
	Mean estimate and uncertainty		Odds ratio / all patients	Mean estimate and uncertainty		Comparative efficacy
	CAR T	SOC		CAR T	SOC	
Sample size	60	61	121	120	122	242
Complete remission	90.0% (81.3%,96.2%)	11.5% (4.7%,20.6%)	97.25 (25.9, 284.0)	90.0% (84.0%,94.7%)	11.5% (6.5%,17.8%)	81.42 (33.1, 177.1)
MRD status	73.4% (61.5%, 83.7%)	1.6% (0.0%,6.0%)	1719.0 (39.26, 7169.0)	73.4% (65.2%, 80.8%)	1.6% (0.0%, 4.5%)	344.4 (54.42, 1462.0)
Probability of HSCT	10.0% (3.8%, 18.7%)	14.8% (7.1%, 24.7%)	0.738 (0.193, 1.913)	10.0% (5.3%, 16.0%)	14.8% (9.0%,21.6%)	0.686 (0.283, 1.379)
Cytokine release syndrome	27.0% (16.6%,38.9%)	1.5% (0.0% ,5.3%)		27.0% (19.5%, 35.3%)	0.9% (0.0% ,3.3%)	-
Encephalopathy	20.0% (11.0%, 31.0%	1.6% (0.0% ,5.8%)		20.0% (13.4%, 27.6%)	0.8% (0.0% ,2.9%)	
Hypotension	27.1% (16.7%, 38.9%)	19.2% (10.7% ,29.6%)		27.0% (19.5%, 35.2%)	18.6% (12.4% ,25.8%)	
Febrile neutropenia	73.0% (61.2%, 83.3%)	48.6% (36.7% ,61.0%)		73.0% (64.8%, 80.5%)	49.1% (40.5% ,58.2%)	
neutropenia/ neutrophil count decreased	1.6% (0.0% ,5.8%)	16.2% (8.0% ,25.9%)		0.8% (0% ,2.9%)	15.5% (9.8% ,22.1%)	
Anemia	1.6% (0.0% ,5.8%)	1.6% (0.0% ,5.8%)		0.8% (0% ,2.7%)	0.8% (0% ,2.9%)	
Thrombocytopenia/ platelet count decreased	1.6% (0.0% ,5.8%)	1.6% (0.1% ,5.7%)		0.8% (0% ,2.9%)	0.8% (0% ,2.9%)	
Leukopenia/ white cell decreased	1.6% (0.0% ,5.8%)	1.6% (0.0% ,6.2%)		0.8% (0% ,2.9%)	0.8% (0% ,2.7%)	
Hypokalemia	1.6%	1.5%		0.8%	0.8%	

	(0.0% ,6.2%)	(0.0% ,5.3%)		(0% ,3.0%)	(0% ,3.0%)	
Hypophosphatemia	1.6% (0.0% ,5.8%)	1.6% (0.0% ,5.8%)		0.8% (0% ,3.0%)	0.8% (0% ,3.0%)	

A summary of the six evidence sets across the 2 separate TPPs is provided in Table 13.

Table 13: Summary of key attributes of the six evidence sets

TPP: Bridge to HSCT			
Attribute	**Evidence set**		
	Minimum	**Intermediate**	**Mature**
Median time to censoring (follow-up)	11.3 months	53.6 months	53.6 months
Overall Survival: Hazard ratio (95% CI) Differences in restricted mean survival times (95% CI)	0.331 (0.203 – 0.539) 5.4 months (3.2 – 7.6)	0.309 (0.190 – 0.503) 22.1 months (12.9 – 31.3)	0.307 (0.218 – 0.434) 22.1 months (15.6 – 28.6)
Complete remission (95% CI) CAR T Standard of care - clofarabine	64.9% (52.6% - 76.5%) 12.5% (6.1% - 21.5%)	64.9% (52.6% - 76.5%) 12.5% (6.1% - 21.5%)	65.7% (57.7% - 73.6%) 12.0% (6.9% - 18.1%)
MRD negative (95% CI) CAR T Standard of care - clofarabine	54.7% (42.8% - 67.7%) 2.7% (0.4% - 7.5%)	54.7% (42.8% - 67.7%) 2.7% (0.4% - 7.5%)	56.0% (47.6% - 64.7%) 2.3% (0.5% - 5.4%)
CRS (95% CI) CAR T Standard of care - clofarabine	28.9% (19.0% - 40.7%) 1.5% (0.0%, 5.3%)	28.9% (19.0% - 40.7%) 1.5% (0.0%, 5.3%)	28.2% (20.8% - 36.1%) 0.9% (0.0% - 3.3%)
Febrile neutropenia (95% CI) CAR T Standard of care - clofarabine	33.8% (22.8% - 45.0%) 48.6% (36.7% - 61.0%)	33.8% (22.8% - 45.0%) 48.6% (36.7% - 61.0%)	33.2% (24.8% - 41.6%) 49.1% (40.5% - 58.2%)
TPP: Curative intent			
Median time to censoring (follow-up)	10.03 months	45.2 months	45.2 months
Overall Survival: Hazard ratio (95% confidence interval) Differences in restricted mean survival times (95% confidence interval)	0.204 (0.113 – 0.370) 10.5 months (7.6 – 13.3)	0.180 (0.099 – 0.327) 32.9 months (24.4 – 41.4)	0.179 (0.117 – 0.272) 32.9 months (26.9 – 38.9)
Complete remission (95% CI) CAR T Standard of care - clofarabine	90.0% (81.3% - 96.2%) 11.5% (4.7% - 20.6%)	90.0% (81.3% - 96.2%) 11.5% (4.7% - 20.6%)	90.0% (84.0% - 94.7%) 11.5% (6.5% - 17.8%)
MRD negative (95% CI) CAR T Standard of care - clofarabine	73.4% (61.5% - 83.7%) 1.6% (0.0% - 6.0%)	73.4% (61.5% - 83.7%) 1.6% (0.0% - 6.0%)	73.4% (65.2% - 80.8%) 1.6% (0.0%-4.5%)
CRS rate (95% CI) CAR T Standard of care - clofarabine	27.0% (16.6% - 38.9%) 1.5% (0.0% - 5.3%)	27.0% (16.6% - 38.9%) 1.5% (0.0% - 5.3%)	27.0% (19.5% - 35.3%) 0.9% (0.0% - 3.3%)
Febrile Neutropenia (95% CI)			

CAR T	73.0% (61.2% - 83.3%)	73.0% (61.2% - 83.3%)	73.0% (64.8% - 80.5%)
Standard of care - clofarabine	48.6% (36.7% - 61.0%)	48.6% (36.7% - 61.0%)	49.1% (40.5% - 58.2%)

7 Review of cost-effectiveness evidence for CAR T-cell therapy and other interventions for ALL

7.1.1 Methods

No previously published studies on the potential cost-effectiveness of CAR T-cell therapy for ALL were identified in our searches. To inform the conceptualisation and development of the economic model, a separate review of published studies evaluating the cost-effectiveness of other treatments for ALL was conducted. The primary aim of the review was to inform key structural assumptions and potential parameter sources required for the model. Hence, the review focussed on the main methodological approaches taken in the studies identified, rather than the specific results reported.

A two–part approach was taken, consisting of a systematic review and a more pragmatic search. Details of the search strategy employed to inform the systematic review are available in Appendix 5, Table 48. The pragmatic search searched for any publicly available reports considering the cost-effectiveness of any intervention in ALL using Google and Google Scholar, in addition, the relevant websites for NICE and the All Wales Medicine Strategy Group's (AWMSG) were searched to identify previous appraisals for ALL.

7.1.2 Results

The systematic search identified 489 records, 11 of which were deemed potentially relevant after a review of their titles and abstracts. However, after obtaining the full articles, none of these studies were found to be full economic evaluations and hence were not subsequently considered within the model conceptualisation stage. The pragmatic search using Google and Google Scholar found two papers deemed relevant to the primary aim of the review.[134, 135]

Costa conducted a cost-effectiveness evaluation of unrelated stem cell transplantation for adults with acute leukemia (ALL and AML) structured around a 20-year Markov model.[134] The study concluded that the two forms of transplantation considered (cord blood and bone marrow/peripheral blood stem cells) were cost-effective when compared to no-transplantation. The study found that despite the high initial cost and short-term mortality associated with the transplantation procedures, the resulting life-year gains achieved by surviving patients were significant.

Lis[135] considered the cost-effectiveness of clofarabine combined with chemotherapy in children and adolescents with ALL who have failed at least two previous therapies, compared to nelarabine and FLAG-IDA, though the use of a lifetime Markov model. After the initial treatments, a proportion of

patients were assumed to subsequently receive HSCT, this proportion varied given response to initial treatment (complete, partial, complete without platelet recovery, or no response) and the treatment arm. A patient who survived for two years post-HSCT was assumed to be cured of ALL, no cure was possible without HSCT. The authors found clofarabine to be cost-effective when compared to both comparators. The result was driven by the success of a therapy in achieving a bridge to HSCT, and thus a potential cure. As clofarabine was associated with a greater proportion of patients experiencing an initial complete response, it had the greatest proportion of HSCTs and thus cured patients.

The search of NICE and AWMSG appraisals found that the only appraisal by NICE in ALL (dasatinib, ID386)[136] was discontinued in 2008 due to a low number of patients anticipated to be treated. By contrast, the AWMSG provides details of 5 separate appraisals in ALL, although 1 of these (imatinib, no.2014) did not receive a formal submission by the manufacturer. Of the remaining AWMSG appraisals, only the final appraisal recommendations (FARs) are made publicly available, limiting the detail available on the evaluative approaches. Only two of the appraisals (clofarabine and nelarabine) provided sufficient detail to review.

Clofarabine[124] was recommended by the AWMSG for children and adolescents with ALL who are relapsed or refractory after at least two prior regimens and where no other treatment is anticipated. Within the FAR, an important restriction was placed on the recommendation such that clofarabine should only be given to patients in whom there is an intention to proceed to HSCT. This recommendation was based on the findings that clofarabine did not appear cost-effective for patients who did not subsequently receive HSCT. In the submission, clofarabine was compared to palliative care alone. Palliative care was assumed to be associated with very short median survival (9 to 10 weeks) based on historic control data.

Although limited details of the modelling approach are reported, it is evident that the primary structural driver within the model is the bridging role of clofarabine to HSCT, with potentially significant gains life year assumed for patients who subsequently receive HSCT. The manufacturer assumed that the success of HSCT in achieving long term remission (and cure) was driven by the achievement of remission (complete, with platelet involvement or partial) at the time of transplantation. Hence, improved rates of remission achieved with clofarabine compared to palliative care directly equate to long term survival. The model submitted assumed that patients who received HSCT and survived for one year were cured, returning to the mortality risks and utilities of the general population.

Nelarabine[137] was recommended by the AWMSG for the treatment of T-ALL and T-LBL whose disease has not responded to, or has relapsed, following treatment with at least two chemotherapy regimens. Best supportive care was used as the main comparator and clofarabine was considered in a separate scenario based on indirect comparisons. In common with the restriction previously applied within their recommendations for clofarabine in ALL, the AWMSG also restricted treatment to patients where there is an intention to proceed to HSCT. This restriction was based on a similar finding that cost-effectiveness of nelarabine was closely related to the assumed increase in the proportion of patients subsequently receiving HSCT (and their related long-term health gains). The base-case analysis presented survival based on within-trial estimates with no extrapolation conducted. This was considered to be an extremely conservative estimate. Separate scenarios were presented considering the long-term survival of post-HSCT patients, and found to have a major impact on the result. The base-case ICER of £102,281 per QALY gained was subsequently reduced to £51,169 if post-HSCT survival was assumed to be 2 years and to £25,523 if normal life expectancy was assumed in patients who survived more than one year (i.e. cure at one year).

7.1.2.1 Implications for model conceptualisation

The systematic and pragmatic searches highlighted a number of potential implications for our evaluation. Within existing studies, it is clear that the main benefit of existing treatments has been related to their ability to provide a 'bridge' to HSCT. The primary factor determining cost-effectiveness in the reviewed literature was the increased likelihood of receiving HSCT with a new treatment and the associated assumptions made regarding subsequent health gains associated with transplantation. Only limited survival gain was attributed to patients who did not subsequently receive HSCT, such that no treatment reviewed appeared cost-effective as a palliative option.

The key structural assumptions employed within these studies are the potentially curative effect of HSCT and the short life expectancy assumed for the comparator treatments (best supportive care/palliative treatment alone) derived from historic controls. The majority of studies assumed a 'cure point' associated with HSCT, although the timing differenced across studies. The 'cure point' was assumed to represent the time at which patients are assumed to no longer be at risk of disease relapse. The study by Costa[134] assumed that 5 years post transplantation the patient will be free of any procedural mortality risk or any risk of disease recurrence. In Lis[135] and nelarabine[137] this cure point is assumed to be 2 years after HSCT, while the clofarabine submission[124] assumes this at 1 year after HSCT.

The studies also differed in the assumptions made concerning subsequent survival after the 'cure point'. Costa acknowledged that long-term ALL survivors are likely to be subject to significant comorbidities over their remaining lifetime despite being leukaemia free.[134] To account for the impact of comorbidities, an assumption was made that the long term survival of ALL patients would be 50% less than the general population. The authors acknowledge this was an arbitrary adjustment due to the lack of data on the long term mortality rate in long-term survivors of ALL reported at the time. In contrast, both of the studies reported by Lis[135] and the clofarabine submission[124] effectively assumed no additional comorbidities (i.e. beyond those experienced by the general population) beyond the 'cure point'. Hence, patients were subsequently assumed to return to the age-adjusted mortality risk and utility of the general population. The AWMSG raised concerns that, not only was this assumption insufficiently justified but that the model was very sensitive to changes in the long-term survival probability.

In the absence of RCT data, each model incorporated historical control data as the basis to inform outcomes associated with the comparator (best supportive care/palliative care and clofarabine within a scenario for the submission for nelarabine). However, insufficient limited details were reported regarding the source of the historic control data used, whether attempts were made to identify possible biases or to formally account for potential confounding.

The existing cost-effectiveness literature is limited in ALL. No completed NICE appraisals of licensed treatments for ALL were identified. Furthermore, of the studies published, none were reported in sufficient detail to provide a suitable basis for informing the exemplar application. In the absence of previous NICE appraisals or sufficient reporting within existing publications, the development of a de-novo model to inform the exemplar application was considered necessary. Full details of this are reported in the next section.

8 The exemplar economic model

8.1 Overview

There are several distinct issues and challenges for the modelling of costs and outcomes that arise from the separate TPPs:

- In the *"bridge to HSCT"* TPP, the primary health benefits of treatment are gained by enabling more patients to successfully undergo HSCT; an established intervention that has known curative potential. For economic modelling purposes, it may therefore be desirable to introduce a structural link between HSCT and overall treatment benefit (i.e. survival) in the model. The introduction of a link between a potential established surrogate outcome or process and final health benefits also enables the use of evidence external to the CAR T-cell evidence sets (i.e. survival post-HSCT). This structural link may also provide decision makers with greater confidence surrounding the modelled health benefits of treatment on survival, given that model projections would depend largely on the established benefits of HSCT. In terms of decision uncertainty, this approach would also mean that the uncertainty surrounding the cost-effectiveness of CAR T-cell is partly determined by the maturity and sample size from the evidence sets, and partly by the maturity, sample size, and acceptability of external evidence obtained from other sources.
- In the *"curative intent"* TPP, the case for introducing a structural link between final health benefits and a surrogate outcome or process such as HSCT is more limited than in the *"bridge to HSCT"* case, given that it is primarily CAR T-cell therapy itself that is expected to provide the curative benefits. In this context, it may be more appropriate to model long-term outcomes via the direct extrapolation of event-free and overall survival data from the CAR T–cell trial evidence sets, as opposed to modelling long-term outcomes through a separate surrogate process. In this case, the decision uncertainty surrounding the cost-effectiveness of treatment would be solely determined by the maturity and sample size of data from the evidence sets.

8.1.1 Patient population

In this evaluation, we evaluate the cost-effectiveness of CAR T-cell therapy in the treatment of children and young adults with two or more relapses or refractory ALL. The baseline demographic characteristics of this patient group is summarised in Table 14.

Table 14: Baseline characteristics of patients

Scenario	Characteristic	Parameter	Source / application

Bridge to HSCT	Mean baseline age	14.0 years	Lee et al[117]
	% Female	33%	
Curative intent	Mean baseline age	14.0 years	Maude et al[118]
	% Female	40%	

8.1.2 Comparator

The comparator treatment to CAR T-cell therapy was defined as standard of care. In the base case, the standard of care treatment was assumed to be clofarabine (Evoltra®). The mean cost for a course of clofarabine treatment is approximately £43,200 per patient.

As part of a separate sensitivity analysis, the standard of care treatment was assumed to be FLAG-IDA. The mean cost for a course of FLAG-IDA treatment is approximately £3,803 per patient.

8.2 Model development

The approaches to modelling the cost-effectiveness of CAR T-cell therapy varied between the separate scenarios. Therefore, two de novo decision models were developed and used to assess the cost-effectiveness of CAR T-cell therapy across the two scenarios:

1) **Bridge to HSCT model** – based on a landmark responder model that comprises two related decision models;
    1. a short-term decision-tree to predict the remission and transplant status of the population in the immediate period following CAR T-cell or comparator therapy, and
    2. a series of partitioned survival (or area under the curve) models to predict the longer-term survival of patients conditional on remission and transplant status.
2) **Curative intent model** – based on a simple three state (alive and event-free, alive post-event, dead) partitioned survival model

The two models share a number of common features, which are outlined in Table 15.

Table 15: Key common features of de novo economic models

Factor	Chosen values	Justification
Time horizon	Lifetime horizon (up to a maximum age of 100 years)	Necessary to capture the potential life time impacts of short-term and potentially ongoing mortality benefit.
Cycle length	One month	Remission status is determined at day 28
Mid or half-cycle correction	Mid-cycle correction employed	To guard against over or under predicting state occupancy in the model
Measure of health effects	QALYs	In accordance with the current NICE reference case for cost-effectiveness. Necessary to quality weight short-term and potentially ongoing mortality benefits and associated adverse events.
Discounting	3.5% for costs and health effects, over the lifetime horizon	In accordance with the current NICE reference case. Alternative discounting rates explored using sensitivity analysis.
Perspective	NHS/PSS	In accordance with the current NICE reference case

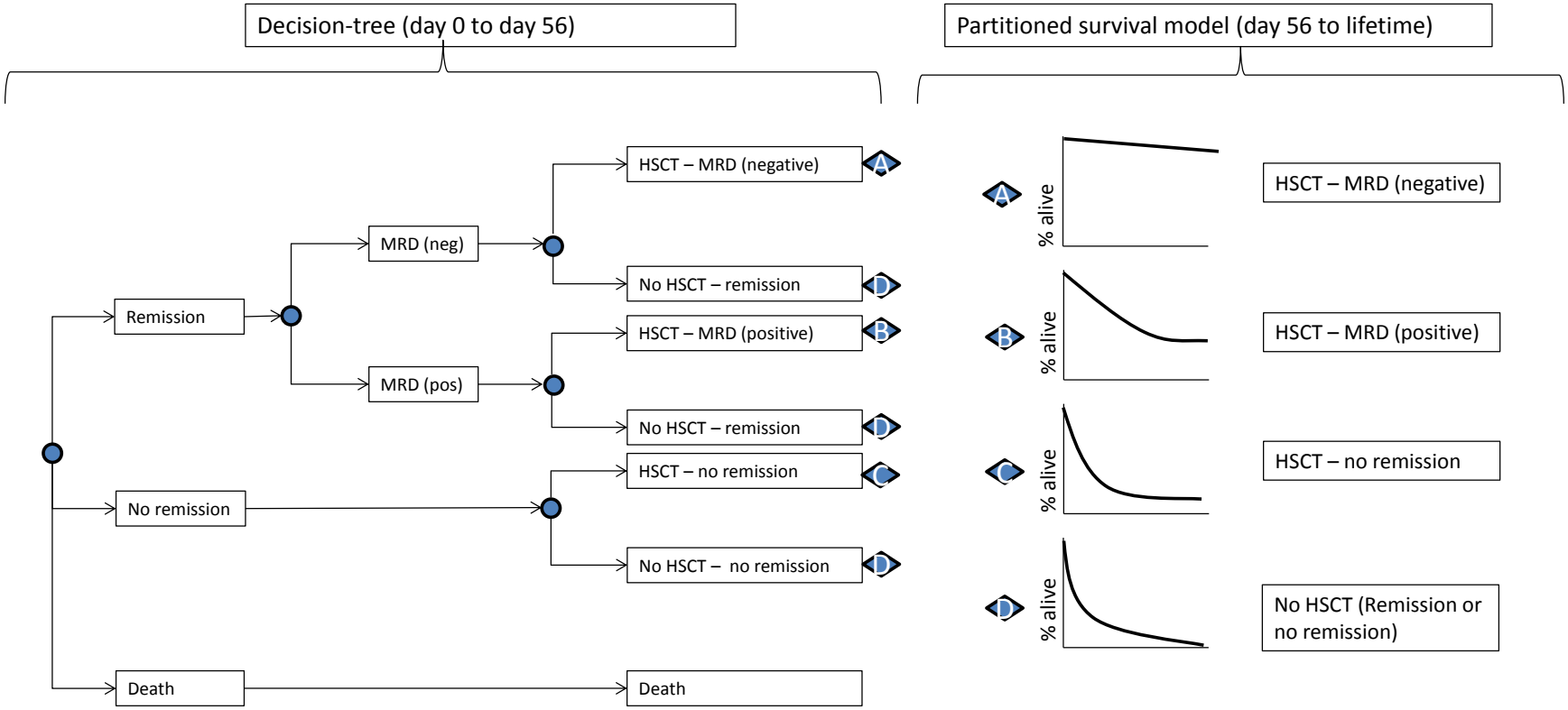
Further details specific to each of the two model structures are reported in the following sections.

8.2.1 Bridge to HSCT scenario

8.2.1.1 Key structural assumptions

The bridge to HSCT model consists of a decision tree model (day 0 to 56) and a series of partitioned survival models (day 56 to lifetime) that when combined, provide an estimate of the lifetime costs and effectiveness of treatment in ALL. An illustration of the structure of the model is provided in Figure 4.

Figure 4: Illustration of bridge to HSCT model



The short-term decision tree component of the model comprises three chance nodes that represent a series of clinically relevant events that may occur during the first 56 days (two months) of treatment:

- Remission status at day 28 (remission, no-remission or death)
- MRD status at day 28 (negative or positive)
- Transplantation status at day 56 (HSCT or no-HSCT)

These events are considered to be prognostic of the duration and quality of life of patients with ALL, and were therefore included in the model to link short-term measures of trial efficacy to longer term health outcomes. The three nodes of the decision tree are sequenced in the order of remission, MRD status and transplantation status.

At the first chance node, the hypothetical cohort are distributed across three states; remission, no remission, or death. In the model, remission is defined using the criteria applied in the CAR T-cell and clofarabine clinical trials.[117, 131] Complete remission is defined as less than 5% marrow blasts by flow cytometry, an absence of circulating blasts, and no extramedullary sites of disease with absolute neutrophil count of 1000 per μL or more and platelets counts of 100 000 per μL or more. In accordance with both Lee et al and Jeha et al studies, remission status is determined at day 28 (month one) of the simulation.

At the second chance node, patients with remission are re-assigned to one of two states; remission and MRD negative, or remission and MRD positive. MRD negative status is defined as less than 0.01% marrow blasts. MRD positive status is determined by marrow blasts of between 0.01% and 5% (at >5% patients are no longer in remission).

At the third and final node (day 56), all patients are assigned to states corresponding to the use of HSCT (HSCT versus no HSCT). The final determination of health status (remission – MRD – HSCT) was assumed to occur at day 56 (month two) of the simulation. This time period was chosen based on the mean time from CAR T therapy to HSCT, estimated from data reported in Lee et al (mean 54 days, 95% confidence interval: 45-77 days).

At the end of the decision tree phase, the cohort is assigned to six mutually exclusive states (presented in order of best prognosis):

- HSCT - Remission and MRD negative
- HSCT - Remission and MRD positive

- HSCT - No remission
- No HSCT - Remission
- No HSCT - No remission
- Death

After day 56, the long-term survival of the cohort is modelled through a series of related partitioned survival models (Figure 4) that are used to model the long-term outcomes of treatment (day 56 to lifetime). The model includes four distinct partitioned survival models that are used to evaluate survival in the following groups:

- HSCT - Remission and MRD negative
- HSCT - Remission and MRD positive
- HSCT - no remission
- No HSCT

HSCT recipients with MRD negative status prior to transplantation are assumed to have the best prognosis in terms of long-term survival. Increasing levels of marrow blasts is assumed to be associated with a lower probability of long-term survival, such that HSCT recipients with MRD positive status have (on average) a worse survival prognosis than MRD negative patients. HSCT recipients who failed to achieve remission prior to transplantation were assumed to have the poorest prognosis of all HSCT patients.

For patients who did not receive HSCT, the probabilities of overall survival were significantly lower than for HSCT patients. It was assumed that complete remission was not associated with improved probabilities of survival in non-HSCT patients. This assumption was made on the basis that in the bridging scenario it is through HSCT (and not remission in the absence of HSCT) that meaningful gains in survival can be achieved. The impact of this assumption on the results of the evaluation was tested in the one-way sensitivity analyses, where it was assumed that non-remission non-HSCT patients had an inferior survival prognosis to remission non-HSCT patients.

At year 5 of the simulation, those who were alive were subsequently assumed to be long-term survivors of ALL. From this point forward, the cohort was considered to be effectively ‘cured’ of ALL, and experienced the mortality risk profile consistent with a long-term survivor of ALL. The mortality risks after year 5 were therefore modelled based on general population age- and gender-adjusted all-cause risk of mortality adjusted for excess morbidity and mortality reported in cohorts of long-term survivors of ALL. The approach is more formally described in Section 8.2.

The model also included treatment-related adverse events. These include events such as cytokine release syndrome, encephalophy, hypotension, febrile neutropenia, neutropenia, anaemia, thrombocytopenia, leukopenia, hypokalemia, and hypophosphatemia. The costs and consequences of these events were assumed to occur at the start of the evaluation. Since prolonged B-cell aplasia did not occur in the Lee study, the costs and consequences of this were not included within the bridge to HSCT scenario. The key structural assumptions applied in the model are outlined in Table 16.

Table 16: Summary of key modelling assumptions for Bridge to HSCT responder model

Input	Assumption
Surrogate relationship between MRD status and HSCT	A lower marrow blast status prior to transplantation (as captured through MRD status) is associated with a higher probability of experiencing a sustained remission and long-term survival benefits in ALL
HSCT	All HSCT events were assumed to occur at day 56 of the simulated time horizon. No further HSCT events were permitted after this point
Survival during the first 5-years of the evaluation time horizon	Survival post-HSCT was modelled based on constant transition probability. There is no difference in survival between remission non-HSCT and non-remission non-HSCT patients.
Survival after the first 5-years of the evaluation time horizon	All patients alive at 5-years post-HSCT are considered to be long-term survivors of the disease. Long-term survivors of ALL experience excess morbidity and mortality compared to the general population.
Treatment / re-treatment	In the base case, it was assumed each patient would receive a single full course of therapy. Re-treatment with CAR T or standard of care therapy was not permitted in the base case, but was considered in the sensitivity analysis.
Treatment effect	CAR T-cell therapy improves the probability of remission, the probability of MRD negative status, and the probability of successful HSCT compared to standard of care therapy The clinical parameter estimates used to inform the models, and TPPs can be generalised to the UK NHS
Patient follow-up	After HSCT, patients receive ongoing care and rehabilitation up to 2-years post-HSCT Patients who do not receive HSCT are assumed to require hospitalisation prior to death
Adverse events	Treatment-related adverse events were considered in the evaluation, and included events such as cytokine release syndrome, whose incidence is expected to increase with the use of CAR T-cell therapy. The costs and health consequences of adverse events were accrued at the start of the evaluation
Health-related quality of life	Patients who achieve remission status are assigned a higher utility weight than patients who do not achieve remission Transplantation is associated with a one-off decrement to health-related quality of life.

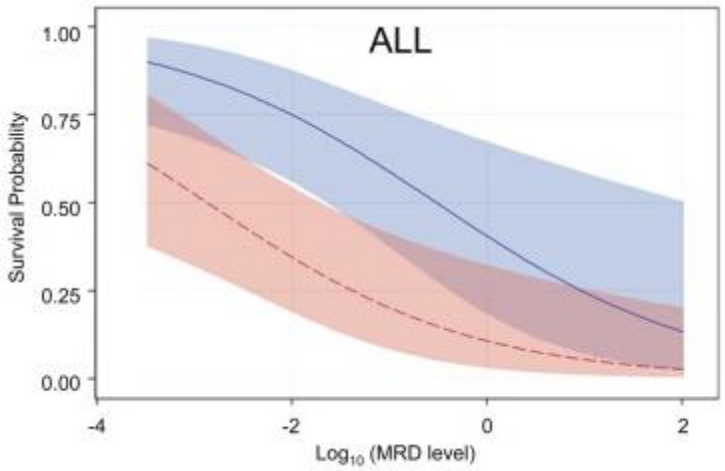
Clinical justification for structure of model

The conceptual structure of the "bridge to HSCT" model is based on an assumed relationship between HSCT use and final clinical benefits, and the assumption that the effectiveness of HSCT is dependent on MRD status prior to transplantation.

Allogeneic HSCT is a potentially curative treatment option in patients with ALL. However, the long-term benefits of HSCT are uncertain, with some patients experiencing long-term benefits, including the effective cure/suppression of ALL, and other patients experiencing relapse and/or mortality shortly after transplantation. In this evaluation, survival benefits are established through remission and MRD status prior to HSCT.

Several studies have investigated the relationship between remission/MRD status prior to HSCT, and the long-term outcomes of HSCT therapy.[70, 71, 138] These studies have shown, to varying degrees, that MRD status prior to HSCT appears an important prognostic determinant of long-term relapse-free and overall survival, with MRD negative (<0.01% marrow blasts) patients experiencing superior survival compared to MRD positive (>0.01% to 5% marrow blast) patients, including within studies of children with relapsed ALL. This relationship is shown graphically in Figure 5.

Figure 5: Post-HSCT survival probability based on level of MRD for patients with ALL (obtained from Leung et al)



In Figure 5, the survival probability 5 years after HSCT is plotted against pre-HSCT MRD level, expressed as a continuous measure and plotted on a logarithmic scale (base 10). The red dashed lines show the relationship observed in patients treated pre-2002, and the blue solid line shows the relationship post-2002. The shaded regions represent the confidence limits for the relationship. In both pre and post-2002 periods, there is a consistent association between MRD level and 5-year survival probability, such that increasing MRD level (% bone marrow blasts) was associated with decreasing probabilities of 5-year survival. These data support the assumption of a continuous relationship between MRD level prior to HSCT, and 5-year survival probability.

For patients who do not receive HSCT, the long-term outcomes of treatment are generally poor. In the study by Von Stackelberg et al,[114] the median survival in refractory patients who failed to respond to induction therapy and went on to receive palliative care was 89 days (3.17 months).

In the model, it was assumed that all patients who did not receive HCST (including remission and non-remission patients), went on to receive palliative care, having exhausted all treatment strategies that may be curative. Remission status was not considered to be prognostic of survival in the non-HSCT population, such that non-HSCT patients who achieved remission were assumed to be at the same risk of mortality as non-HSCT patients who failed to achieve remission. However, as discussed later in the report, all patients with remission were assigned an improved health utility compared to those who failed to achieve remission. These benefits were, however, assumed to not extend to improved life expectancy.

In previous economic evaluations in ALL (reviewed in Section 7), it had been assumed that survivors of ALL experience the same mortality risk profile as the general population. This assumption implies that there is no excess mortality or morbidity risk associated with their previous illness. This assumption is not supported by the published literature, which generally report excess mortality and morbidity amongst the long-term ALL survivor population, when compared to match-adjusted individuals without ALL (i.e. siblings).[139, 140] In the model, the risk of mortality assigned to survivors of ALL was set equal to the general population background all-cause mortality risk profile, with an adjustment for an increased mortality risk amongst survivors of ALL.

The point at which patients were assumed to be long-term survivors of ALL (5-years) was based on the definition used in a number of published studies reporting long-term survival data in ALL. None of these studies provide explicit rationale for selecting 5-years as the cure point, and to our knowledge, there appears to have been no published attempts to empirically justify the widespread use of the 5-year cure

point. However, across a number of studies, the KM curves for post-HSCT survival appear to stabilise within the 5-year time frame, such that the curve becomes flat and the incidence of death reduces to near zero.

8.2.1.2 Efficacy parameter estimates

In the decision-tree component of the model, the data for the remission, MRD, and HSCT status of the modelled cohort were derived from the separate evidence sets reported in 6.7.1. The key assumptions required to generate the estimates for the evidence sets are outlined in Table 17.

Table 17: Parameter estimates for bridge to HSCT scenario

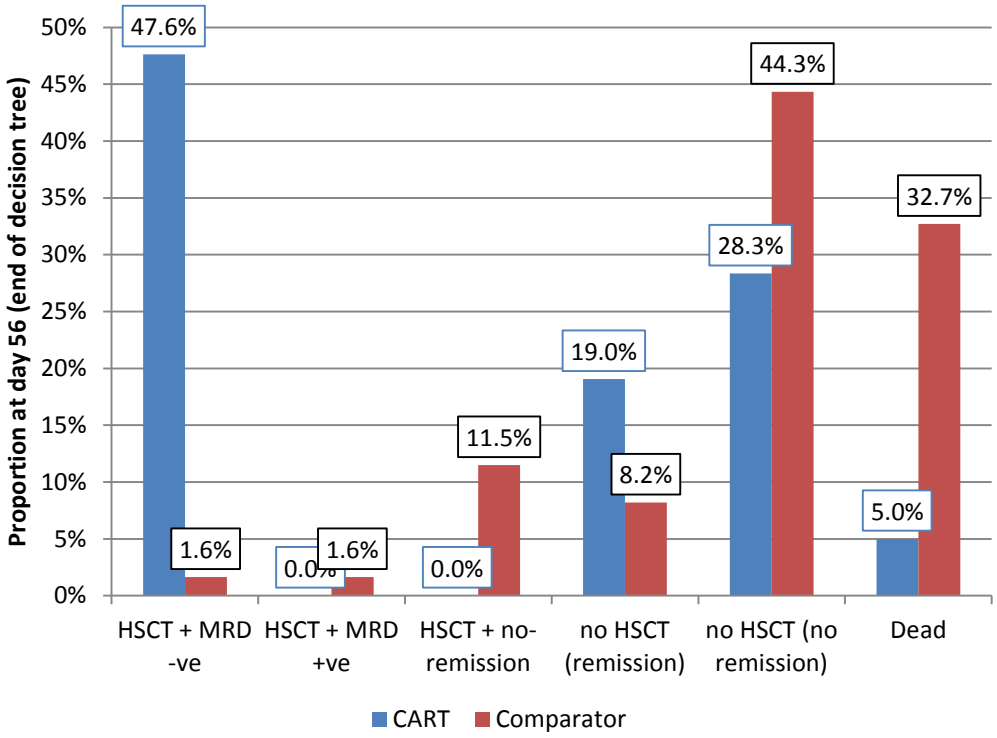
Node parameter 1	Node parameter 2	Node parameter 3	Estimate	Probabilistic distribution	Minimum (n)	Intermediate (n)	Mature (n)	Key assumptions
CAR T-cell								
Remission No Remission Death (day 0 -56)			66.7% 28.3% 5.0%	Dirichlet distribution	63	63	126	Remission probability based on 14 of 21 patients achieving remission in Lee et al. By day 56, 1 of 21 patients had died (5%). All remaining patients (28.3%) were assumed to have non-remission
Remission	MRD negative MRD positive		85.7% 14.3%	Beta distribution				12 patients in Lee et al had a MRD negative status. In total, 14 patients were in remission. Thus, 12 of 14 patients were MRD negative and in remission
Remission	MRD negative	HSCT No HSCT	83.3% 16.7%	Beta distribution				12 patients were MRD negative in Lee et al, of which 10 had HSCT
Remission	MRD positive	HSCT No HSCT	0% 100%	Beta distribution				In Lee et al, no MRD positive patients received HSCT
No remission	HSCT No HSCT		0% 100%	Beta distribution				In Lee et al, none of the patients who failed to achieve remission received HSCT
Clofarabine								
Remission No Remission Death (day 0 -56)			11.5% 55.8% 32.7%	Dirichlet distribution	61	61	122	Remission based on 7 of 61 patients achieving complete remission in Jeha et al. By day 56, 32.7% of patients had died (based on digitisation of published Kaplan-Meier curve). All remaining patients (55.8%) were assumed to have non-remission
Remission	MRD negative MRD positive		14.3% 85.7%	Beta distribution				1 patient with remission had undergone HSCT and was considered to be in long-term remission (>200 days alive). This patient was assumed to have MRD negative status (MRD not reported in Jeha et al). Thus, 1 of 7 remission patients were MRD negative
Remission	MRD negative	HSCT No HSCT	100.0% 0.0%	Beta distribution				Assumption that all patients who were MRD negative went on to HSCT
Remission	MRD positive	HSCT No HSCT	16.7% 83.3%	Beta distribution				7 patients had remission, of which 1 was assumed to be MRD negative and had HSCT. Of the remaining 6 patients (MRD

								positive), a further one patient had HSCT. Thus, 1 of 6 MRD positive patients had HSCT.
No remission	HSCT No HSCT		20.6% 79.4%	Beta distribution				In total, 9 patients had HSCT in Jeha at al. Two HSCT patients were in remission, with the remaining 7 HSCT patients having no remission. During the initial 56 days, an estimated 34 (55.8%) patients had no remission. Thus, 7 of 34 no-remission patients had HSCT

Source: Lee et al[117], Jeha et al[131]

Figure 6 presents the proportion of patients occupying each state at the end of decision tree model. The model predicts that 48% of patients receiving the CAR T-cell therapy and 15% of patients receiving standard of care treatment will receive HSCT. All patients who underwent HSCT following CAR T-cell therapy were assumed to have a MRD negative status. In contrast, most patients who underwent HSCT after receiving clofarabine had not achieved remission (11.5%) prior to transplantation, with only a small proportion of patients receiving HSCT after complete remission (1.6% MRD negative, 1.6% MRD positive).

Figure 6: Proportion of patients occupying each state at the end of the decision tree model (day 56), by CAR T and comparator group



With the structural link included within the model, it was necessary to use external data rather than the evidence sets themselves for the purposes of extrapolation and estimating life-time mortality. This was necessary as the existing survival data for CAR T-cell therapy were not reported conditional on remission, MRD or HSCT status. Hence, the parameter estimates for the partitioned survival analyses were sourced from two external studies; Leung et al[71] for the post-HSCT survival probabilities, and Von Stackelberg et al[114] for the non-HSCT survival probabilities. A summary of the survival rates is provided in Table 18.

Table 18: Survival rates in patients who receive HSCT based on MRD and remission status

Treatment status	Status prior to treatment	Exponential rate parameter (standard error)	Sampling distribution used in probabilistic analysis	Proportion alive and considered "effectively" cured at year 5	Mean time to death following HSCT (years)*	Notes	Source
HSCT	MRD negative (<0.01% bone marrow blasts)	-	N/A	99.0%	43.70	Based on all-cause mortality, with adjustment for excess mortality in ALL survivors	Assumption
	MRD positive (>0.01% to 5% bone marrow blasts)	0.0121 (0.0232)	Log-normal (applied to rate)	48.5%	22.43	Leung et al reports 5-year post-HSCT survival probability of 48.5% in patients with MRD positive status. 5-year cumulative probability converted to monthly rate using equation: (-(1/60)*log(0.485))	Leung et al
	No remission (>5% to 25% bone marrow blasts)	0.0175 (0.0232**)	Log-normal (applied to rate)	35.1%	16.74	5-year post-HSCT survival probability estimated by fitting linear regression model to survival data by MRD status, reported in Leung et al. The independent variable in the regression was the cumulative hazard rate at year 5, and the dependent variable was the midpoint of each MRD category (on the log to base 10 scale). To predict the cumulative hazard at year 5 for patients without remission prior to HSCT, the midpoint MRD level of 15% was used.	
No HSCT	All patients	0.2425 (0.2085)	Log-normal (applied to rate)	0%	0.35	See main text	Von Stackelberg et al

Note: * after 5-years, all patients alive are assumed to be long-term survivors of ALL
** assumed to be the same standard error as MRD positive analysis

In the base case, it was assumed that all transplant recipients with remission and MRD negative status prior to HSCT, reverted to the same mortality rates as long-term survivors of ALL, from the time point of HSCT. Employing this assumption, as opposed to using the data reported in Leung et al for this population, provided a more consistent prediction of survival data from Lee et al, where it was reported that all 10 HSCT recipients with MRD negative status were leukaemia-free and alive at the end of study follow-up.

Transplant recipients with remission and MRD positive status were assumed to have an inferior long-term survival prognosis compared to those who were MRD negative. Similarly, recipients who failed to respond to therapy were assumed to have an inferior long-term prognosis compared to those who responded (including MRD positive and negative). Parameter estimates were obtained from Leung et al, and modelled assuming an exponential distribution for time to death.

The Leung et al data was used in the base case analysis, as this was the only study identified in the literature review that reported post-HSCT survival in patients who failed to achieve remission (marrow blasts >5.0%). The parameter estimate for no-remission HSCT patients forms an important part of predicting the long-term survival benefits of standard of care therapy, as approximately 11% of the standard of care population had HSCT despite having failed to achieve complete remission (versus 0% of the CAR T-cell trial population).

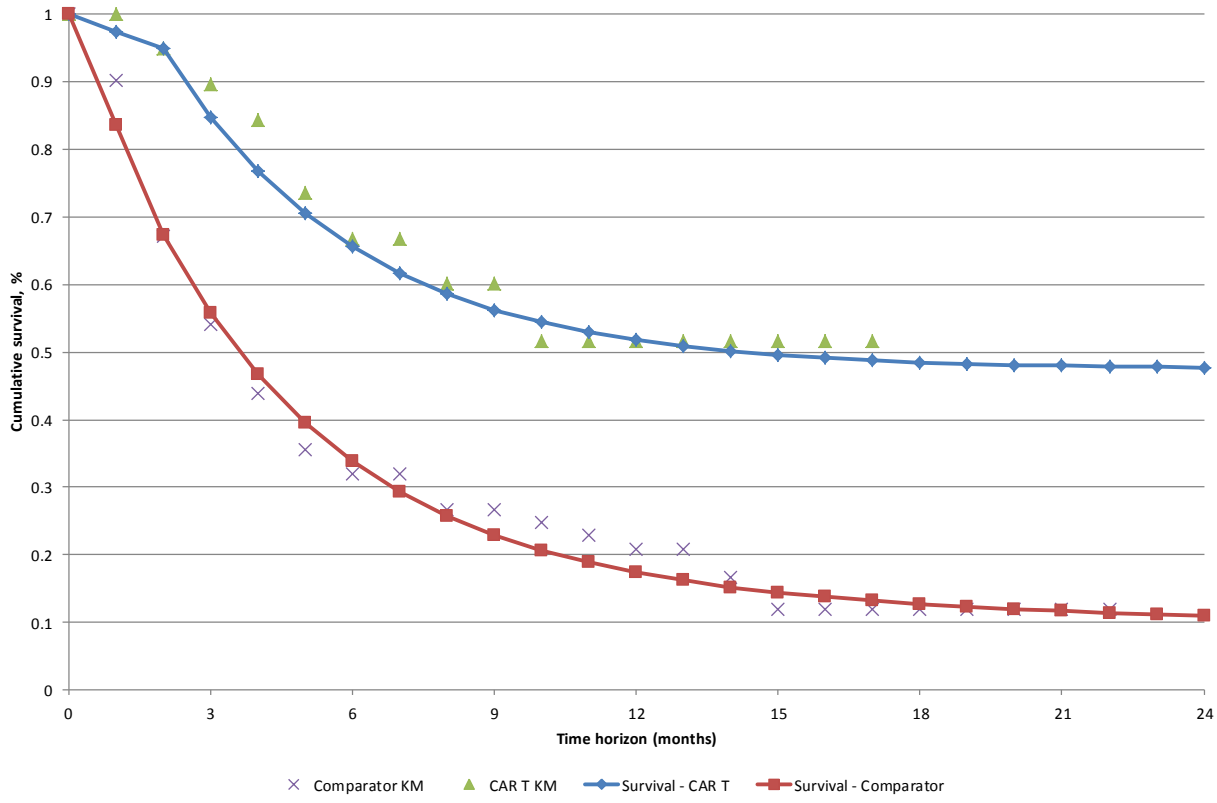
For patients who do not receive HSCT (including remission and non-remissions patients), long-term survival was modelled using data from Von Stackelberg et al.[114] A series of parametric survival functions were fitted to estimates of patient-level data generated from the published Kaplan-Meier curve. According to goodness of fit statistics, the best fitting distribution according to visual fit and goodness of fit statistic was the lognormal. However, when the function is applied in the model, the predicted overall survival for the total CAR T-cell and standard of care populations became visibly disjointed, with the risk of mortality in the decision tree phase being significantly greater than the risk being applied at the start of the partition survival phase. Consequently, there was an uncharacteristic "plateau" in the modelled survival curve between days 56 (month 2) and 84 (month 3). This plateau effect was caused by an initially low probability of death that was being predicted from the Von Stackelberg et al data.

Because of the implausible nature of the survival curve, an alternative survival distribution for Von Stackelberg et al was selected in the base case. To be consistent with the approach used in modelling post-HSCT survival (Leung et al[71]), the exponential distribution was chosen for the base case analysis.

The mean time to death with the exponential function was 0.35 years, which is consistent with the mean time to death estimated using the lognormal (0.34 years).

The validation of the responder model in predicting the outcomes of the Lee et al and Jeha et al studies was assessed by comparing the predicted survival probabilities from the model, versus the KM data extracted from these studies. As shown in Figure 7, the final model appears to provide an accurate prediction of reported survival for both CAR T-cell therapy and the comparator.

Figure 7: Model prediction (lines) versus reported KM curves of overall survival



The background all-cause mortality risks were obtained from the interim lifetables published by the UK Office for National Statistics (ONS). The ONS data report annual all-cause mortality rates by gender and by age (yearly increments from age 0 to 100 years). A gender-averaged mortality risk was derived based on a cohort that was 33.3% female (7 of 21, Lee et al)[117]. An adjustment factor for excess mortality in ALL survivors was also incorporated and modelled using data from MacArthur et al,[140] (SMR=9.1, 95% confidence interval = 7.8-10.5). These data were combined using the following equation:

$$TP(age) = 1 - e^{(-(MR(female,age)\times P(female)+MR(male,age)\times P(male))\times\frac{1}{12}\times SMR)}$$

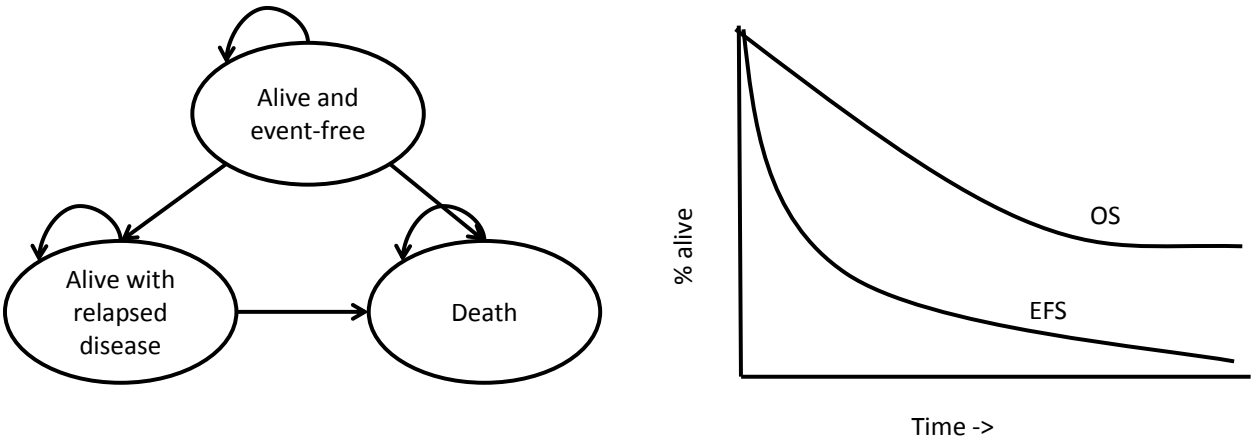
Where TP(x) is the monthly transition probability for the cohort with average age x, MR(y, x) is the ONS all-cause mortality rate for gender y and average age x, P(y) is the proportion of cohort with gender y, and SMR is the standardised mortality ratio for long-term ALL survivors versus the general population. The factor of 1/12 was included so as to convert the annualised mortality rates from the ONS to monthly rates, and probabilities, for use in the model. The mortality risk was assumed to remain constant within each year of the cohort's age.

8.2.2 Curative intent model

8.2.2.1 Structural assumptions

A simple three-state partitioned survival model was developed to assess the cost-effectiveness of CAR T-cell therapy used with curative intent. The three health states included in the model are "alive and event free", "alive with relapsed disease", and "death". An illustration of the structure of the model is provided in Figure 8.

Figure 8: Illustration of partitioned survival model structure (left, illustration of state structure; right, illustration of partitioned survival)



The health state of alive and event-free comprised of all patients who had either stable disease or had responded to therapy. The health state of alive with relapsed disease comprised patients who had either failed induction therapy, had relapsed after previously responding to treatment, or developed second malignancies. This definition is based on the criteria used in the UK ALL study.[75]

State occupancy in the model was derived using the partitioned survival technique. This involves the direct extrapolation of EFS and OS curves, which are then used to estimate the proportion of patients occupying each of the three states, via the following equations:

- Alive and event free (t) = P (EFS,t)
- Alive with relapsed disease (t) = P (OS,t) – P(EFS,t)
- Death (t) = 1- P (OS,t)

Where P(event,t) is the cumulative survival probability for event at time t.

Data on EFS were not available from either the CAR T-cell or clofarabine studies. In the absence of data, the EFS curve was derived from the available OS data, through assuming a proportional relationship between EFS and OS. This relationship is justified on the basis that EFS is highly correlated to OS, as it includes death prior to recurrence.

In the short-term, it was assumed that the cumulative hazard function for EFS would be proportional to the cumulative hazard function for OS. This was modelled based on data from the UK ALL study.[75] The proportional relationship between EFS and OS is not expected to continue indefinitely, given the potential for cure of disease, and the expectation that after a finite period of time all patients alive in the simulation would also be free of relapsed disease (EFS=OS). This is equivalent to saying that at some point in time, all patients who are alive are long-term survivors of ALL. Therefore, in the model, the proportional relationship between EFS and OS was assumed to continue up to year 5 of the simulation (the assumed point of "effective" cure in ALL). After year 5, the cumulative survival probabilities for EFS were assumed to be flat up to the point where EFS equals to OS. In all cases, EFS was always assumed to be less than or equal to OS to avoid a negative number of patients being assigned to the relapsed disease state.

In common with the 'Bridge to HSCT' scenario, at year 5 of the simulation, those who were alive in the 'curative intent' model were also subsequently assumed to be long-term survivors of ALL. From this point forward, the cohort was considered to be effectively 'cured' of ALL, and experienced the mortality risk profile consistent with a long-term survivor of ALL. The mortality risks after year 5 were also modelled based on general population age- and gender-adjusted all cause risks of mortality adjusted for excess morbidity and mortality reported in cohorts of long-term survivors of ALL.

The model evaluation also included the costs and consequences of treatment-related adverse events, which included cytokine release syndrome and B-cell aplasia, whose occurrence is specifically associated

with CAR T-cell therapy. Other events captured in the model, include encephalophy, hypotension, febrile neutropenia, neutropenia, anaemia, thrombocytopenia, leukopenia, hypokalemia, and hypophosphatemia. All events, with the exception of B-cell aplasia, were assumed to occur at the time of treatment initiation and to resolve within the first year of therapy. The cost consequences of these events were therefore captured at the start of the evaluation.

The occurrence of B-cell aplasia in patients treated with CAR T-cells is an expected consequence of CAR T-cell therapy, and is linked to the proliferation of CAR T-cells and the associated durability of the clinical effect. Consequently, for some patients, treatment of B-cell aplasia is expected to persist beyond the first year of post-CAR T therapy. To capture this in the model, a series of survival models were fitted to data on the time to CDLT-19 positivity or relapse reported in Maude et al, and used to predict the proportion of patients requiring treatment for B-cell aplasia.

Table 19: Summary of key modelling assumptions

Input	Assumption
Survival during the first 5-years	Survival was modelled based on a weighted average survival distribution.
Survival after the first 5-years	All patients alive at 5-years are considered to be long-term survivors of the disease. Long-term survivors of ALL experience excess morbidity and mortality compared to the general population.
Treatment / re-treatment	In the base case, it was assumed that all patients received a single full course of therapy. Re-treatment with CAR T-cell therapy or standard of care therapy was not permitted in the base case, but was considered in the sensitivity analysis
Treatment effect	Treatment with CAR T-cell therapy is assumed to lead to an increase the number of patients achieving a sustained cure for ALL, and therefore extend the life expectancy of patients with ALL The clinical parameter estimates used to inform the models, and TPPs can be generalised to the UK NHS
Adverse events	Treatment-related adverse events were considered in the evaluation, and included cytokine release syndrome and B-cell aplasia. The costs and health consequences of all adverse events except B-Cell aplasia were accrued at the start of the evaluation. The costs of B-cell aplasia was modelled by estimating the probability of patients with B-cell aplasia over time, using data from Maude et al.
Health-related quality of life	Patients who are event-free are assigned a higher utility weight than patients who have relapsed disease. Transplantation is associated with a one-off adjustment to utilities.

8.2.2.2 Efficacy parameter estimates - Partitioned survival model ("curative intent")

The primary data sources for overall survival in the "*curative intent*" model were the same imputed patient data used to derive the evidence sets reported in Section 6.7.2. Each separate evidence set was then analysed using parametric survival modelling to inform the 5-year survival estimates and projections applied within the cost-effectiveness analyses. The parametric analyses were undertaken using the *FlexSurv* package in the statistical programming platform R.

A series of survival distributions were considered in the analysis, including the Exponential, Lognormal, Weibull, and Gompertz. Because of the potential curative nature of CAR T-cell therapy (and therefore the

potential for an unconventional hazard function), a series of flexible cubic spline models were also considered in the analysis. The cubic spline models were based on those developed by Royston and Parmar.[141] Cubic spline models expressed on the proportional odds scale were used as they appeared to converge to an optimised solution more frequently than the proportional hazards or probit variants of the cubic spline model. A series of one-, two-, three-, and four-knot spline models were considered. The knots were evenly distributed across the time scale of the study, as per the default settings for the *FlexSurv* package in R.

Separate curves were fitted to the hypothetical CAR-T cell data and the comparator data to allow both the shape and scale of the distribution to vary between these. Alternative options include fitting proportional hazard models to a dataset containing both treatments and including a covariate in the regression for treatment assignment. This alternative approach was not considered here given that an earlier assessment of the validity of the proportional hazards assumption illustrated that this assumption may not consistently hold across all evidence sets.

Within cost-effectiveness studies, it is common practice to use a single survival distribution in the base-case analysis. This is chosen based on goodness of fit statistics, the fit of each distribution to the Kaplan Meier curves, and the clinical plausibility of subsequent model projections over the full time horizon. However, it is unlikely that a single survival distribution can adequately characterise uncertainties over the longer-term extrapolation period. The robustness of the ICER estimates to alternative distributions can be considered within separate sensitivity analyses or scenarios. However, transparency concerns may exist regarding this approach if their weighting is not explicitly specified in subsequent policy decisions.

To more formally account for the uncertainty surrounding choice of survival distribution, a model averaging approach was adopted using the methods outlined in Jackson et al.[142] This technique involves the parameterisation of uncertainty surrounding the choice of distribution, through including all plausible survival functions as part of a weighted distribution, and sampling both the parametric uncertainty associated within each distribution and the uncertainty (or weights) surrounding the choice of preferred method. Through the probabilistic analysis, it is therefore possible to estimate the joint distribution of uncertainty around the parameter estimates and the choice of survival function.

Each model is assigned a weight that represents the adequacy of that distribution in predicting the lifetime survival of the modelled cohort, in comparison to all other distributions considered in the model. There are a number of measures of model adequacy that can be considered. Examples include statistical

adequacy measures such as the Akaike Information criterion (AIC) and Bayesian information criterion (BIC), and expert judgement. The weights considered in this evaluation were based on AIC scores. As outlined in Jackson et al, the AIC values reported from each survival distribution was converted to a probability weight (w_k) using the following equations:

$$A_k = e^{(-0.5 \times AIC)}$$

$$w_k = \frac{A_k}{\sum A_k}$$

The weighted distribution was then applied in the base case analysis. Different model weights and parameter estimates were considered across the three different data sets, as outlined in the following sections.

Minimum data set

A summary of the goodness of fit statistics for each distribution fitted to the imputed survival data across each of the evidence sets is provided in Table 20.

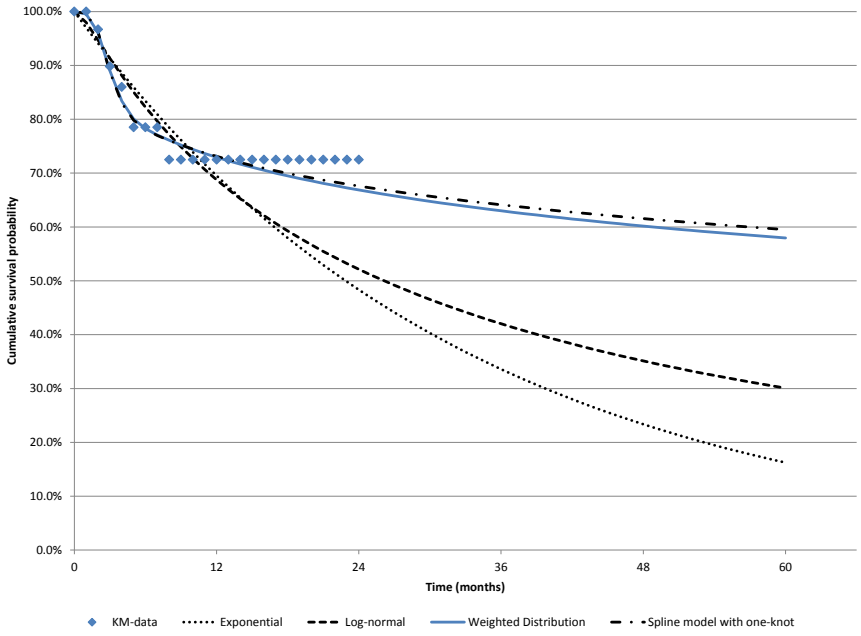
Table 20: Summary of goodness of fit statistics and weights for survival distributions (minimum set)

Distribution	CAR T- cell		Standard of care	
	AIC	AIC based weight	AIC	AIC based weight
Exponential	127.91	1.9%	302.15	0.1%
Weibull	129.88	0.7%	303.31	0.0%
Gamma	129.91	0.7%	304.09	0.0%
Gompertz	139.21	0.0%	303.01	0.0%
Lognormal	128.70	1.3%	291.00	13.8%
Spline with a single knot	121.02	60.1%	288.65	44.6%
Spline with two knots	122.97	22.7%	290.41	18.5%
Spline with three knots	124.93	8.5%	291.32	11.7%
Spline with four knots	126.44	4.0%	291.42	11.2%

According to the AIC statistic, the distribution with the best goodness of fit to the CAR-T cell data was the spline model with a single knot (AIC=121.02), followed by the spline with two knots (AIC=122.97). The spline model with a single knot was assigned the highest single weight of 60.1%, and was followed by the two- (22.7%), three- (8.5%) and four (4.0%) -knot spline configurations respectively. A visual

comparison of the survival data based on the weighted distribution and several single distributions are reported in Figure 9.

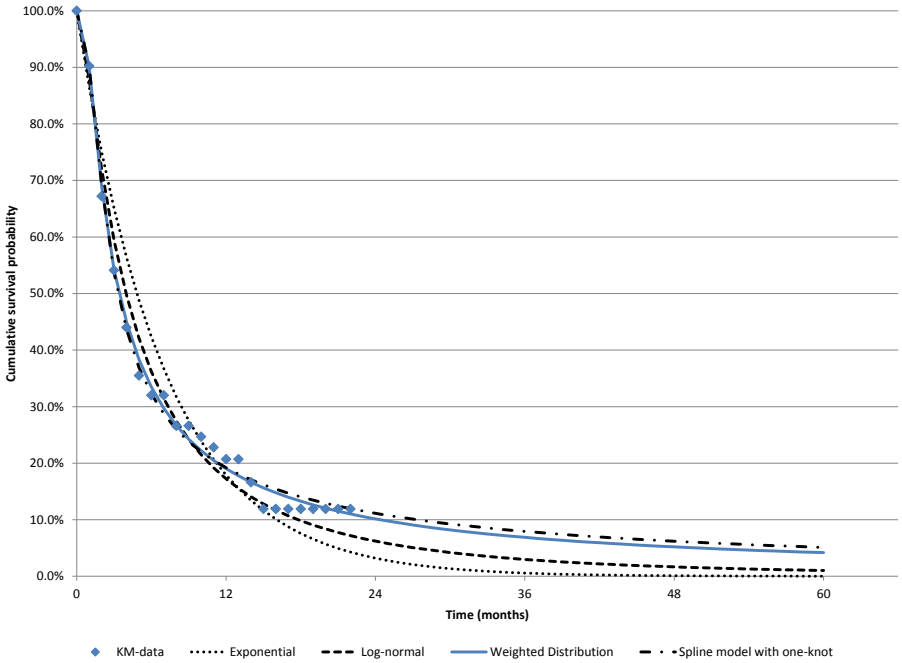
Figure 9: Plot comparing fit of weighted distribution and key single distributions to CAR-T data



Because of the limited maturity in the minimum set, there was considerable variation in the predicted long-term survival of the modelled cohort, as shown by the spread of survival trajectories in Figure 9. Whilst the "best-fitting" spline models appeared to generate a robust fit to the data over the first 3 months of the study, the functions were not able to accurately predict the tail of distribution. In this case, the "best fitting" model underestimated the KM probabilities from month 18 of the simulated time horizon. The weak fit of the model to the tail of the KM is partly driven by the limited data available to support the continued flattening of the curve. As shown later in this section, with additional data maturity, the parametric models tend to provide a better prediction of the tail of the KM as there is more data to support the long-term flattening of the survival curve.

In the standard of care group, the distribution with the optimal predictive validity as judged via AIC, was also the spline model with a single knot (AIC=288.65). A weight of 44.6% was assigned to the spline model with a single knot, followed by weights of 18.5% for the spline model with two knots and 13.8% for the lognormal. A visual comparison of the survival data based on the weighted distribution and several single distributions are reported in Figure 10.

Figure 10: Plot show predictive fit of weighted distribution, spline model with one knot, lognormal, and exponential functions to Kaplan-Meier curves



Intermediate and mature data sett

A summary of the goodness of fit statistics and weights is reported for the intermediate and mature evidence sets in Table 21 and Table 22.

Table 21: Summary of goodness of fit statistics and AIC-based weights for survival distributions (intermediate set)

Distribution	CAR T- cell		Standard of care	
	AIC	AIC based weight	AIC	AIC based weight
Exponential	157.43	0.0%	345.45	0.0%
Weibull	147.51	0.0%	329.54	0.0%
Gamma	148.64	0.0%	337.70	0.0%
Gompertz	161.81	0.0%	343.38	0.0%
Lognormal	143.41	0.0%	308.08	0.1%
Spline with a single knot	125.18	61.2%	296.00	56.0%
Spline with two knots	126.95	25.3%	298.08	19.7%
Spline with three knots	128.84	9.8%	299.58	9.4%
Spline with four knots	130.84	3.6%	298.66	14.8%

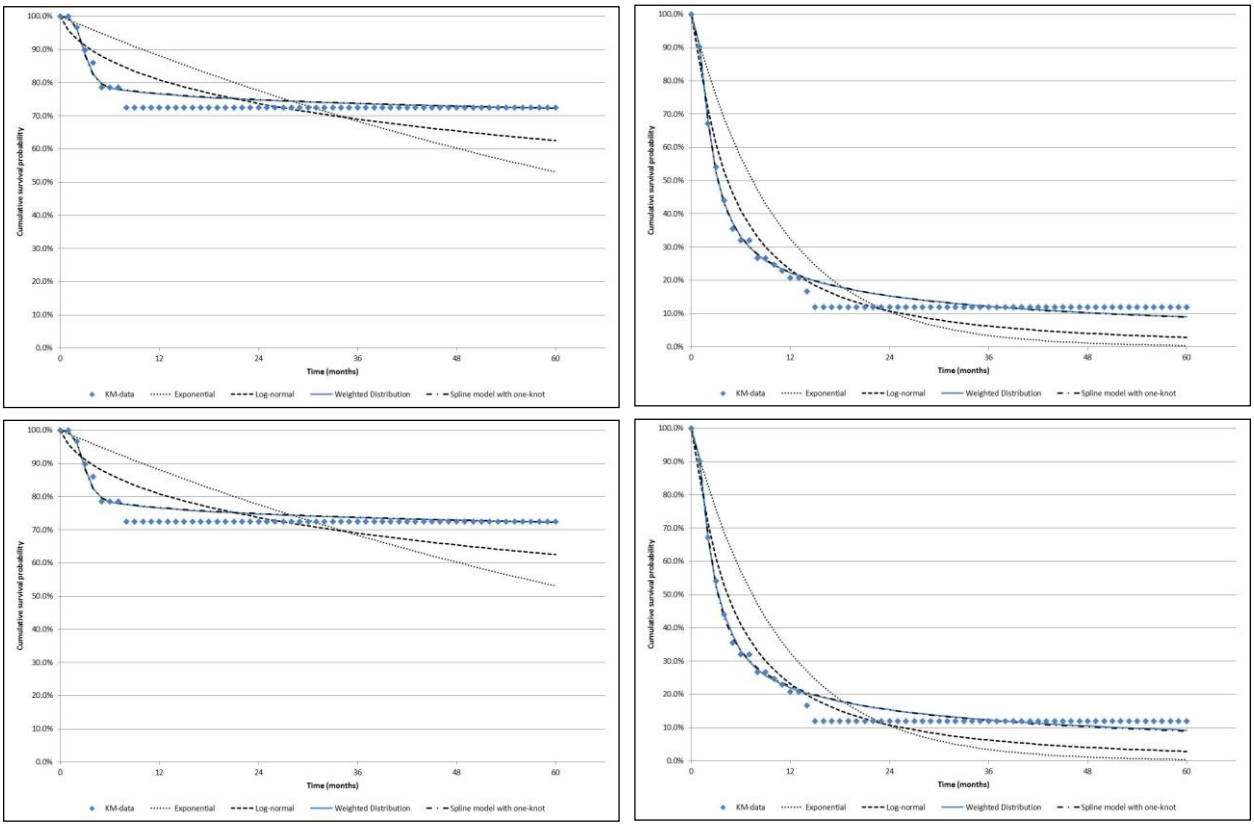
Table 22: Summary of goodness of fit statistics and AIC-based weights for survival distributions (mature set)

Distribution	CAR T- cell		Standard of care	
	AIC	AIC based weight	AIC	AIC based weight
Exponential	312.85	0.0%	688.89	0.0%
Weibull	291.02	0.0%	655.07	0.0%
Gamma	293.28	0.0%	671.39	0.0%
Gompertz	339.75	0.0%	644.99	0.0%
Lognormal	282.81	0.0%	612.15	0.0%
Spline with a single knot	244.36	60.3%	586.00	34.0%
Spline with two knots	245.89	28.0%	588.17	11.5%
Spline with three knots	247.65	11.7%	589.17	7.0%
Spline with four knots	NA	NA	585.32	47.6%

The additional maturity of the data in these evidence sets and the superior AIC statistics associated with the flexible spline models, resulted in none of the standard distributions being assigned a weighted greater than 0.1%. The different levels of precision resulted in small difference in the weights assigned to the spline models across the intermediate and mature evidence sets.

Visual comparisons of the survival data based on the weighted distribution and several single distributions are reported in Figure 11.

Figure 11: Plot show predictive fit of weighted distribution, spline model with one knot, lognormal, and exponential functions to Kaplan-Meier curves



In comparing across evidence sets, the survival models fitted to the intermediate and mature evidence sets appear to have a shallower slope than those fitted to the minimum evidence set, resulting in a longer tail to the predicted survival curves. This is driven by the assumption that in the more mature evidence sets there is greater certainty over the "curative" benefit of treatment because of additional evidence on patient survival up to month 60 of the hypothetical evidence set (versus maximum survival of approximately 24 months in the minimum set). This is broadly equivalent to saying that, in the intermediate and mature evidence sets, there is greater certainty over the flattening of the KM curve.

When comparing across competing survival models, the intermediate and mature evidence sets are also associated with a more consistent set of survival projections than in the minimum set. This leads to a narrower range of potential survival probabilities being predicted at later time points in both intermediate and mature sets. Therefore, unlike the bridge to HSCT model, additional evidence maturity in the curative model leads to a different projection of survival benefit, as well as impacting on the parametric uncertainty surrounding model extrapolations.

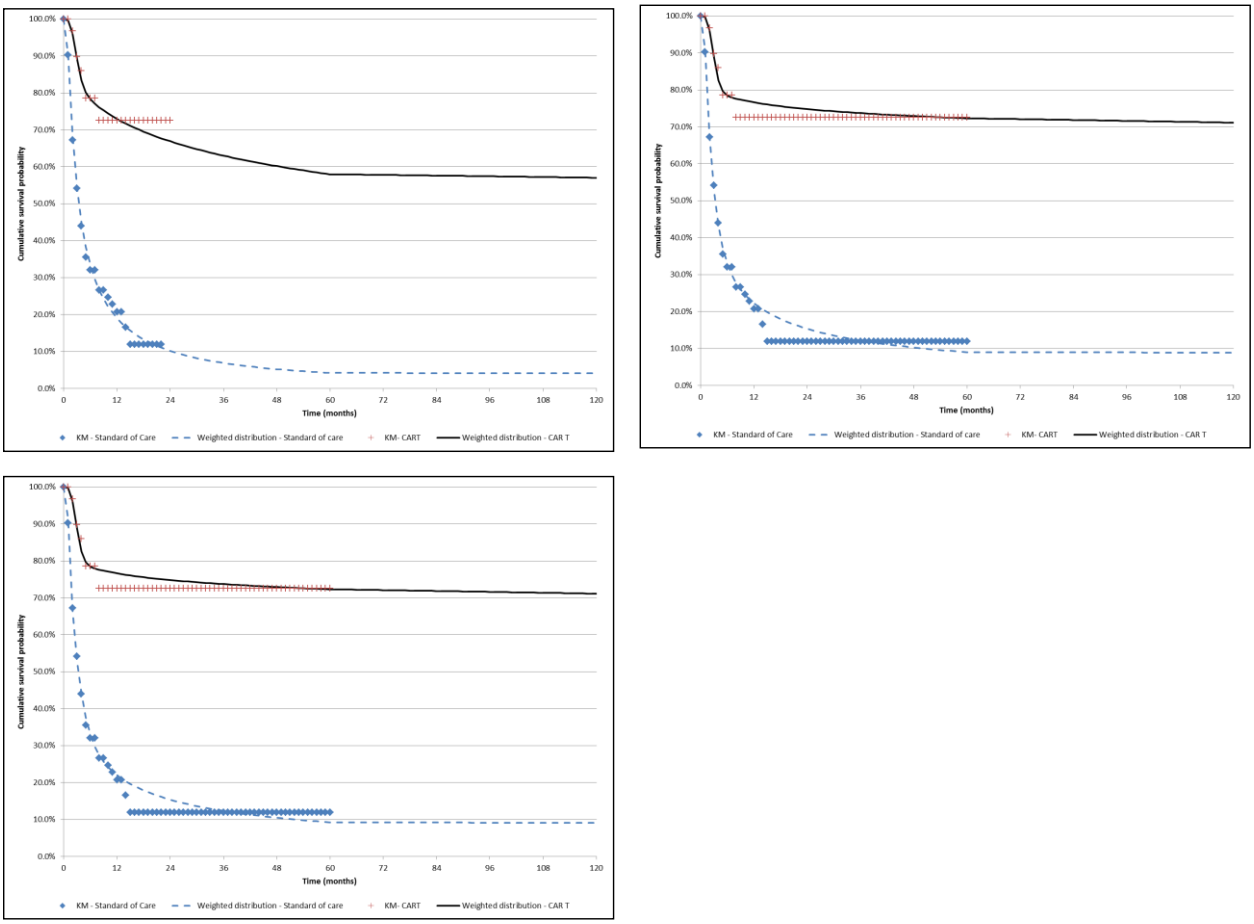
There are slight differences in the survival curves predicted from the intermediate and mature evidence sets because of differences in the weights applied to different functions. These differences cannot be clearly seen on the plots, as the difference in weights is marginal. The key difference between these evidence sets is the additional sample size assigned to the mature set, which primarily impacts the uncertainty/precisions surrounding survival estimates, which is not shown on these plots.

Adverse events – B-cell aplasia

A series of survival models were fitted to data on the time to CDLT-19 positivity or relapse reported in Maude et al, and used to predict the proportion of patients requiring treatment for B-cell aplasia. The best fitting distribution was the Weibull distribution.

The accuracy of the partitioned survival model in predicting the outcomes of the Maude et al[118] and Jeha et al[131] studies was assessed by comparing the predicted survival probabilities from the model, versus the Kaplan-Meier data. As shown in Figure 12, the final models appear to provide an accurate prediction of the extracted KM curve for overall survival in both studies.

Figure 12: Model prediction (lines) versus extracted Kaplan-Meier curves (markers) of overall survival in all patients



8.3 Resource use and costs – Bridging and curative models

The resource use and costs incorporated within each separate model are based on the following components:

1. Treatment acquisitions costs
2. Administration and monitoring costs
3. Adverse events
4. HSCT
5. Long term costs

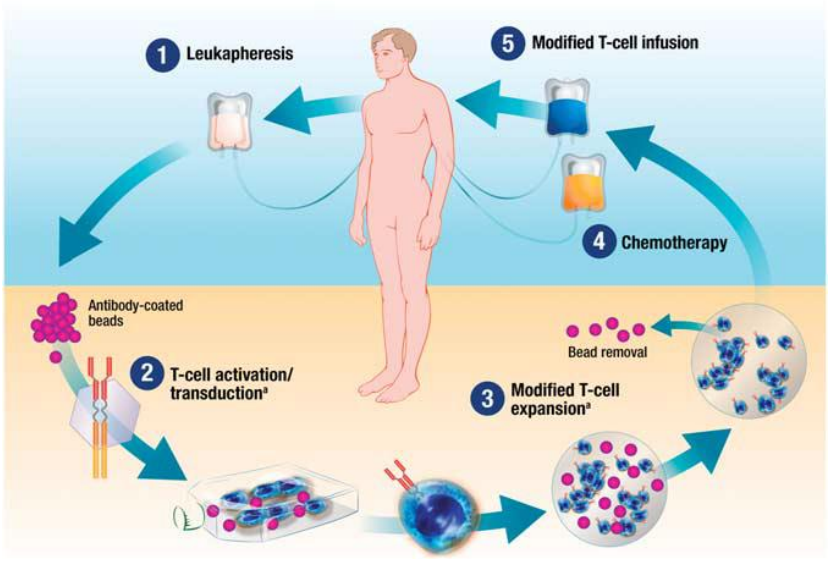
8.3.1 Acquisition costs

8.3.1.1 CAR T-cell therapy

The complex nature of regenerative medicines and the treatment pathway makes it necessary to disentangle the separate procedural elements of the CAR T-cell treatment process and to make assumptions concerning which elements would be included within the acquisition cost of the therapy itself and which might represent additional procedural costs which would need to be separately provided and funded by the NHS itself.

Levine et al[99] provides an overview of the CAR T-cell process, reproduced below in Figure 13.

Figure 13: CAR T-cell therapy process



The overview separates the processes of leukapheresis, conditioning chemotherapy and infusion (steps 1, 4 and 5 in the figure), from the transduction and expansion (steps 2 and 3). We assume the same split to represent those components of care which would be provided (and funded separately) by the NHS and those which would be undertaken by the manufacturer and included within the acquisition cost of CAR T-cell therapy. Hence, we assume the acquisition cost of CAR T-cell therapy would not include the cost to the NHS of providing leukapheresis, conditioning chemotherapy or cell infusion, and that these are assumed to represent additional costs to the NHS.

In the absence of licensed products being available, there are currently no commercially available estimates of the acquisition cost of CAR T-cell therapy. Informal sources have indicated future

acquisition costs may be in the region (US $) of $150,000- $500,000.[143] Within the exemplar, we have assumed that the manufacturer would employ a value-based approach to pricing, such that the acquisition cost would be set at a level such that the resulting cost-effectiveness (ICER) estimates would be close to NICE's current threshold range. In the context of the specific population considered, we have assumed that this would be in line with the £50,000 per QALY estimate based on NICE's current approach to treatments at the end of life. We subsequently explore the impact of alternative prices and payment schemes using separate scenarios. Full details of the hypothetical prices assumed across the separate scenario are reported in Section 9.

The acquisition cost of conditioning therapy (£3,803) is estimated from the regimen used in the Lee study,[117] which represents fludarabine 25 mg/m² per day on days –4, –3, and –2 and cyclophosphamide 900 mg/m² per day on day –2.

The acquisition cost associated with clofarabine is derived from the AWMSG report for clofarabine, which reported a cost of £43,200 per patient treated, based on the average costs of the drug volumes used in study CLO-212 (based on 1.8 cycles of treatment, a patient body surface area of 1.2m^2 and the licensed dose of 52mg/m^2/five-day treatment cycle).

The acquisition costs of FLAG-IDA are considered as part of a separate sensitivity analysis and are estimated by applying unit costs from the BNF to a dosing guide published by the Royal Surrey NHS Trust.[144] Assuming an average body surface area of 1.2m^2 and an average of 1.76 cycles of treatment[145] gives the estimate of £3,808 per patient.

8.3.1.2 Administration and monitoring costs

In addition to the acquisition costs, it is important to consider the resource use and costs associated with administration and subsequent monitoring. All patients regardless of subsequent treatment are assumed to require an initial non-elective hospitalisation. For clofarabine and FLAG-IDA it is assumed that the costs of this hospitalisation also include all costs associated with monitoring and administration of treatment. For CAR T-cell therapy, the same initial hospitalisation is assumed to occur for the administration of the conditioning therapy. However, due to the additional production period required to manufacture the CAR T-cells (in the region of 11 days currently) an additional, elective hospitalisation is also assumed during which CAR T-cells are subsequently administered and the patient monitored. The cost of a single leukapheresis procedure is also applied to CAR T-cell patients.

Table 23 reports these per patient costs, the sources and associated assumptions.

8.3.1.3 Adverse events

The individual costing of each adverse event for the alternative treatments could entail double counting, since some aspects of these may already be included in the hospitalisation costs used for the administration and monitoring costs of each treatment. Therefore, an assumption is made that all grade 3 and 4 adverse events (except CRS and B-cell aplasia as discussed below) require an extension of hospitalisation by one day with a cost based on the excess bed day HRG cost as shown in Table 23.

For CRS, a combination of the acquisition cost of cytokine inhibitor drugs and an admission to a paediatric intensive care unit are assumed for all cases of grade 4 or severe CRS.

B-cell aplasia is assumed to be treated with a regimen of IVIG, given at a dose of 0.5g/kg every 4 weeks until the patient is no longer in need of treatment (i.e. CD19 positivity, relapse or death). We assumed the population treated to have an average weight of 49.5kg. Rounding down of each dose to the nearest vial (i.e. 20g vial per dose), the cost per vial is estimated as £850. In addition, an administration cost of £225 per dose is assumed.

8.3.1.4 HSCT

Three potential sources of cost estimates of HSCT were identified and considered:

1. **NHS Reference Cost**. This provides estimates of completed HRG activity and unit costs across six different paediatric allogeneic transplantation categories. While intuitively appealing due to the relevance to our population and UK context, concerns have been raised[146] that these do not capture the full cost of HSCT, due to their focus on a single admission period.

2. **The London Specialised Commissioning Group Report**.[147] This report estimated a national tariff for adult blood and bone marrow transplants based on the phases of transplantation from decision to transplant to 100 day post-transplantation follow-up care. However, no details are given as to how the estimate was derived. In addition, the estimate only considers an adult population.

3. **UK Stem Cell Strategy Oversight Committee Report**.[146] This report used results from a Dutch study published in 2002, reporting the cost of allogeneic adult unrelated bone marrow transplantation. This estimate includes all initial costs of the transplantation as well as follow-up

costs for up to two years after the transplantation. The inclusion of the longer-term follow up costs addresses the primary concern around existing Reference Costs. However, there exists uncertainty about the generalisability of the cost to the specific population considered here.

To take account of the limitations around each of the three data sources, the model combines estimates from both the Reference Costs and the UK Stem Cell Strategy Oversight Committee Report. The London Specialist Commissioning Group Report was discounted due to a lack of details as to how the estimate was derived.

The cost of HSCT is considered in two parts: the cost of the procedure and associated hospitalisation, and the cost of long term care. While all three sources provide an estimate of the cost of the procedure, both the London Specialist Commissioning and UK Stem Cell Committee focus on adult populations. Existing HRG costs report that a higher cost of the procedure for paediatric patients, with paediatric HRG costs between £21,622 and £74,434 more than the equivalent adult HRG costs across the four different forms of allogeneic transplantation reported.[148] Therefore, the cost of the procedure has been estimated as the weighted average (by frequency of HRG) of all paediatric allogeneic transplantations from the HRG costs to estimate the cost of initial transplantation.

As previously noted, the HRG costs only include the costs accrued in the admission in which the transplantation occurred. Hence, any longer-term costs will not be included. To estimate the longer-term costs, an estimate of post-transplantation costs from the UK Stem Cell Strategy Oversight Committee Report is used. No further adjustment is made to the estimate. The use of this estimate makes the same assumptions about the appropriateness of the original source of the costs.[149] The use of this estimate in our population additionally assumed that, unlike the cost of the procedure, long-term costs are independent of type of transplantation and age of patient at time of transplantation.

Table 23: Model inputs - costs

Parameter	**Cost**	**Source/assumption**
1. Acquisition costs		
1a. CAR T-cell		
Acquisition cost of CAR T	Threshold analysis	Threshold price analysis based on three approaches detailed above
Conditioning therapy	£329.86 per patient	Acquisition costed directly from Lee assuming full use of 2x50mg fludarabine vials and 1x500mg and 1x1g vial of

		cyclophosphamide and a body surface area of 1.2m^2,[145] infusion costs assumed included in CAR T admin costs below
1b. Clofarabine		
Acquisition cost of clofarabine	£43,200 per patient	Cost presented in AWG FAR for clofarabine, excluding costs of administration
1c. FLAG IDA		
Acquisition cost of FLAG-IDA	£3,808.57 per patient	Cost per cycle estimated from the Royal Surrey guide [144], average body surface area of 1.2m^2 and the average number of cycles of FLAG-IDA of 1.76[145]
2. Administration and monitoring costs		
2a. CAR T-cell		
Leukapheresis	£1,627 per patient	Weighted average of HRGs for stem cell and bone marrow harvest
Initial hospitalisation for conditioning	£7,179.99	HRG paediatric ALL admissions weighted average non-elective long stay
Additional hospitalisation for CAR T treatment	£5,831.72	HRG paediatric ALL admissions weighted average elective inpatient
2b. Clofarabine		
Hospitalisation over treatment period	£7,179.99	HRG paediatric ALL admissions weighted average non-elective long stay
2c. FLAG IDA		
Hospitalisation over treatment period	£7,179.99	HRG paediatric ALL admissions weighted average non-elective long stay
3. Adverse events		
CRS	£2,857.99 per patient per grade 4 or severe CRS event	Combination of the cost of drug (£1,193 HRG for cytokine inhibitor drugs) plus ICU hospitalisation (£1,664.99 weighted all advanced critical care paediatric ICUs)
B-cell aplasia	£1075/month per patient for the first 3 months	Dose of 0.5g/kg every 4 weeks until the patient is no longer in need of treatment (i.e. CD19 positivity, relapse or death)
Febrile neutropenia	£0	Assumed included in CRS costs
Encephalopathy	£539.24 per patient per adverse event	HRG paediatric ALL admissions weighted excess bed day non-elective inpatient stay
Hypotension		
neutropenia/ neutrophil count decreased		

Anemia		
Thrombocytopenia/ platelet count decreased		
Leukopenia/ white cell decreased		
Hypokalemia		
Hypophosphatemia		
4. HSCT		
Transplantation	£89,879.15 per patient	Weighted average of paediatric transplant HRGs, elective inpatients only
Follow-up costs	£61,965 per living patient	Sum of follow-up costs from UK Stem Cell Oversight committee report[146] (<6 months = £28,390, 6-12 months = £19,502, 12-24 months = £14,073). In the model these will be included as time and OS dependent.
4. Long term costs		
Post non-HSCT population	£7,179.77, at point of death	HRG paediatric ALL admissions weighted average non-elective long stay
Curative model population	£7,179.77, at point of recurrence	HRG paediatric ALL admissions weighted average non-elective long stay

8.3.2 Model inputs – utilities

8.3.2.1 Literature Review

A pragmatic approach was taken to identify potentially relevant sources for health utilities. Google and Google Scholar were used to search for publicly available utility estimates, alongside a search of known economic evaluations and HTA appraisals in ALL. The search focussed on utility estimates of children with ALL, regardless of treatment provided. Two systematic reviews of utility studies in paediatric ALL were identified.[150, 151]

Van Listenburg et al.[151] reviewed the measurement of health related quality of life (used synonymously with utilities) in paediatric patients with ALL using the Health Utilities Index (HUI). The study identified 15 studies reporting utilities in this population using both HUI2 and HUI3. The Van Listenburg review has several issues that limit its relevance to our model. Firstly, no attempt was made to meta-analyse the results, with the review only summarising the individual utility estimates from each study. In addition, the results were reported by phase of care, often focussing on specific time points in the treatment pathway rather than to specific health states relevant to our modelling. Given the time constraints in our work, a more detailed consideration of each study was not considered feasible.

Kelly et al.[150] undertook a decision analysis of cranial radiation therapy for paediatric T-cell ALL patients, including a systematic review of utility studies to inform this. While the study focussed on T-cell ALL, the review of utilities did not stipulate type of ALL and hence included all forms of ALL. The study used existing mapping functions to convert generic HRQoL measures (SF36 and CHRIs) to preference based utility estimates (HUI2 and EQ-5D). Of particular relevance to our model are the states of *in the state of relapse* and *cured after relapse*, with mean utility estimates reported of 0.75 (range 0.44 to 1) and 0.91 (0.87 to 0.95) respectively. Since the exemplar considers patients who have previously failed prior lines of therapy, the state for *cured after relapse* was considered to be more consistent with expected utility than the state for *cured after initial treatment* (mean utility 0.92; range 0.90 to 0.94).

In addition, the pragmatic search also identified a number of published economic evaluations which had used utility estimates.[124, 134, 135, 137]

Of the three AWMSG Final Appraisal Recommendations (FAR) related to ALL, one did not report any utility results from the manufacturer's submission (Dasatinib). The clofarabine FAR reported that all patients who survived post 1 year after HSCT were assumed to have the utility of the general population. All other states modelled were varied between 0.2 and 1 as scenario analyses to demonstrate the results were not sensitive to the utility values of those who do not survive long-term. The nelarabine FAR[137] reported that non-responders and untreated patients were assumed to have utility of 0.64. This value was referenced from Health Outcomes Data Repository (HODaR) data of patients with lymphoid leukaemia, and as such represents patients in secondary care. In addition, all patients who undergo successful transplantation were assumed to have a utility of 0.92 based on a study by Sung et al.[152]

The Sung study considers physician elicited estimates of utility for AML patients who have survived without recurrent disease post transplantation. Sung additionally presents an estimate of disutility (i.e. decrement associated with an event) associated with treatment with chemotherapy and transplantation, estimated as 0.42 (plausible range 0.16 to 0.83) and 0.57 (0.31 to 0.87), respectively. No estimate of the duration of these disutilities are presented.

Similar to Sung, the economic evaluation of clofarabine for paediatric ALL conducted by Lis et al[135] conducted an elicitation exercise of physicians due to a lack of relevant utility estimates available at the time. Lis reported utility estimates for during treatment with palliative care (0.26) clofarabine without HSCT (0.34), clofarabine with HSCT but surviving less than 1 year (0.48), as well as survival post HSCT for 1 year (0.80), 2 years (0.85) and beyond (0.88).

While these values appear generally consistent with the results reported within the systematic reviews, the magnitude of the treatment disutilities appear higher. It is plausible that this discrepancy may be the result of the use of physician rather than patient utility elicitation.

8.3.2.2 Informing the model states

All model utility inputs applied in the model are summarised in Table 24.

Table 24: Model inputs - utilities

Parameter	Utility (95% CI)	Source/assumption
Treatment disutilities		
HSCT disutility	0.57 for one year (0.33-0.87)	Sung 2003 'disutility of undergoing BMT' expert VAS elicitation
Adverse events		
CRS	0 for one week	Assume severity of ICU hospitalisation associated with utility of 0
Short-term utility		
Relapse	0.75 (0.44–1)	Kelly 2015 'in the state of relapse' mapped value from CHRIs to EQ5D
Remission	0.91 (0.87–0.95)	Kelly 2015 'cured after relapse-all relapsed patients treated with CRT' mapped value from SF-36 to HUI2, need to assume no long terms disutility AEs from CRT
Long-term utility		
Long term disutility	Remission utility (0.91) with age adjusted decrement	

8.3.2.3 Treatment disutilities

Due to a lack of literature considering the short-term impact on health utility associated with both chemotherapy and HSCT, we based our estimates on the study by Sung.[152] A decrement in utility of 0.57 for HSCT and 0.42 for all forms of chemotherapy is assumed. Both estimates are assumed to incorporate all short term adverse events associated with both treatments. However, Sung fails to report any estimate of duration associated with the estimated disutilities for either treatment. Therefore we assume disutilities

apply for one year post treatment initiation. As the disutility estimate for all forms of chemotherapy is the same in both treatment arms, the impact will cancel out and is therefore excluded from our model.

8.3.2.4 Adverse events

As discussed in the previous section, all HSCT and chemotherapy adverse events are assumed to be incorporated in the treatment disutility estimates applied. The only additional adverse events to consider are those specifically associated with CAR T-cell therapy. As discussed in the cost section, only CRS and B-cell aplasia are expected to be associated with potential additional burden not considered elsewhere in the model. The pragmatic literature review was unable to find any specific estimates of disutility or duration associated with either adverse event.

For severe (grade 4) CRS it is assumed that, due to the severity of initial onset of the event and associated intensive care admission, a utility of 0 is incurred for one week. For B-cell aplasia, while there is a large cost burden associated with its management there is little evidence of any significant impact on patient utility. In existing CAR-T cell studies. B-cell aplasia appears to be either well managed or short lived, with no reported cases of associated intensive care hospitalisation. Therefore, no disutility is assumed for cases on B-cell aplasia.

8.3.2.5 Short-term health-related quality of life

The model considers short-term response as either relapse or remission. The utility estimates to inform these estimates are derived from the Kelly et al.[150] study, with a utility of 0.75 assigned to the relapse state and 0.91 to the remission state.

8.3.2.6 Long-term health-related quality of life

Patients with the severe form of ALL considered in the model are likely to experience long-term comorbidities associated with the disease and associated disutility. As such the utility score estimated for the state of remission is applied with an additional age related decrement.

8.4 Conclusions

Two *de novo* decision models were developed to assess the cost-effectiveness of CAR T-cell therapy within the 2 separate TPPs (Bridge to HSCT and Curative intent) across each of the separate evidence sets. Although, a number of common inputs and assumptions were employed across both models, the 2 models had important structural differences which led to differences both in the underling modelling approach as well as in the use of external evidence.

In the *"Bridge to HSCT"* scenario, the primary health benefits of treatment with CAR-T cell therapy were assumed to be driven by an increase in the proportion of patients receiving HSCT and the subsequent success of HSCT itself (based on remission and MRD status). The introduction of an epidemiological 'link' between a potential established surrogate outcome and/or process (i.e. MRD and HSCT status) and final health benefits (i.e. OS and QALYs) also enabled the use of external evidence to be utilised alongside the separate hypothetical evidence sets generated. A landmark response model was developed utilising evidence from the hypothetical evidence sets to inform short-term outcomes on remission, HSCT and MRD status and external evidence to estimate overall survival conditional on these shorter-term outcomes. Hence, the key assumption employed within this scenario is that external evidence substantiating the relationship between MRD and HSCT status in studies in which CAR T-cells have not been used can be generalised to patients in whom CAR-T cells have been used. Importantly, results of our validation work appears to demonstrate that, with minor calibration and adjustment, the combination of trial reported evidence on short-term outcomes (remission, HSCT and MRD status) and external evidence on their relationship to OS appeared to closely match the OS estimates directly reported within the studies used to generate the evidence sets for CAR T-cell therapy and the comparator (clofarabine).

In the "*Curative intent*" model, a different assumption was employed; specifically that the CAR T-cell therapy itself potentially confers longer-term and potentially curative benefits without the need to bridge to HSCT. In this context, the case for use a structural link between final health benefits and a surrogate outcome or process such as HSCT appears more limited. Instead, a simple three state partitioned survival model was developed to model long-term outcomes via the direct extrapolation of overall survival data from the evidence sets. An important consideration within this model was whether the use of conventional parametric survival functions (e.g. exponential, Weibull, log-normal etc) would adequately capture the potential for a less conventional hazard function that might be observed for a curative treatment; and how this might be affected by different levels of precision and maturity of evidence. Consequently, our work considered the goodness of fit of conventional survival functions and more flexible survival models (e.g. spline-based models developed by Royston and Parmar[141]). A key finding was that the more flexible survival models appeared to more closely approximate the observed hazard function across each of the evidence sets. To our knowledge, although the use of these more flexible survival models are briefly discussed within existing NICE technical support documents (TSD14),[153] we are not aware of any examples of their use to date by manufacturers or AGs within the NICE TA process. Consequently, further research may be required to more formally consider the appropriateness of

alternative survival modelling approaches to regenerative medicines and cell-based therapies, including more flexible models and cure fraction models.[154]

The importance of the level of data maturity in deriving robust survival projections for the economic model was evident in our results. Whilst the "best-fitting" spline models appeared to generate a robust fit to the data over the first 3 months of the KM estimate used in the minimum dataset, the functions were not able to accurately predict the tail of distribution. Furthermore, considerable variation was evident in the predicted long-term survival of the modelled cohort with a significant spread in the projected survival trajectories employing different parametric functions. Consequently, we concluded that it was unlikely that a single survival distribution could adequately characterise uncertainties over the longer-term extrapolation period. Although the robustness of the ICER estimates to alternative distributions can be explored separate sensitivity analyses or scenarios, concerns may exist regarding the transparency of subsequent decisions if the weighting of these is not explicitly specified in subsequent policy decisions.

To more formally account for the uncertainty surrounding choice of survival distribution, a model averaging approach was adopted. This technique involves the parameterisation of uncertainty surrounding the choice of distribution, combining results from a series of alternative survival functions as part of a weighted distribution. This approach samples both the parametric uncertainty associated within each distribution and the uncertainty (or weights) surrounding the choice of preferred method. Through the probabilistic analysis, it is therefore possible to estimate the joint distribution of uncertainty around the parameter estimates and the choice of survival function.

In contrast to the minimum set, the additional data maturity in the intermediate and mature evidence sets results in greater certainty over the long-term survival benefits of treatment. This leads to reduced variability in the potential trajectories for the survival benefits of treatment. In addition, with more mature evidence, the fitted survival models are better able to predict the tail of the KM. Therefore, unlike the bridge to HSCT model, additional evidence maturity in the curative model leads to different projections of survival benefit, as well as impacting on the parametric uncertainty surrounding model extrapolations. The weights in the exemplar model were based on standard measures of statistical fit. However, these weights could also be informed by clinical judgement and the committee's deliberations.

Given the inevitable uncertainties which are likely to exist regarding the longer-term benefits of regenerative medicines and cell-based therapies and their implications for the robustness of subsequent cost-effectiveness estimates, further methodological research could be usefully undertaken to help inform

how these uncertainties might be appropriately quantified in a transparent manner to inform subsequent decisions. A key consideration here would be the extent to which these weights can be defined prior to the Committee's deliberations or should be more directly informed by them. Given the potential complexity in both undertaking these analyses and communicating the results, more efforts should be made to ensure models are developed to ensure that informal judgements can be more explicitly incorporated in a timely and transparent manner.[155]

A key assumption employed within both models is that from year 5 onwards in the model, all patients who remained alive were assumed to experience a similar mortality risk profile consistent with a long-term survivor of ALL. Hence, the mortality risks assumed in both models after year 5 were based on matched general population estimates of the all cause risk of mortality adjusted for excess morbidity and mortality reported in longer-term cohorts of long-term survivors of ALL. Since data were not assumed to be available beyond 5-years, it is not possible to determine the possible direction and/or magnitude of any possible bias that this approach might introduce. However, this period is consistently utilised within existing studies of ALL and appears clinically to represent an important time point for patients to reach without subsequent relapse. Hence, for the purposes of extrapolation and the exemplar, it was considered a reasonable basis for informing subsequent longer-term extrapolations. This assumption also impacted on reducing some of the longer-term uncertainties that would inevitably arise from the extrapolation of the data beyond the maximum reported follow-up across the evidence sets considered for CAR T-cell therapies. Clearly if additional follow-up data were available then the validity of such an approach could be more formally considered and any claims of longer-term benefits could be more robustly substantiated.

Our searches to inform other model parameters identified other important uncertainties. The existing HRQoL data in ALL was limited and several assumptions were required. Importantly, no existing CAR T-study had incorporated measures of HRQoL that could be considered directly in the model. In the absence of this data, assumptions were made based on external studies to account for the possible magnitude of HRQoL benefits of achieving remission, alongside any negative impacts due to the model of therapy (i.e. HSCT, chemotherapy) and other specific adverse events. Our model focused specifically on the impact of CRS and B-cell aplasia. Importantly, no studies were identified on the potential HRQoL impact of these specific events which are likely to be associated with CAR-T therapy necessitating the use of potentially arbitrary assumption. Further research to generate more robust estimates of HRQoL appropriate for cost-effectiveness analysis are clearly required, together with more specific research which more formally demonstrates the impact of specific therapeutic modalities (including CAR T-cells).

Finally, our research also identified important uncertainties regarding both the likely acquisition costs of CAR T-cells and other key elements of the process (e.g. leukapheresis, conditioning therapies, level of hospitalisation required for different aspects such as conditioning, subsequent administration and monitoring etc). Furthermore, no account was taken of the potential costs incurred by patients and their families. Based on prior NICE TA appraisals, additional evidence would need to be provided by manufacturers to more robustly determine the potential costs to the NHS in order to avoid similar uncertainties regarding the costing assumptions to be raised. An important uncertainty identified related to the costs of HSCT and any additional costs that may arise due to longer-term management of patients. A variety of possible sources were identified in our review and important differences observed across these. Further studies would be useful to more formally cost the short and longer-term implications of HSCT in paediatric populations and to also determine the generalisability of studies reporting estimates from outside the UK.

Although the existence of possible learning curves was identified as an important issue in the conceptual review, these were not directly considered within the exemplar. Some aspects of these may become more apparent as larger studies report, particularly involving centres with different levels of expertise. Hence, some aspects of learning may be reflected within the results from larger studies and/or specific factors may become more apparent in terms of how these might be incorporated within cost-effectiveness assessments. For example, as experience with using CAR T-cell therapies develop, this may have important implications for both the identification and management of potential AEs, as well as provision of the therapy itself. An assumption is made in the exemplar model is that the different stages of the process for CAR T-cells would require separate hospitalisation (i.e. for the initial conditioning therapy and later for the subsequent administration of the CAR T-cells and subsequent monitoring). However, as experience and knowledge continues to develop, aspects of the process may evolve over time such that subsequent administration and monitoring may be undertaken in a less resource intensive setting. Although the existence of learning curves has received significant attention in the clinical literature, to date the implication for and application within cost-effectiveness analysis remains limited and warrants further investigation.[92]

Finally, an important assumption made within the exemplar relates to the acquisition cost of CAR T-cell therapy itself. In the absence of a commercially available product and published price, an assumption was made that the manufacturer would employ a value-based approach to their decision such that the resulting cost-effectiveness (ICER) estimate was close to NICE's cost-effectiveness threshold. In the context of the

exemplar, this was assumed to be based on the maximum range of the threshold considered by NICE assuming the existing End of Life criteria are met. Importantly, this price is not considered to be indicative of the final acquisition cost that might be set when commercially available products are available. Neither are we making the presumption that NICE's current End of Life criterion would apply. Rather, the basis for setting the price on the basis of existing cost-effectiveness threshold was to enable different interested parties to better understand the potential impact of other uncertainties (e.g. precision and maturity of evidence) within NICE's current decision making process, identifying potential trade-offs that may exist and illustrating how these uncertainties might be more explicitly addressed within different MEAs (i.e. evidence generation and/or pricing schemes). Although it is clearly possible to examine a range of different possible prices for the CAR T-cell therapies within the exemplar, it was considered that this approach may result in the subsequent Panel decision process becoming unmanageable (i.e. multiple pricing scenarios) and would lessen the generalisability learning which the exemplar was developed to highlight.

9 Assessment of cost-effectiveness, uncertainty and the value of alternative policy options

9.1 Overview

The exemplar in Section 8 has been developed to highlight some of the specific challenges that may present themselves to manufacturers and AGs in terms of developing and populating a cost-effectiveness model. Consideration is now given to how such estimates could be presented and communicated to the Committee. In doing this we consider the analyses routinely requested within NICE's existing methods guide[83] but also consider whether additional analyses may provide useful additional insights to help inform subsequent committee deliberations.

Based on the scoping review reported in Section 4.4, we also consider analyses relating to some of the broader issues and approaches identified previously (e.g. alternative payment mechanisms) which, although potentially outside the existing remit of NICE, may provide additional insights to other interested bodies and manufacturers.

Importantly, the use of non-reference case approaches and additional analyses undertaken beyond those requested within NICE's existing process and methods guide are not intended to be prescriptive. Neither are they comprehensive given the multiplicity of issues and challenges raises. Instead these have been provided to help explore whether additional information and analyses may be helpful in informing the Committee's deliberations and the nature of such analyses.

Consideration will be subsequently given regarding whether particular analyses helped inform particular considerations within NICE's deliberations within the exemplar appraisal and to identify areas where further methodological and applied work may be required.

9.2 Acquisition costs of CAR T-cell therapies

As noted in Section 8.3.1.1 of the report, the acquisition cost for CAR T-cell therapy in the exemplar was assumed to be based on a value-based approach from the manufacturer, such that this would be priced at a level such that the ICER for CAR-T cell therapy would be close to the upper limit of NICE's end of life threshold range (circa £50,000 per QALY gained). Because of differences in the projected survival benefits of treatment across the separate TPPs, the subsequent cost of CAR T-therapy varied across these, with one-off acquisition costs of £356,100 assumed in the "Bridge to HSCT" scenario and £528,600 in the "curative intent" scenario.

A full summary of CAR T acquisition costs assumed across the separate pricing scenarios described in subsequent sections (one-off fixed cost, monthly leasing price, discounted list price via PAS) is provided in Table 25.

Table 25: Estimated acquisition costs for CAR T-cell therapies (excluding costs for conditioning therapy and leukapheresis)

Scenario	**One-off acquisition cost per patient**	**Monthly leasing price**	**Discounted list price (10%)**
Bridge to HSCT	£356,100	£2,756.27	£320,490
Curative intent	£528,600	£3,282.66	£475,740

9.3 Bridge to HSCT TPP

9.3.1 Per-patient analyses – minimum evidence set

The sequence of assessments starts with a conventional assessment of cost-effectiveness at the patient level based on the minimum evidence set reported within the TPP section. Disaggregated costs and outcomes are presented in Table 26.

Table 26: Summary of costs and outcomes

Outcome	CAR T	Standard of care	Incremental
Costs			
Course of treatment (including conditioning)	£358,057	£43,200	£314,857
Hospitalisation for treatment	£13,012	£7,180	£5,832
AE costs	£2,750	£442	£2,308
HSCT and related follow-up costs	£71,918	£21,380	£50,538
Non-HSCT follow-up costs	£3,391	£3,759	-£368
Total costs	£449,128	£75,962	£373,166
QALYs			
Decision tree	0.14	0.11	0.03
Post-HSCT MRD -ve	8.82	0.30	8.52
Post-HSCT MRD +ve	0.00	0.16	-0.16
Post-HSCT no remission	0.00	0.72	-0.72
No-HSCT remission	0.06	0.03	0.03
No-HSCT no-remission	0.07	0.11	-0.04
QALY loss due to HSCT	-0.27	-0.08	-0.19
Total QALY	8.82	1.36	7.46
Total life years	10.60	1.77	8.83
Proportion of patients receiving HSCT (undiscounted)	48%	15%	33%

The mean incremental costs of CAR T-cell therapy over a patient's lifetime was estimated to be £373,166 and resulted in an additional 7.46 QALYs. A summary of the incremental cost per QALY gained (£49,995) over a lifetime horizon are reported in Table 27 which can be compared against the cost-effectiveness threshold. Equivalently this can be also expressed as the per-patient net-health effect (NHE); including benefits, harms and NHS/Personal Social Services costs. The NHE is the difference between any health gained with the intervention and health foregone elsewhere in the health care system and can be expressed in both monetary and QALY terms. With an ICER of approximately £50,000 per QALY, the incremental NHE at a threshold of £50,000 NHE is close to zero (i.e. 0.001 QALYs or £41 per patient).

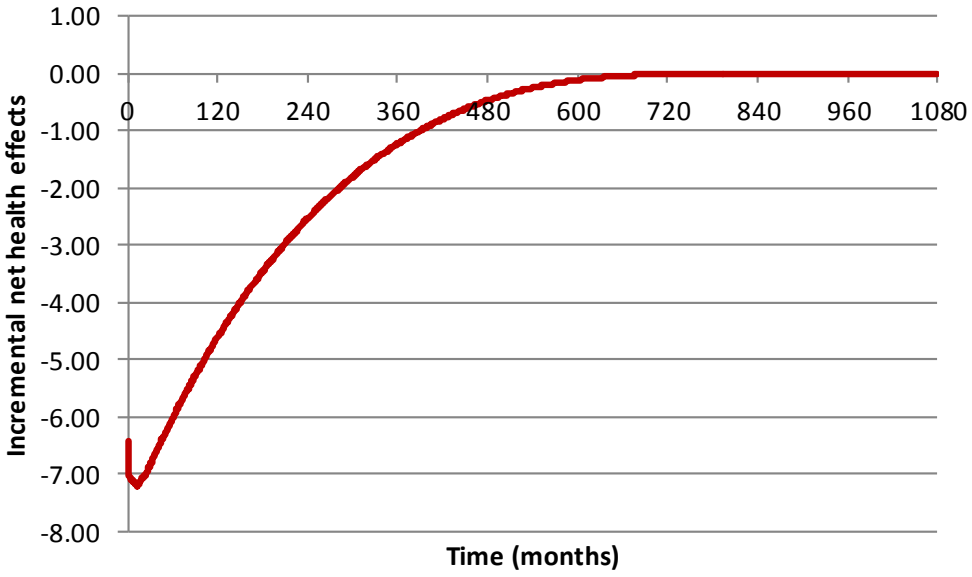
That is, the additional health gained with the intervention is almost exactly offset by health foregone elsewhere.

Table 27: Expected cost-effectiveness of CAR T therapy per patient treated (lifetime horizon)

Per patient-level				Cost-effectiveness threshold of £50,000 per QALY gained	
Treatment	Costs	QALYs	ICER	NHE, QALY (£)	Incremental NHE, QALY (£)
CAR T-cell therapy	£449,128	8.82	£49,995	-0.16 (-£7,919)	0.001 (£41)
Standard of care	£75,962	1.36		-0.16 (-£7,960)	-

Given the uncertainties surrounding longer-term outcomes, it may also be informative to consider how incremental NHEs accumulate over time or the 'investment profile' with CAR T-cell therapy, shown in Figure 14. The initial per-patient cost for CAR T-cell patients is due to the additional acquisition and administration costs of the CAR T-cells and associated HSCT costs. These negative NHEs are gradually offset by positive NHEs in later periods due to the ongoing mortality benefits assumed from successfully bridging to HSCT. However it is only after 60 years that the initial losses are sufficiently compensated by later gains, that CAR T-cells appear to be close to break-even (i.e. NHE>=0).

Figure 14: Investment profile per patient treated with NHE over a lifetime horizon



9.3.2 Population-level analyses – minimum evidence set

NHEs can also be presented for a population of patients over time. Although the presentation of population NHE is not formally requested within the existing NICE methods guide, population-based

analyses are requested to be submitted to assess population impacts within Section 5.12 (Impact on the NHS). In addition, Section 6.4.1. states that in situations in which the evidence of clinical effectiveness evidence is either "*absent, weak or uncertain*" the Committee is requested to "*balance the potential net benefits to current NHS patients of a recommendation not restricted to research with the potential net benefits to both current and future NHS patients of being able to produce guidance and base clinical practice on a more secure evidence base*".

Analyses of population NHE may therefore provide additional information to help inform the Committee's deliberations regarding possible research recommendations (Section 6.4 of the current methods guide). Population NHE requires information about the prevalence and future incidence of the target population and a judgement about the time horizon over which the technology will be used in clinical practice. As outlined in Appendix 8, the expected incidence of eligible cases for the exemplar is estimated to be approximately 38 patients per annum. The technology time horizon is set to 10-years in the base case.

Table 28 reports population NHE for CAR-T therapy over the 10-year technology time horizon. Over this period, the use of CAR-T cell therapy is estimated to generate an additional 2356 QALYs (discounted values) within the population considered compared to the current standard of care. However, since the additional lifetime costs of £117.78 million (£141.75 million - £23.97 million; discounted values) require other treatments to be displaced and health foregone by other patients in the NHS, overall the additional QALYs are exactly offset by health foregone elsewhere. Hence, the incremental population NHE at a £50,000 per QALY threshold is 0.26 QALYs (£12,813).

Table 28: Expected cost-effectiveness of CAR T therapy at population level (including incident patients)

Population-level				Cost-effectiveness threshold of £50,000 per QALY gained	
Treatment	Costs	QALYs	ICER	NHE, QALY (£)	Incremental NHE, QALY (£)
CAR T-cell therapy	£141,751,559	2785.04	£49,995	-49.99 (-£2,499,490)	0.26 (£12,813)
Standard of care	£23,974,719	429.25		-50.25 (-£2,512,303)	

A series of one-way sensitivity analyses were undertaken to assess the sensitivity of model results to changes in assumptions. The results of this sensitivity analysis are presented in Table 29.

Table 29: Results of one-way sensitivity analysis (population level) – Bridge to HSCT TPP

Scenario	Incremental cost	Incremental QALY	ICER	Incremental NHE at willingness to pay of £50,000, QALY (£)
Base case	£117,776,840	2355.79	£49,995	0.26 (£12,813)
Repeat CAR T treatment – monthly probability of 0.5%	£193,649,693	2355.79	£82,201	-1517.20 (-£75,860,040)
Repeat CAR T treatment – monthly probability of 0.1%	£132,951,410	2355.79	£56,436	-303.24 (-£15,161,757)
Discounting – 0% costs and health effects	£117,863,631	4608.43	£25,576	2251.16 (£112,557,826)
Discounting – 6% costs and health effects	£117,718,706	1662.50	£70,808	-691.87 (-£34,593,729)
Discounting – 0% costs and 6% health effects	£117,863,631	1662.50	£70,895	-694.77 (-£34,738,654)
Discounting – 6% costs and 0% health effects	£117,718,706	4608.43	£25,544	2254.06 (£112,702,751)
Discounting – 3.5% costs and 1.5% health effects	£117,776,840	3350.89	£35,148	995.35 (£49,767,620)
UK treasury recommended step discounting 3.5% up to year 30, 3% thereafter (both costs and health effects)	£117,776,840	2374.47	£49,601	18.94 (£946,799)
Standard of care costs based on FLAG-IDA	£130,211,131	2355.79	£55,273	-248.43 (-£12,421,478)
Hazard rate for death in non-remission no-HSCT patients increased from 0.2425 (mean time to death =0.34 years) to 0.6075 (mean time to death = 0.14 years)	£117,775,723	2363.47	£49,832	7.95 (£397,705)

The results of the one-way sensitivity analyses indicate that the results of the evaluation are sensitive to assumptions on the potential for re-treatment of CAR T-cell and the assumed discounting rate for health effects in the model. The results of the evaluation are less sensitive to assumptions on the discounting rate for costs, assumptions on the impact of remission status on survival in non-HSCT patients, and to reducing the cost of standard of care treatment to values consistent with treatment using FLAG-IDA (assuming similar efficacy to clofarabine).

If the committee were to consider the criteria met for applying the non-reference-case discount rate of 1.5% for both costs and health effects (i.e. when treatment restores people who would otherwise die or have a very severely impaired life to full or near full health, and when this is sustained over a very long

period, normally at least 30 years), then the ICER would reduce to £35,148 per QALY and CAR T-cell therapy would be associated with additional population NHE equivalent to 995 QALYs (£49.77 million) in comparison to health foregone elsewhere.

Employing the stepwise discounting recommended by the UK treasury to all public sector bodies, makes only a small difference to the ICER results with an ICER of £49,601. The incremental population NHE increases to 18.94 QALYs (£946,799).

Although the results of the evaluation appear sensitive to assumptions on the potential for re-treatment of CAR T-cell therapy, this was not considered to represent such a challenge in this TPP. CAR T-cells were assumed to be used as a one-off therapy to induce remission and to improve the likelihood and outcomes of HSCT. It was assumed that patients would not receive a repeat treatment in the event of not achieving remission, nor would patients who were successfully treated with HSCT receive further treatments with CAR T-cell therapy.

Probabilistic analysis

The results of the probabilistic analysis therapy are shown in Table 30.

Table 30: Results of base case probabilistic analysis, presented at population-level (Bridge to HSCT)

		Population-level		Cost-effectiveness threshold of £50,000 per QALY gained			
Treatment	E[Costs]	E[QALYs]	ICER	E[NHE], QALY (£)	Incremental NHE, QALY (£)	Probability cost-effective	Consequences of decision uncertainty, QALY (£)
CAR T-cell therapy	£141,556,652	2716.4	£55,090	-114.8 (-£5,738,274)	-215.9 (-£10,794,902)	26.1%	56.3 (£2,813,197)
Standard of care	£24,728,297	595.7		101.1 (£5,056,627)			

The probabilistic ICER increased to £55,090 due to the model non-linearities. Consequently, population NHE were now negative with an overall loss to the health system of 215.9 QALYs (£10.79 million). The cost-effectiveness acceptability planes and curves are presented in Figure 15 and Figure 16. At a £50,000 cost-effectiveness threshold, the probability that CAR T-cell therapy was the most cost-effective option was 26.1%.

Figure 15: Cost-effectiveness Acceptability plane: **Bridge to HSCT TPP– minimum evidence set**

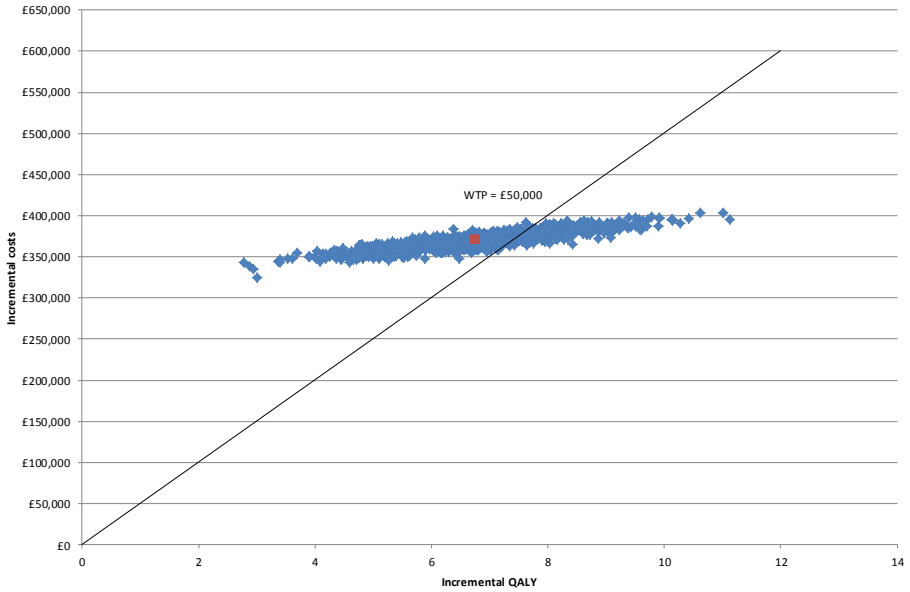
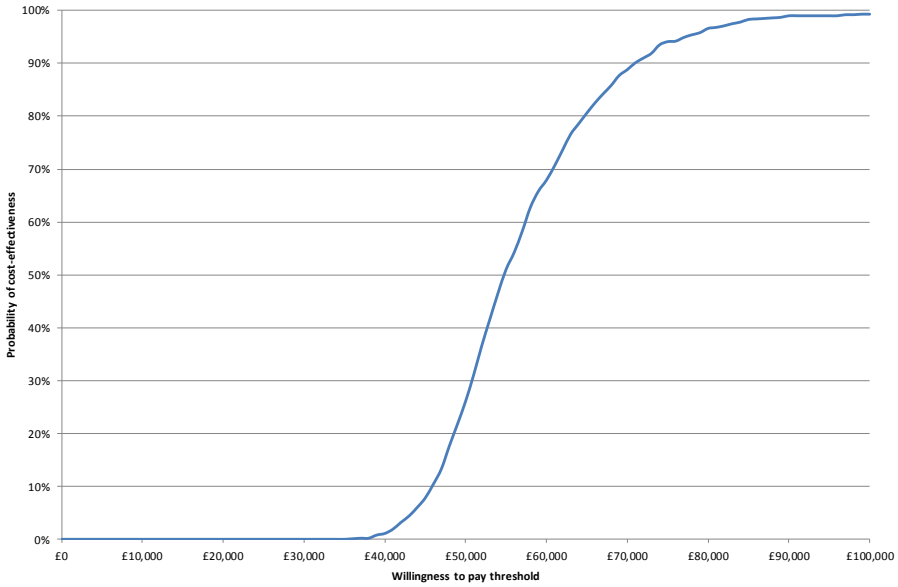


Figure 16: Cost-effectiveness acceptability curve: Bridge to HSCT TPP– minimum evidence set



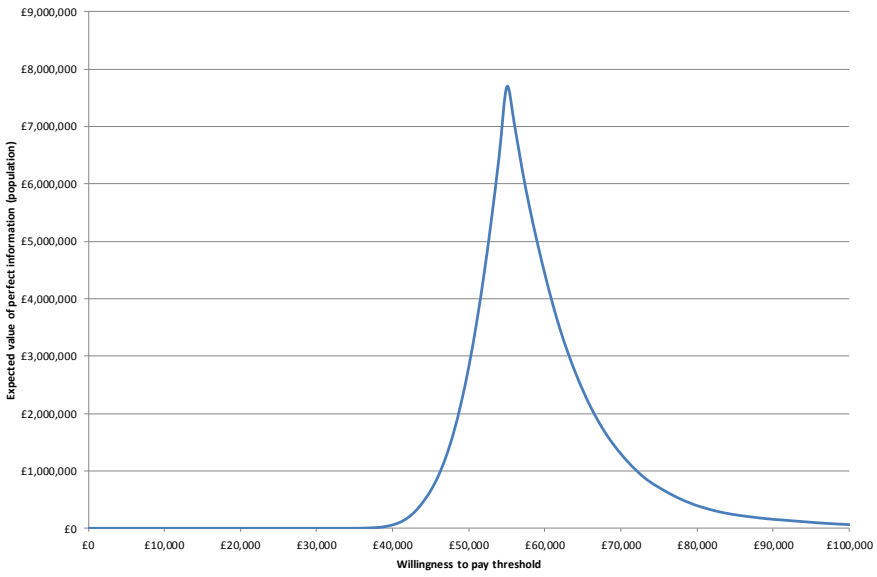
In addition to considering how uncertain a decision is to approve or reject a technology based on expected cost-effectiveness, an assessment of the scale of the likely consequences may also be potentially informative to the committee, particularly in deliberations related to possible research recommendations. An assessment of the potential consequences of uncertainty is important because it indicates the scale of the population NHEs that could be gained if uncertainty surrounding this decision could be resolved

immediately.[88] This estimate also represents an expected upper bound to the benefits of more research. This may help inform subsequent research recommendations. For example, if the maximum potential benefits of further research are considered unlikely to sufficiently justify the research costs, then it may not be worthwhile to issue further research recommendations.

These same consequences are referred to using the term Payer Uncertainty Burden (PUB) in the draft DSU report on Managed Access. Elsewhere in the literature, these have been defined as the Expected Value of Perfect Information (EVPI) and the overall Expected Opportunity Loss. Within the DSU report this is further defined as value of the risk of making a particular decision due to uncertainty (expressed in either monetary or health units); combining two key concepts: first, the probability that the strategy with the highest expected NHE may not be the optimal strategy (i.e.1-probability the intervention is cost-effective based on the probabilistic results), and second, the consequences of a 'wrong' decision in terms of QALYs and NHS costs that could have been saved if the truly optimal strategy had been selected instead.

Assuming a 10-year technology horizon, the consequences of decision uncertainty in the minimum evidence set are estimated to be 56.3 QALYs (£2.83 million). Figure 17 shows how the scale of the consequences of decision uncertainty varies across different cost-effectiveness thresholds, reaching a peak at a £55,000 threshold.

Figure 17: Consequences of decision uncertainty – Bridge to HSCT TPP (minimum evidence set)



A summary of the population level incremental net health effects, net monetary benefits, probability of cost-effectiveness and consequences of decision uncertainty across a range of willingness to pay thresholds is presented in Table 31.

Table 31: Consequences of decision uncertainty across different thresholds - Bridge to HSCT TPP (minimum evidence set)

Cost-effectiveness threshold	Incremental NHE, QALY	Incremental NMB, £	Probability cost-effective	Consequences of decision uncertainty, QALY (£)
£20,000	-3720.75	-£74,414,973	0%	0 (£0)
£30,000	-1773.61	-£53,208,283	0%	0 (£0)
£50,000	-215.90	-£10,794,902	26.1%	56.3 (£2,813,197)
£75,000	562.96	£42,221,825	94.1%	9.5 (£710,894)
£100,000	952.39	£95,238,551	99.3%	0.6 (£63,592)

At conventional cost-effectiveness thresholds of between £20,000 and £30,000 per QALY gained, the probability that CAR T-cell therapy is cost-effective versus standard of care is 0%. Consequently, because of the high certainty that CAR T-cell therapy is not cost-effective at conventional thresholds (i.e. assuming end of life criteria do not apply) there are no consequences of decision uncertainty. At a cost-effectiveness threshold of £50,000 per QALY gained, the probability that CAR T-cell therapy is cost-effective versus standard of care is 26.1%. In this case, the expected population health consequence of decision uncertainty is 56.3 QALYs (10-years). The corresponding expected monetary cost of decision uncertainty is approximately £2.8 million. At thresholds of £75,000 and £100,000 per QALY gained the probability that CAR T is cost-effective increases to over 94%. Because there is now high certainty that CAR T is cost-effective at these thresholds, the corresponding consequences of decision uncertainty reduces to fewer than 10 QALYs (or less than £1 million in monetary terms).

Alternative pricing scenarios

A series of alternative pricing schemes have been generated to explore their potential impact on cost-effectiveness and decision uncertainty. The schemes considered include:

- A leasing scheme approach based on the approach outlined by Edlin et al.[95] In this scenario, the technology is assumed to be leased from the company. The monthly 'lease' payment was established by calculating a stream of payments over the expected survival duration of the patients that has the same expected net present value as the agreed price. Hence, payment was assumed to continue on a monthly basis while a patient remained alive.

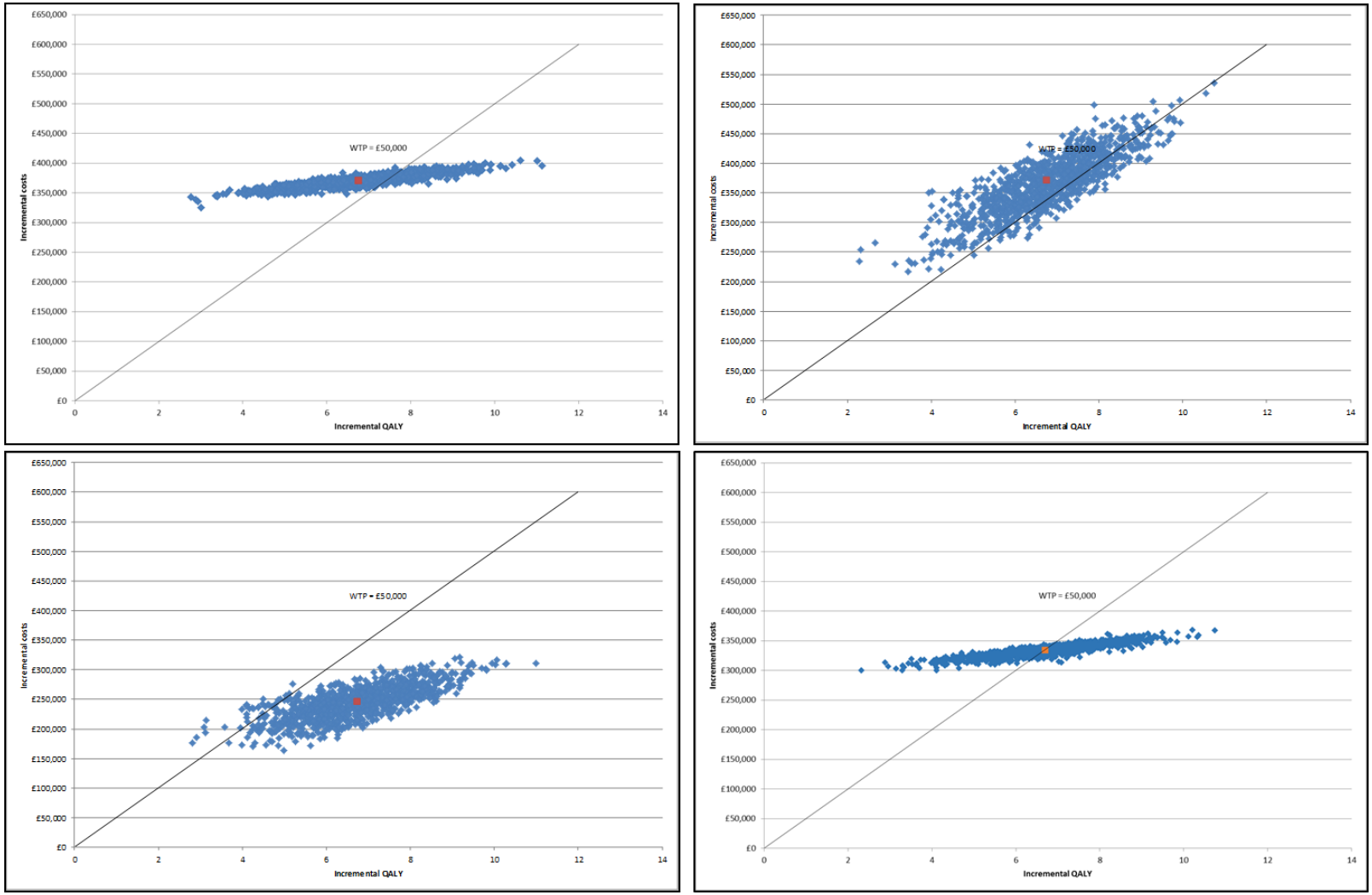
- A pay for performance scheme in which payment is made retrospectively only for patients who achieve remission (CR) within a specified period (e.g. 28 days). Alternatively, an initial upfront payment could be made to all with a separate 'clawback' agreed for patients who do not achieve remission.
- A more conventional PAS scheme providing a fixed percentage discount (e.g. 10%).

The probabilistic results based on alternative hypothetical pricing scenarios are shown in Table 32. The scatter plots showing each iteration of incremental costs and incremental effects considered in the PSA is provided in Figure 18.

Table 32: Impact of different pricing schemes on the cost-effectiveness of CAR T therapy, and the associated consequences of decision uncertainty: Bridge to HSCT TPP (minimum evidence set)

Pricing scenario	Population-level				Cost-effectiveness threshold of £50,000 per QALY gained			
	Treatment	E[Costs]	E[QALYs]	ICER	E[NHE], QALY (£)	Incremental NHE, QALY (£)	Probability cost-effective	Consequences of decision uncertainty, QALY (£)
Base case	CAR T-cell therapy	£141,556,652	2716.4	**£55,090**	-114.77 (-£5,738,274)	-215.9 (-£10,794,902)	26.1%	**56.3 (£2,813,197)**
	Standard of care	£24,728,297	595.7		101.13 (£5,056,627)			
Leasing method	CAR T-cell therapy	£140,082,600	2727.04	**£54,227**	-74.61 (-£3,730,478)	-179.94 (-£8,997,139)	22.1%	**22.5 (£1,123,900)**
	Standard of care	£24,678,802	598.91		105.33 (£5,266,662)			
Payment for remission patients only (average of 70%)	CAR T-cell therapy	£102,099,708	2708.82	**£36,430**	666.83 (£33,341,351)	577.2 (-£28,861,808)	96.8%	**3.9 (£195,152)**
	Standard of care	£24,614,048	581.87		89.59 (£4,479,543)			
Fixed pricing discount (10%)	CAR T-cell therapy	£130,229,928	2707.12	**£49,857**	102.52 (£5,125,971)	6.05 (302,586)	51.8%	**131.2 (£6,558,209)**
	Standard of care	£24,818,199	592.83		96.47 (£4,823,385)			

Figure 18: Scatter plots of incremental costs and incremental effectiveness across the four pricing scenarios (upper left – base case, upper right –

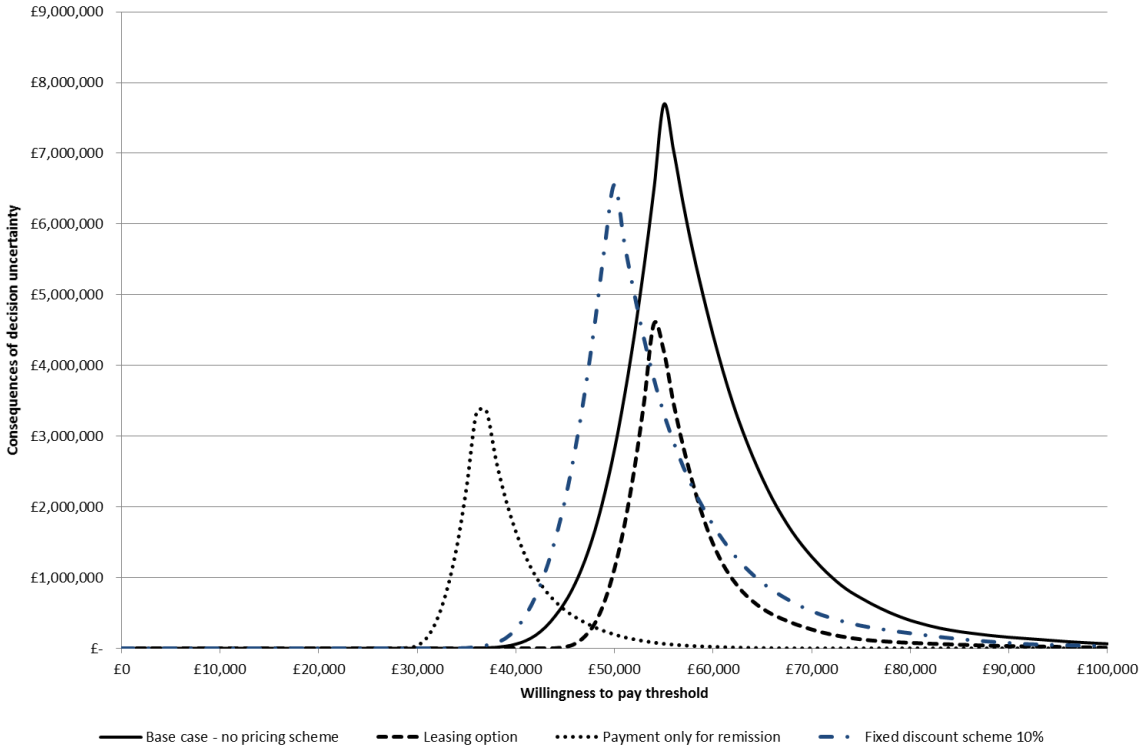


Leasing method, lower left – payment for remission only, lower right – fixed discounting; Red Box = mean of PSA values)

The impact of the different pricing schemes on the sampled outputs of the PSA is shown graphically in Figure 18. In the base case (fixed cost for CAR T), the cloud of simulated outcomes from the PSA is flat, such that there is considerable variability around the QALY gains of treatment, but little relative variability around the incremental costs. By introducing a leasing method, the costs of CAR T-cell therapy becomes more closely linked to the effectiveness of treatment, such that the cloud of simulated outcomes from the PSA is re-orientated around the willingness to pay threshold. With both the remission and discounted schemes, the cost-effectiveness of treatment is improved, and the cloud of simulated outcomes is shifted downwards on the chart.

A comparison plot of the consequences of decision uncertainty across the alternative pricing scenarios for different cost-effectiveness thresholds is shown in Figure 19.

Figure 19: Comparison plot of consequences of decision uncertainty across alternative pricing schemes: Bridge to HSCT TPP (minimum evidence set)



Under a fixed one-off acquisition cost approach, assumed in the main analyses, the NHS bears all the risks associated with uncertainty surrounding whether the expected benefits of therapy will be realised in routine clinical practice. Hence, the consequences of decision uncertainty to the NHS appear highest with this scheme (56.2 QALYs; £2.81 million). The alternative schemes result in reductions in decision uncertainty and associated consequences to the NHS. However, the impact and mechanism in which this is achieved differs across the separate approaches.

The leasing approach results in only a minor difference in the ICER. Similar levels of decision uncertainty also remain (i.e. the probability that the intervention is cost-effective is similar to that under a fixed one-off acquisition cost approach). However, the scale of the consequences of the uncertainty to the NHS is significantly reduced via this scheme. This scheme limits the risk to the NHS for over-paying for a technology which doesn't achieve the expected outcomes, significantly lowering the consequences of decision uncertainty to 22.5 QALYs (£1.12 million).

The use of a pay for performance scheme improves the expected cost-effectiveness and as a result reduces both the level of decision uncertainty and the scale of their consequences. Restricting payment to only patients who achieve remission, improves expected cost-effectiveness (£36,430 per QALY), leading to a higher probability of being cost-effective (96.8%), thereby reducing the consequences of uncertainty to 3.9 QALYs (£195,000). The use of a more conventional PAS scheme, based on an assumed 10% reduction in the acquisition cost, works in a similar manner by improving both expected cost-effectiveness (£49,857 per QALY) and the likelihood a treatment is cost-effective (51.8%), however, as the ICER now lies closer to the threshold in absolute terms the consequences are increased to 131.2 QALYs (£6,558,209).

The comparison plot more clearly shows the impact of the alternative pricing scheme. The alternative schemes affect both the shape of the distribution of the consequences across the separate cost-effectiveness thresholds as well as their position.

Results of probabilistic analysis across evidence sets

All the analyses reported previously have been based on the minimum evidence set. The impact of the alternative evidence sets on expected cost-effectiveness, the level of decision uncertainty and the scale of the consequences are reported in Table 33.

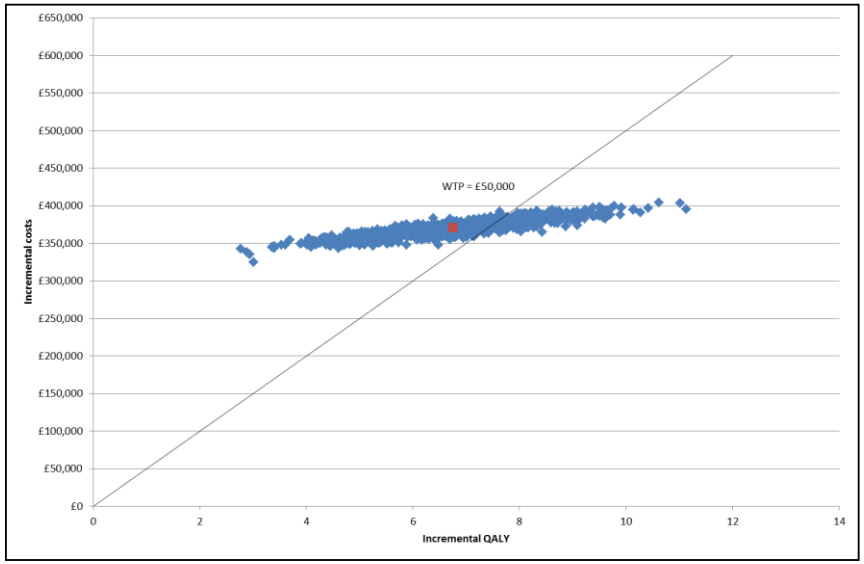
As highlighted in Section 6, the use of a separate structural/surrogate link within the 'bridge to HSCT' TPP was employed to allow the incorporation of external evidence on the relationship between remission MRD and HSCT status. A limitation of our analysis is that the same external evidence is then used across each of the separate evidence sets. This means that the additional follow-up assumed in both the intermediate and mature evidence sets are not adequately reflected in the results. Consequently, the ICER and associated decision uncertainty are identical across the minimum and intermediate datasets. Furthermore, the differences in results based on these evidence sets and the mature evidence set is driven entirely by the increased precision (i.e. due to higher patient numbers) in the short-term remission, MRD and HSCT rates, as opposed to the additional maturity of follow-up data which may be available.

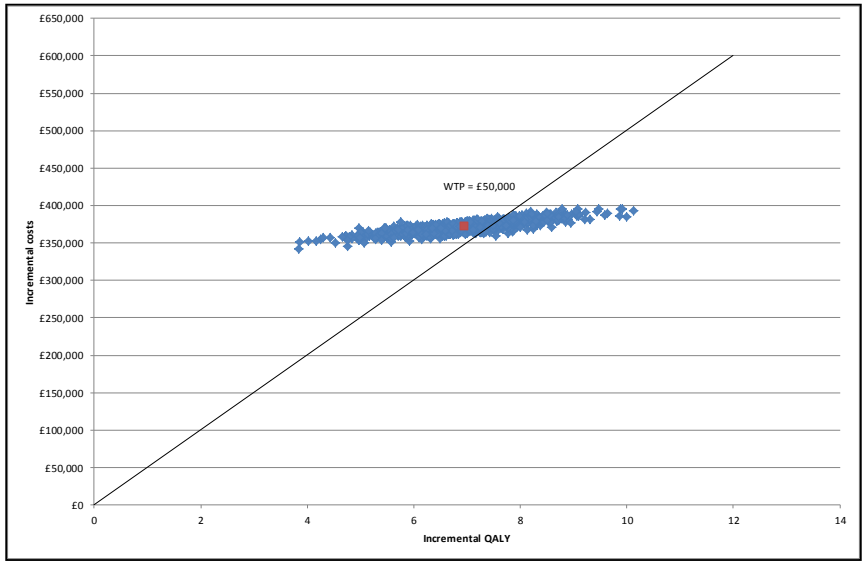
In practice, the additional follow-up reported in more mature follow-up could either replace the existing surrogate relationship employed here or be synthesised and combined with the external evidence. Hence, the value that the additional follow-up brings in terms of either confirming an assumed surrogate relationship, or increasing the precision around this relationship, are not adequately captured in these analyses.

Table 33: Probabilistic results comparing across evidence sets: Bridge to HSCT TPP

Evidence set	Population-level				Cost-effectiveness threshold of £50,000 per QALY gained			
	Treatment	E[Costs]	E[QALYs]	ICER	E[NHE], QALY (£)	Incremental NHE, QALY (£)	Probability cost-effective	Consequences of decision uncertainty, QALY (£)
Minimum (Base case)	CAR T-cell therapy	£141,556,652	2716.4	£55,090	-114.77 (-£5,738,274)	-215.9 (-£10,794,902)	26.1%	56.3 (£2,813,197)
	Standard of care (clofarabine treatment)	£24,728,297	595.7		101.13 (£5,056,627)			
Intermediate	CAR T-cell therapy	£141,556,652	2716.4	£55,090	-114.77 (-£5,738,274)	-215.9 (-£10,794,902)	26.1%	56.3 (£2,813,197)
	Standard of care (clofarabine treatment)	£24,728,297	595.7		101.13 (£5,056,627)			
Mature	CAR T-cell therapy	£141,680,276	2764.76	£53,462	-68.84 (-£3,442,181)	-151.92 (-£7,595,782)	28.1%	48.1 (£2,406,886)
	Standard of care (clofarabine treatment)	£24,375,545	570.58		83.07 (£4,153,600)			

Figure 20: Scatter plots of incremental costs and incremental effectiveness across evidence sets (upper left – minimum and intermediate evidence sets, upper right –mature evidence set)



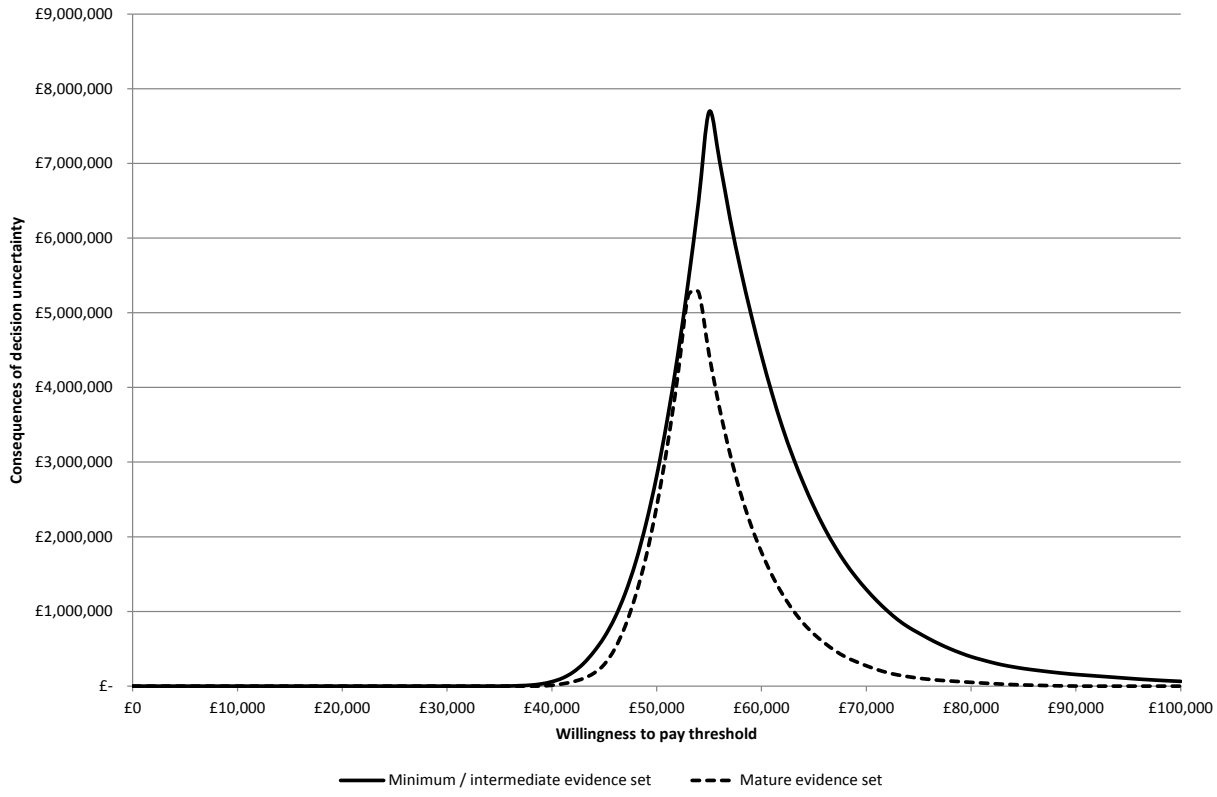


Despite these limitations, the separate evidence sets may still provide an important comparison for the Committee to consider. Specifically in relation to how their deliberations might be affected in situations in which the same ICER and decision uncertainty were reported but under different circumstances i.e. situations in which the results are based entirely on external surrogate relationships versus when these are based on actual observed data from a longer-term trial or follow-up.

As expected, the health consequence of decision uncertainty in the mature evidence set (48.1 QALYs; £2.41 million) is lower than that reported in the minimum set (56.2 QALYs; £2.81 million), at a threshold of £50,000 per QALY gained. These consequences are reduced by the increased precision associated with the larger sample in terms of the short term remission, MRD and HSCT rates.

A comparison plot of the consequences of decision uncertainty between the minimum/intermediate and mature evidence sets across a range of cost-effectiveness thresholds in Figure 21.

Figure 21: Comparison plot of consequences of decision uncertainty across TPPs



Presentation of the scale of consequences using population NHE allows some important comparisons to be made across the separate pricing approaches and the difference evidence sets. More specifically these comparisons could provide a more explicit basis for considering the value of direct price reductions that might be realised via a conventional PAS (or less conventional schemes which work by indirectly lower the effective price) compared to the provision of additional evidence (both precision and maturity), in terms of reducing decision uncertainty and its consequences.

In the 'Bridge to HSCT' TPP, significant reductions in the level and scale of the consequences of decision uncertainty (i.e. the risk faced by the NHS), appear to be achieved by more innovative pricing approaches such as pay for performance and leasing approaches than that which might be realised by the provision of further evidence. Such information might provide an important basis for discussions between manufacturers and NICE in terms of how the existing uncertainties that exist might be appropriately managed ensuring risks and benefits are more appropriately shared.

9.4 Curative intent TPP

A similar sequence of assessments and analyses were conducted based on the curative intent TPP. In contrast to the 'Bridge to HSCT', differences in the results across the evidence sets are more evident since the results are directly informed by the data assumed within these rather than employing evidence from external sources.

9.4.1 Per-patient analyses – minimum evidence set

Again, the sequence of assessments starts with a conventional assessment of cost-effectiveness at the patient level based on the minimum evidence set. Disaggregated costs and outcomes are presented in Table 34. The mean incremental costs of CAR T-cell therapy over an individual patient's lifetime was estimated to be £503,256 and resulted in an additional 10.07 QALYs.

Table 34: Summary of costs and outcomes in the Curative intent TPP (minimum evidence set)

Outcome	CAR T	Standard of care	Incremental
Costs			
Course of treatment (including conditioning)	£530,557	£43,200	£487,357
Hospitalisation for treatment	£13,012	£7,180	£5,832
AE costs	£20,513	£442	£20,070
HSCT and related follow-up costs	£15,092	£22,267	-£7,175
Non-HSCT follow-up costs	£4,189	£7,016	-£2,827
Total costs	£583,362	£80,106	£503,256
QALY			
EF	10.62	0.83	9.79
Recurrent disease	0.62	0.37	0.25
AE	0.00	0.00	0.00
QALY loss due to HSCT	-0.06	-0.08	0.03
Total QALY	11.18	1.11	10.07
Total life years	13.42	1.47	11.95

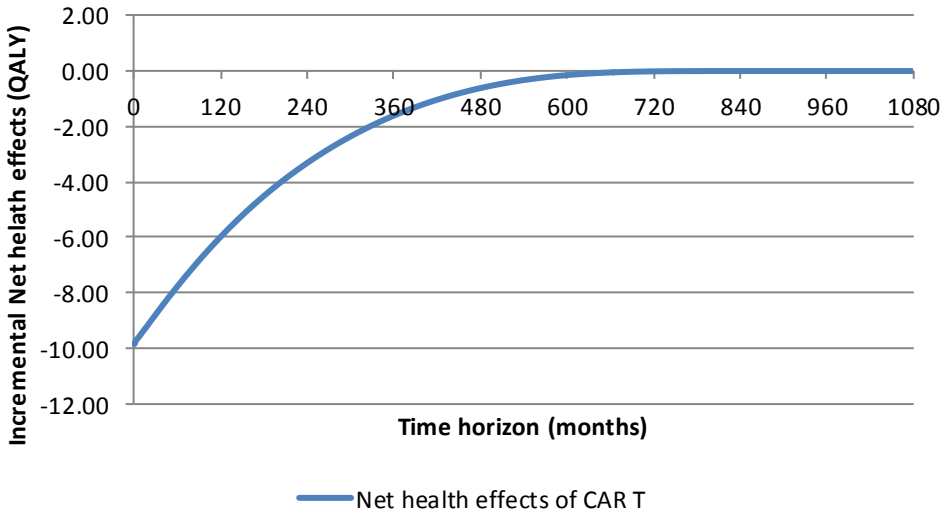
The expected cost-effectiveness of CAR T-cell therapy and per-patient NHE is shown in Table 35. In common with the previous TPP, the acquisition cost was set such that the ICER (£49,994) was close to the upper limit of NICE's end of life threshold range (circa £50,000 per QALY gained).

Table 35: Per patient expected cost-effectiveness: Curative intent TPP (minimum evidence set)

Per patient-level				Cost-effectiveness threshold of £50,000 per QALY gained	
Treatment	Costs	QALYs	ICER	NHE, QALY (£)	Incremental NHE, QALY (£)
CAR T-cell therapy	£583,362	11.18	£49,994	-0.49 (-£24,509)	0.001 (£61)
Standard of care	£80,106	1.11		-0.49 (-£24,570)	-

The accumulation of NHEs over time or equivalently the 'investment profile' per patient is shown in Figure 22.

Figure 22: Investment profile – Curative intent TPP



At the start of the time horizon, the initial high costs of treatment are far in excess of the immediate health benefits of treatment, leading to a negative NHE. Over time, the initial negative NHEs are gradually offset by the accrual of the residual health benefits of treatment (i.e. of cure). In common with the 'Bridge to HSCT' TPP, it is only after approximately 60 years that the initial losses are sufficiently compensated by later gains that CAR T-cells appear to be close to break-even (i.e. NHE>=0).

The shape of the investment profile differs slightly across the separate TPPs, the early kink that was shown in the previous TPP is not evident here. The lack of the kink is due to the small number of patients who are assumed to receive HSCT in the curative TPP. Hence, the resulting investment profile is smoother, although higher initial negative NHEs are reported due to the higher acquisition cost assumed within this TPP.

9.4.2 Population-level analyses – minimum evidence set

The expected per patient effects of treatment are also extended to a population level based on similar assumptions concerning the incidence (approx. 38 patients per annum) and technology time horizon (10-years).

Table 36 reports population NHE for CAR-T cell therapy over the 10-year technology time horizon. Over this period, the use of CAR-T cell therapy is estimated to result in an additional 3,177 QALYs (discounted values) within the population considered compared to the current standard of care. However, since the additional lifetime costs of £158.84 million (discounted values) require other treatments to be displaced and health foregone by other patients, overall the additional QALYs are almost exactly offset by health foregone elsewhere. The resulting incremental population NHE is 0.39 QALYs expressed in health terms and £19,269 in monetary terms.

Table 36: Expected cost-effectiveness of CAR T therapy (population level): Curative intent TPP (minimum evidence set)

Population-level				Cost-effectiveness threshold of £50,000 per QALY gained	
Treatment	Costs	QALYs	ICER	NHE, QALY (£)	Incremental NHE, QALY (£)
CAR T-cell therapy	£184,117,952	3527.65	£49,994	-154.71 (-£7,735,298)	0.39 (£19,269)
Standard of care	£25,282,579	350.56		-155.09 (-£7,754,567)	

A series of one-way sensitivity analyses were conducted to assess the sensitivity of model results to changes in assumptions or model settings. The results of this sensitivity analysis are presented in Table 37.

Table 37: Results of one-way sensitivity analysis at population level: Curative intent TPP (minimum evidence set)

Scenario	Incremental cost	Incremental QALY	ICER	Incremental NHE, QALY (£)
Base case	£158,835,372	3177.09	£49,994	0.39 (£19,269)
Repeat CAR T treatment – monthly probability of 1%	£429,511,483	3177.09	£135,190	-5413.14 (-£270,656,842)
Repeat CAR T treatment – monthly probability of 0.5%	£294,173,428	3177.09	£92,592	-2706.38 (-£135,318,786)
Repeat CAR T treatment – monthly probability of 0.1%	£185,902,983	3177.09	£58,514	-540.97 (-£27,048,342)
Discounting – 0% costs and health effects	£160,095,703	6127.15	£26,129	2925.24 (£146,261,823)
Discounting – 6% costs and health effects	£158,456,968	2272.68	£69,723	-896.46 (-£44,823,167)
Discounting – 0% costs and 6% health effects	£160,095,703	2272.68	£70,444	-929.24 (-£46,461,901)
Discounting – 6% costs and 0% health effects	£158,456,968	6127.15	£25,861	2958.01 (£147,900,557)
Discounting – 3.5% costs and 1.5% health effects	£158,835,372	4478.47	£35,466	1301.76 (£65,087,945)
Step discounting (3.5% up to year 30, 3.0% thereafter (both costs and health effects)	£158,853,044	3202.28	£49,606	25.22 (£1,260,898)
Standard of care costs based on FLAG-IDA	£171,269,663	3177.09	£53,908	-248.3 (-£12,415,022)

Again, the results of the one-way sensitivity analyses indicate that the results of the evaluation in the Curative intent TPP are sensitive to assumptions on the potential for re-treatment of CAR T-cell, and the assumed discounting rate for health effects in the model. The results of the evaluation are relatively insensitive to assumptions on the discounting rate for costs, the use of stepped discounting rate (versus constant discounting rates), and to reducing the cost of standard of care treatment to values consistent with treatment using FLAG-IDA (keeping the same efficacy).

If the committee were to consider the criteria met for applying the non-reference-case discount rate of 1.5% for both costs and health effects, then the ICER would reduce to £35,466 per QALY and CAR T-cell therapy would be associated with additional population NHE equivalent to 1302 QALYs (£65.1 million) in comparison to health foregone elsewhere.

Employing the stepwise discounting recommended by the UK treasury, again makes only a small difference in the ICER results with an ICER of £49,606. The incremental population NHE is 25 QALYs (£1.26 million) in comparison to health foregone elsewhere.

The sensitivity of the results to assumptions on the potential for re-treatment of CAR T-cell therapy was considered to represent a more important issue within this TPP. That is, the longer term survival benefits are directly linked to the curative potential of the CAR T-cells themselves rather than to an intermediate treatment such as HSCT. Consequently, the potential need to re-administer CAR T-cell therapy over a longer period represents an important additional source of uncertainty within this TPP, particularly for the minimum data set with relatively short follow-up.

Probabilistic analysis

The results of the probabilistic analysis therapy are shown in Table 38. The probabilistic ICER increased to £50,906 due to the model non-linearities. Consequently, population NHE were now negative with an overall loss to the health system of 56 QALYs (£2.8 million). At a £50,000 cost-effectiveness threshold, the probability that CAR T-cell therapy was the most cost-effective option was 50.7%.

Table 38: Base case probabilistic analysis results: Curative intent TPP (minimum evidence set)

	Population-level			Cost-effectiveness threshold of £50,000 per QALY gained			
Treatment	E[Costs]	E[QALYs]	ICER	E[NHE], QALY (£)	Incremental NHE, QALY (£)	Probability cost-effective	Consequences of decision uncertainty, QALY (£)
CAR T-cell therapy	£183,931,590	3501.50	£50,906	-177.13 (-£8,856,695)	-56.4 (-£2,823,943)	50.7%	304.6 (£15,229,876)
Standard of care	£25,270,727	384.76		-120.66 (-£6,032,752)			

The cost-effectiveness acceptability planes and curves are presented in Figure 23 and Figure 24.

Figure 23: Incremental cost-effectiveness plane: Curative intent TPP (minimum evidence set)

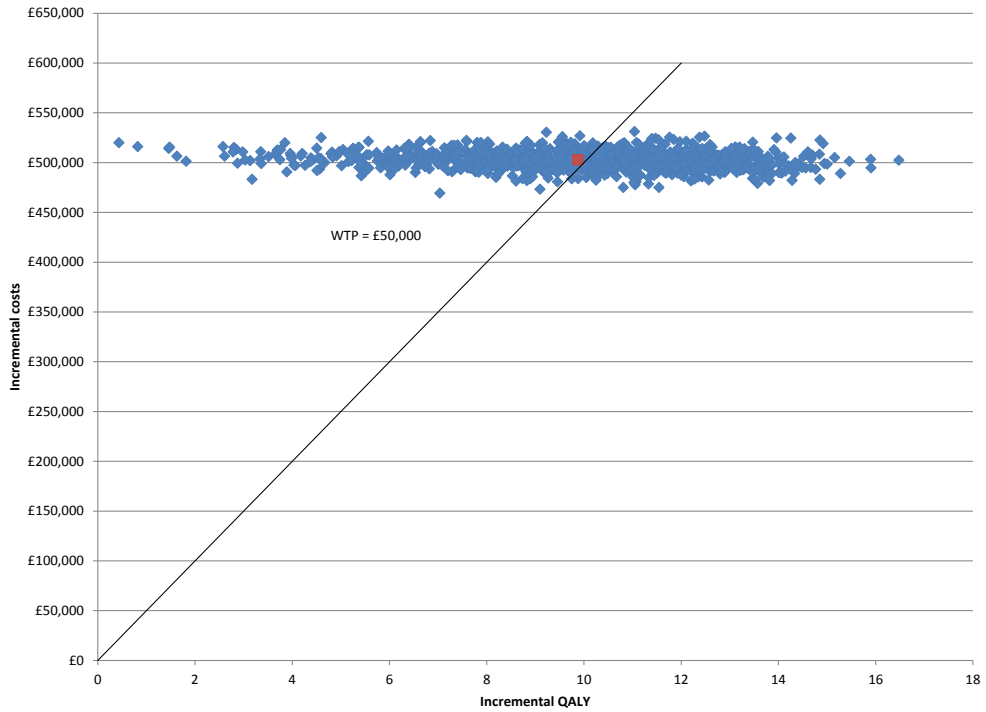
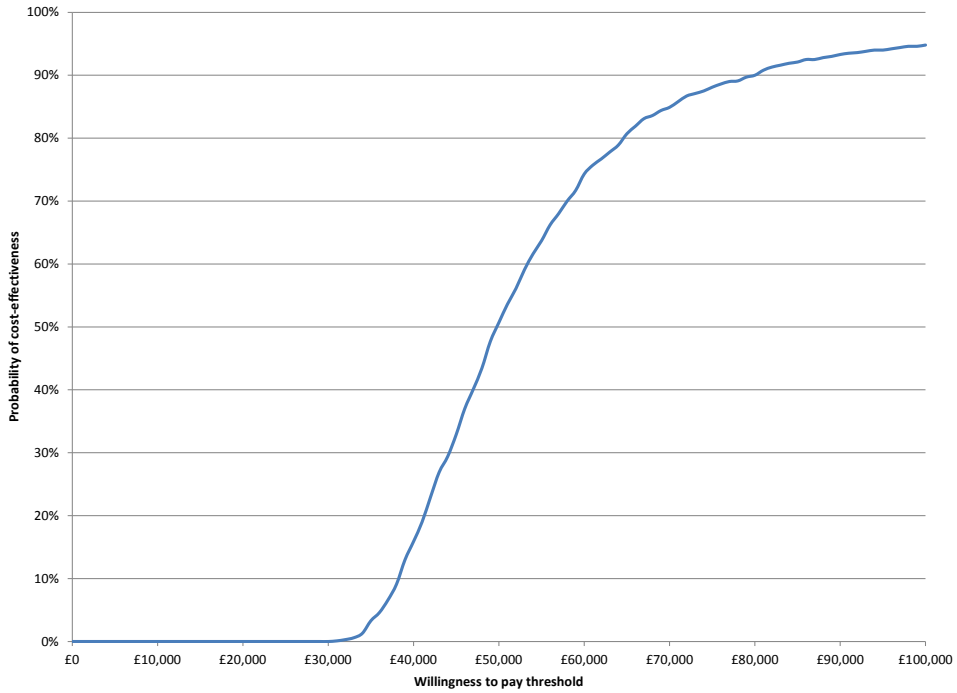
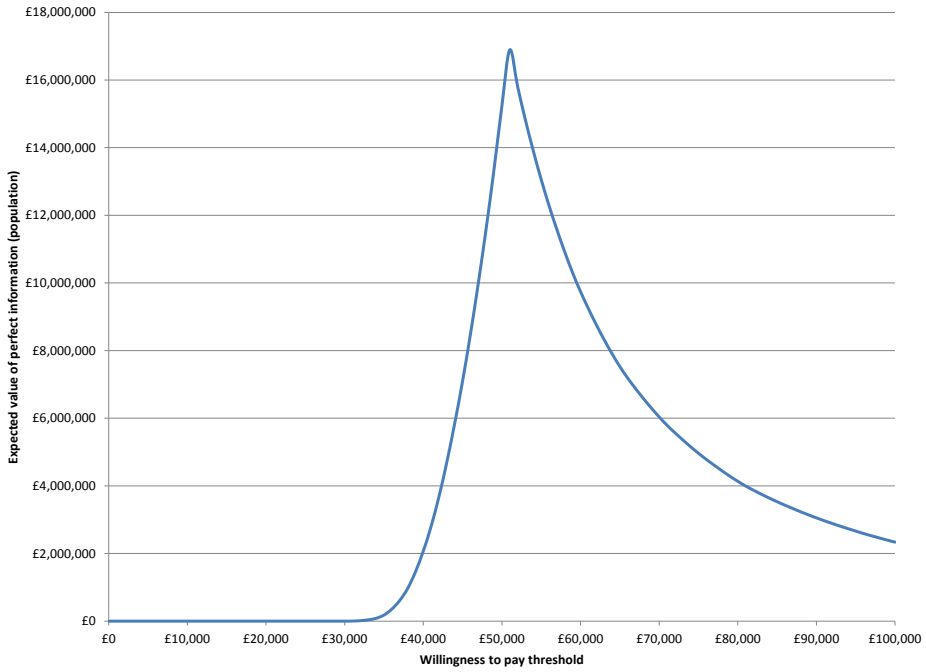


Figure 24: Cost-effectiveness acceptability curve: Curative intent TPP (minimum evidence set)



The consequences of decision uncertainty in the minimum evidence set are estimated to be 304.6 QALYs (£15.23 million). Figure 25 shows how the scale of the consequences of decision uncertainty varies across different cost-effectiveness threshold, reaching a peak at the £50,000 threshold.

Figure 25: Consequences of decision uncertainty: Curative intent TPP (minimum evidence set)



A summary of the population level incremental net health effects, net monetary benefits, probability of cost-effectiveness and consequences of decision uncertainty across a range of cost-effectiveness thresholds is presented in Table 39.

Table 39: Consequences of decision uncertainty across different thresholds - Curative TPP (minimum evidence set)

Cost-effectiveness threshold	Incremental NHE, QALY (£)	Incremental NMB, £	Probability cost-effective	Consequences of decision uncertainty, QALY (£)
£20,000	-4816.30	-£96,326,095	0.0%	0 (£0)
£30,000	-2171.96	-£65,158,711	0.0%	0 (£0)
£50,000	-56.48	-£2,823,943	50.7%	304.6 (£15,229,876)
£75,000	1001.26	£75,094,517	88.1%	66.2 (£4,963,418)
£100,000	1530.13	£153,012,977	94.8%	23.4 (£2,336,731)

At conventional thresholds of between £20,000 and £30,000 per QALY gained, the probability that CAR T-cell therapy is cost-effective versus standard of care is 0%. Consequently, there are no consequences of decision uncertainty at these threshold values. At a willingness to pay threshold of £50,000 per QALY gained, the probability that CAR T-cell therapy is cost-effective versus standard of care is 50.7%. In this case, the expected population health consequence of decision uncertainty is 305 QALYs (10-years). The corresponding expected monetary cost of decision uncertainty is approximately £15.2 million. At thresholds of £75,000 and £100,000 per QALY gained the probability that CAR T-cell therapy is cost-effective increases to over 88%. Despite there being high certainty that CAR T-cell therapy is cost-effective at these thresholds, the corresponding consequences of decision uncertainty remain relatively high at 66 (£4.9 million in monetary terms) QALYs, and 23 QALYs (£2.3 million) respectively.

Alternative pricing scenarios

Alternative pricing scenarios probabilistic analysis

The probabilistic results based on alternative hypothetical pricing scenarios are shown in Table 40. A comparison plot of the consequences of decision uncertainty across the alternative pricing scenarios is shown in Figure 27.

Table 40: Impact of different pricing schemes on the cost-effectiveness of CAR T therapy, and the associated consequences of decision uncertainty: Curative intent TPP (minimum evidence set)

Pricing scenario	Population-level				Cost-effectiveness threshold of £50,000 per QALY gained			
	Treatment	E[Costs]	E[QALYs]	ICER	E[NHE], QALY (£)	Incremental NHE, QALY (£)	Probability cost-effective	Consequences of decision uncertainty, QALY (£)
Base case	CAR T-cell therapy	£183,931,590	3501.50	£50,906	-177.13 (-£8,935,381)	-56.48 (-£2,902,629)	50.7%	304.6 (£15,229,876)
	Standard of care	£25,270,727	384.76		-120.66 (-£6,032,752)			
Leasing method	CAR T-cell therapy	£181,832,300	3488.85	£50,618	-147.79 (-£7,389,708)	-38.21 (-£1,910,653)	49.2%	65.6 (£3,277,969)
	Standard of care	£25,317,596	396.77		-109.58 (-£5,479,055)			
Payment for remission patients only (90% on average)	CAR T-cell therapy	£167,127,512	3510.80	£45,708	168.25 (£8,412,636)	266.50 (£13,325,042)	63.9%	236.1 (£11,803,131)
	Standard of care	£25,219,827	406.15		-98.25 (-£4,912,407)			
Fixed pricing discount (10%)	CAR T-cell therapy	£167,054,363	3535.49	£45,131	194.40 (£9,719,974)	305.88 (£15,293,860)	64.2%	209.1 (£10,456,541)
	Standard of care	£25,301,914	394.56		-111.48 (-£5,573,886)			

Figure 26: Scatter plots of incremental costs and incremental effectiveness across the four pricing scenarios (upper left – base case, upper right – Leasing method, lower left – payment for remission only, lower right – fixed discounting)

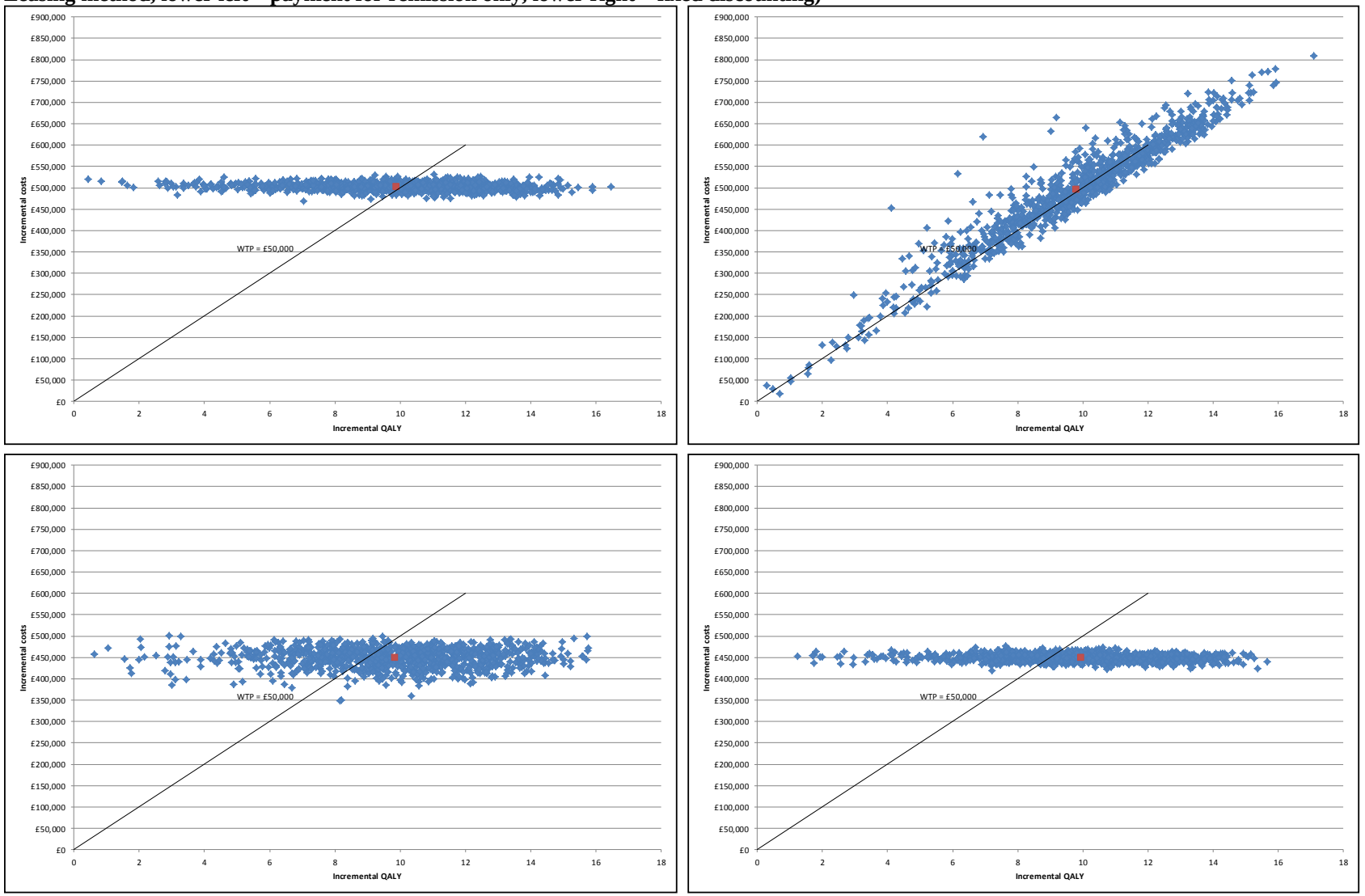
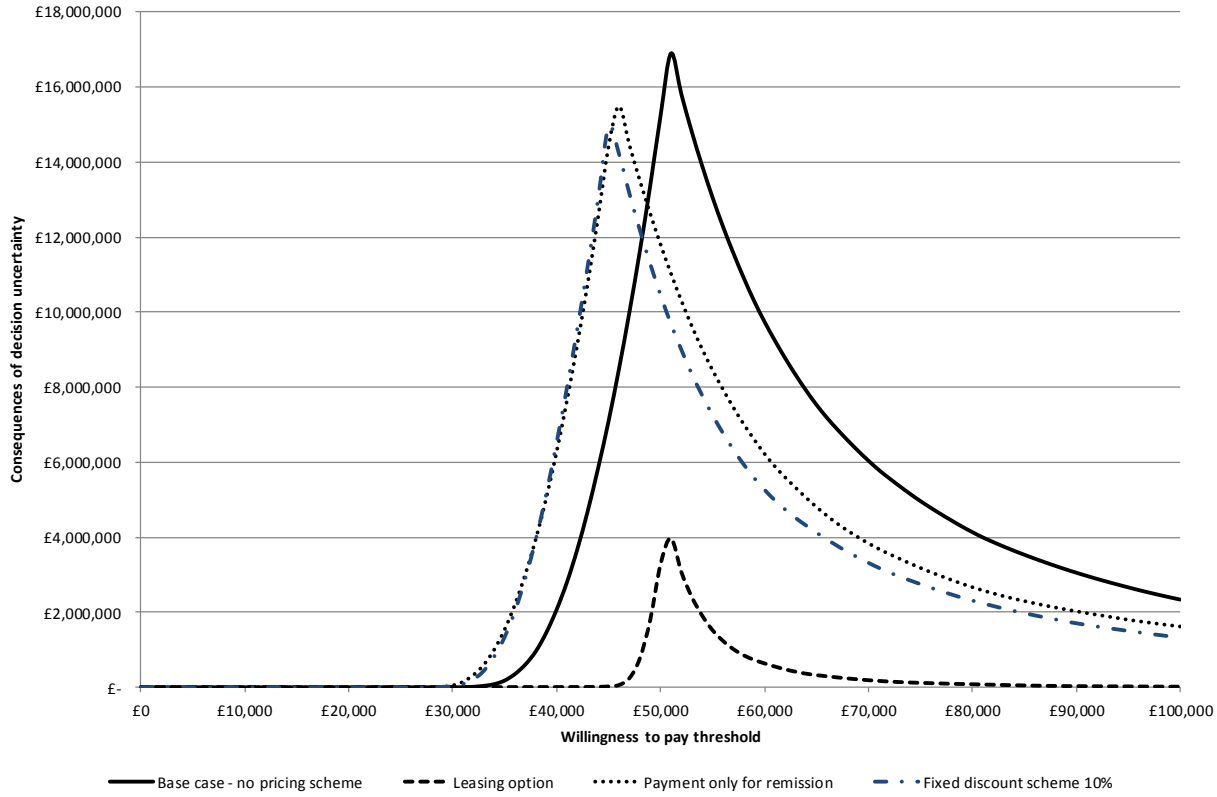


Figure 27: Comparison plot of consequences of decision uncertainty across alternative pricing schemes: Curative TPP (minimum evidence set)



As observed in the previous analysis, the fixed one-off acquisition cost approach is associated with the highest potential consequences due to decision uncertainty (304.6 QALYs; £15.2 million). As before, the alternative schemes result in reductions in decision uncertainty and associated consequences to the NHS. However, the impact and mechanism in which this is achieved differs across the separate approaches.

The leasing approach results in only a minor difference in the ICER. Similar levels of decision uncertainty also remain (i.e. the probability that the intervention is cost-effective is similar to that under a fixed one-off acquisition cost approach). However, the scale of the consequences of the uncertainty to the NHS is significantly reduced via this scheme. This scheme limits the risk to the NHS for over-paying for a technology which doesn't achieve the expected outcomes, significantly lowering the consequences of decision uncertainty from over 300 QALYs in the base case to 65.6 QALYs (£3.2 million) with the leasing approach.

Alternative evidence sets probabilistic analysis

The results of the probabilistic analysis therapy are shown in Table 41. A comparison plot of the consequences of decision uncertainty across the evidence sets is shown in Figure 28.

As expected, the health consequence of decision uncertainty in the mature evidence set (14.1 QALYs; £707,000) is lower than that reported in the minimum set (304.6 QALYs; £15.2 million), at a threshold of £50,000 per QALY gained. These consequences are reduced by the increased certainty surrounding the trajectory of the parametric survival curves, and the effect of increased maturity on improving the cost-effectiveness of CAR T therapy. As is evident from Figure 28, the increased certainty over the longer-term survival benefits of treatment (and represented by the longer-follow up assumed in the intermediate and mature evidence sets), has a proportionately greater effect in reducing decision uncertainties within the minimum dataset than the increased precision of greater patient numbers (i.e. only reflected in the mature evidence set).

Figure 28: Comparison plot of consequences of decision uncertainty across alternative evidence sets: Curative TPP (minimum evidence set)

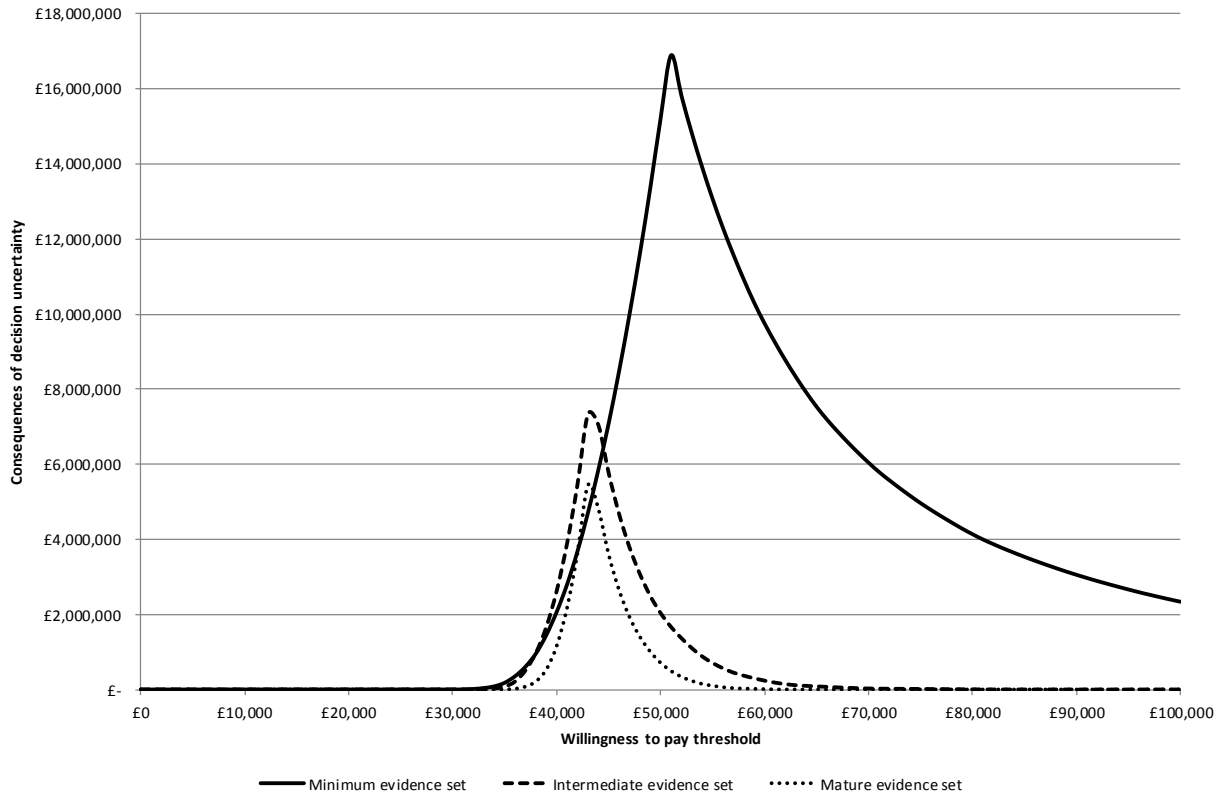
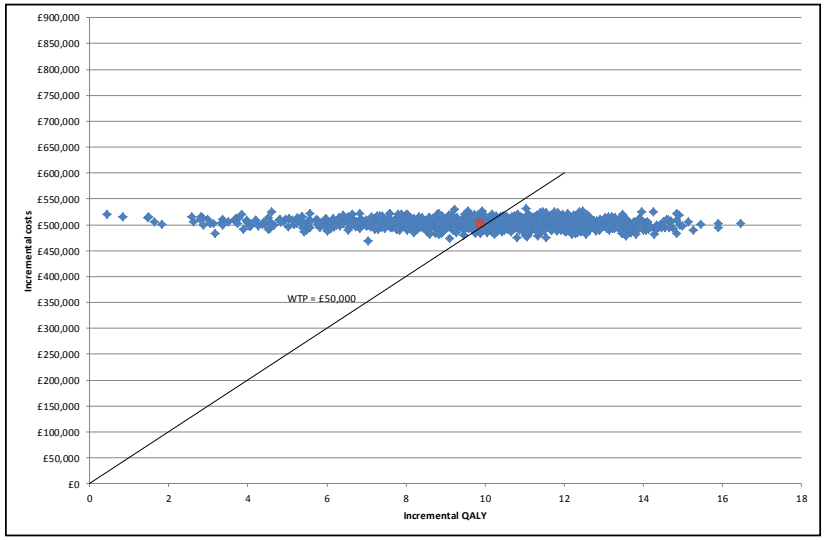
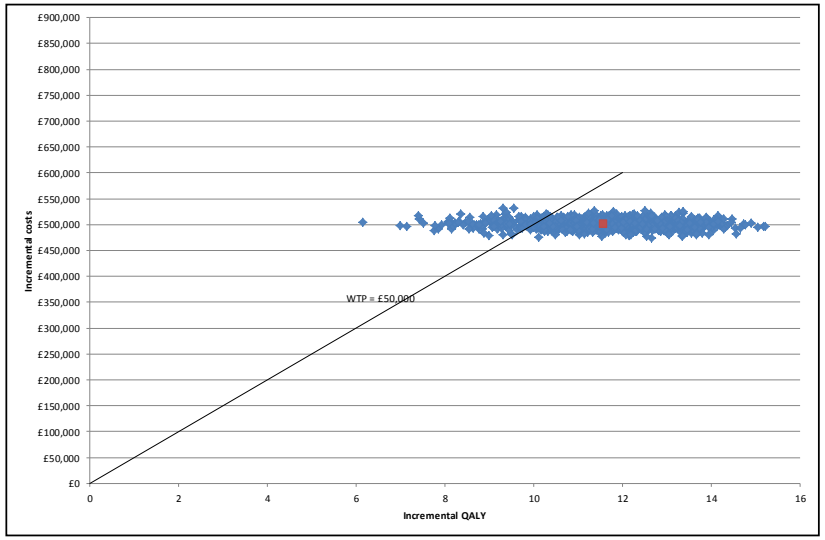


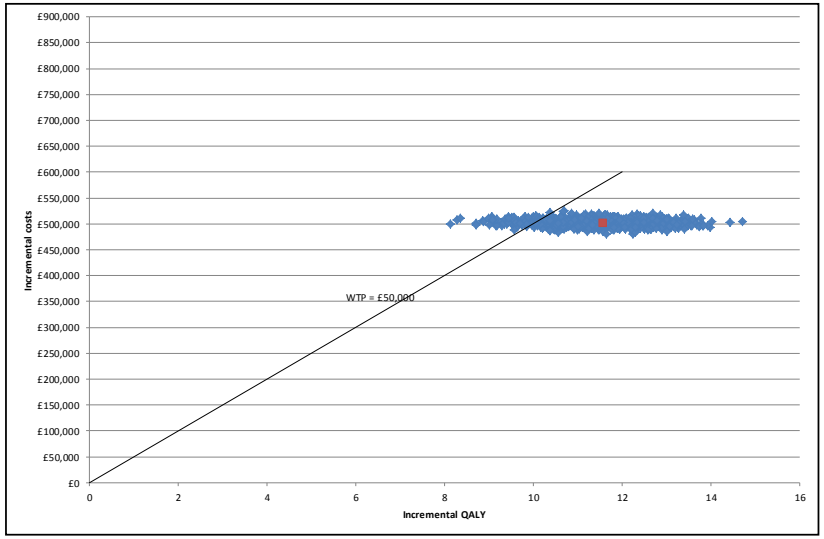
Table 41: Probabilistic results comparing across evidence sets: Curative intent TPP

Evidence set	Population-level				Cost-effectiveness threshold of £50,000 per QALY gained			
	Treatment	E[Costs]	E[QALYs]	ICER	E[NHE], QALY (£)	Incremental NHE, QALY (£)	Probability cost-effective	Consequences of decision uncertainty, QALY (£)
Minimum	CAR T-cell therapy	£183,931,590	3,501.50	£50,906	-177.13 (-£8,935,381)	-56.48 (-£2,902,629)	50.7%	304.6 (£15,229,876)
	Standard of care (clofarabine treatment)	£25,270,727	384.76		-120.66 (-£6,032,752)			
Intermediate	CAR T-cell therapy	£183,586,917	4,296.77	£43,344	625.03 (£31,251,488)	486.22 (£24,311,227)	85.9%	40.6 (£2,031,623)
	Standard of care (clofarabine treatment)	£25,264,818	644.10		138.81 (£6,940,262)			
Mature	CAR T-cell therapy	£183,560,268	4,307.12	£43,252	635.91 (£31,795,547)	494.47 (£24,723,328)	91.5%	14.1 (£707,136)
	Standard of care (clofarabine treatment)	£25,103,273	643.51		141.44 (£7,072,220)			

Figure 29: Scatter plots of incremental costs and incremental effectiveness across evidence sets (upper left – base case, upper right – intermediate evidence set, lower – Mature set)







9.5 Conclusions

The primary purpose of this section was to report the potential cost-effectiveness of CAR T-cell therapy within the separate scenarios considered and to highlight key uncertainties surrounding these results. An important aspect of this work was also to consider how these estimates could be presented and communicated to the Committee to inform their deliberations. In doing this we presented analyses based on approaches routinely requested within NICE's existing methods guide. We also undertook additional analyses may provide useful additional insights to help inform subsequent committee deliberations and the potential nature of such analyses.

The sequence of assessments presented started with a conventional assessment of cost-effectiveness at the patient level based on the minimum evidence set. Disaggregated estimates of the costs and outcomes were estimated, together with resulting cost-effectiveness estimates based on the ICER. These results were also expressed using NHEs, representing the difference between any health gained with the intervention and health foregone elsewhere in the health care system, expressed either in monetary and QALY terms. The impact of uncertainties was explored using conventional one-way sensitivity analyses (i.e. varying individual parameters or specific assumptions) and probabilistic approaches (i.e. exploring the impact of joint uncertainty across all parameters). Conventional scatter-plots and acceptability curves were utilised to graphically show the impact of parameter uncertainties and other more methodological uncertainties (e.g. the appropriate discount rate). The analyses also explored the potential impact if the committee were to consider the criteria met for applying the non-reference-case discount rate of 1.5% for costs and health effects.

In addition to the analyses undertaken using the conventional reference case approaches, a series of more exploratory analyses were also undertaken. In particular, the per-patient assessments were subsequently scaled up to population assessments, requiring an estimate of the number of potentially eligible patients (assumed to be approximately 38 patients per annum) and an assessment of the 'technology time horizon' i.e. the period over which the therapy might be utilised within clinical practice (assumed to be 10 years in the exemplar). Although the presentation of population-level analyses are not formally requested within the existing NICE methods guide for reporting cost-effectiveness results, an assessment of population impact is required within Section 5.12 (Impact on the NHS). Hence, these exploratory analyses were considered to be consistent with the requirement to consider population impact and the specific requests within Section 6.4.1 (Research recommendations) for the committee to balance the potential NHEs of current and future NHS patients when considering making research recommendations.

The results of the population based analyses were summarised in terms of incremental NHE (both in terms of QALYs and equivalent monetary value) together with an assessment of the probability that CAR T-cells were cost-effective. Alongside these more conventional assessments, an assessment of the scale of the likely consequences was considered to be potentially informative to the committee, particularly in deliberations related to possible research recommendations. An estimate of the consequences of existing decision uncertainty was subsequently derived reflecting the possible scale of NHEs that could be gained if uncertainty surrounding this decision could be resolved

Using the different analyses, the impact of alternative pricing scenarios were explored, including conventional PAS type schemes (i.e. equivalent to a fixed price reduction) as well as more sophisticated schemes based on pay for performance and leasing approaches. Similarly, the impact of the alternative evidence sets was explored to establish the implications of increased precision and maturity assumed in the intermediate and mature evidence sets.

An important consideration within this work is the extent to which current NICE methods and processes are likely to appropriately quantify the potential uncertainties surrounding regenerative medicines and cell-based therapies to ensure that appropriate policy decision are made regarding adoption and spread of potentially promising technologies. Our findings show that the conventional assessments requested within the current TA process may not be sufficient. Estimates of the ICER and associated uncertainty (e.g. probability a technology is cost-effective) were shown to be similar in one of the TPPs despite being based on 3 different evidence sets with varying levels of precision and maturity. Consequently, it is unclear how these differences would be reflected within the current deliberative process. Whilst it is acknowledged that different conclusions might be reached based on informal judgements, the importance of ensuring transparency in subsequent decisions remains a key principle of the Institute and appears critical for manufacturers in developing appropriate R&D and pricing strategies.

Presentation of the scale of consequences using population NHE provided a clearer distinction between the different evidence sets and an assessment of the impact of alternative pricing schemes. Consequently, their more routine application within the TA process for regenerative and cell-based therapies may be an important consideration for the Institute. Furthermore, such comparisons could also provide a more transparent and explicit basis for considering the value of direct price reductions that might be realised via a conventional PAS (or less conventional schemes which work by indirectly lower the effective price) compared to the provision of additional evidence (both precision and maturity), in terms of reducing decision uncertainty and its consequences. Such information might provide an important basis for discussions between manufacturers, NICE and other relevant parties in

terms of how the existing uncertainties that exist might be appropriately managed ensuring risks and benefits are more appropriately shared.

10 Issues arising from the NICE panel meeting

A separate panel and meeting were convened by NICE to discuss the findings from Sections 1 to 9 of the report. The panel included clinical experts and current and past NICE committee members and was chaired by Professor Andrew Stevens (current chair NICE TA committee). A full list of panel members is provided in Appendix 9. The objective of the panel meeting was to assess the clinical and cost-effectiveness evidence informing the separate TPPs and to identify potential issues and challenges for the NICE TA appraisal process and methods.

A summary of the clinical and cost-effectiveness evidence was presented to the panel, together with an overview of key technical and process issues for consideration. The panel was then presented with a series of separate decision scenarios reflecting: the 2 separate TPPs ('Bridge to HSCT' and 'Curative Intent'); the 3 evidence sets (minimum, intermediate and mature); and the impact of different pricing approaches based on the minimum evidence set. The panel was requested to deliberate on the scenarios and to provide 'hypothetical' decisions and to outline the main considerations for these. The panel were requested to focus particularly on the role of uncertainty (clinical and cost-effectiveness) in order to: (i) identify key areas of uncertainty; (ii) understand the nature of assessments/analyses that could help inform deliberations and (iii) explore the impact of different pricing approaches and different evidence sets.

The main clinical and cost-effectiveness issues discussed by the panel are summarised below. This is followed by a summary of the panel discussions related to the separate scenarios.

Clinical issues

When asked for their thoughts, following a presentation of the clinical effectiveness and safety issues, the panel clinical experts commented that although the data for CAR T-cell therapies are limited, the results nevertheless appeared to be *very* encouraging when compared with the best available alternative (clofarabine). They added that manufacturers nevertheless need guidance on how to account for the uncertainty of trial results, given the availability of only short-term data and potential long-term effects. It is likely that future cell therapies will be aimed at larger populations (which ties in with the EMA's adaptive pathways approach – see Section 3). They also highlighted that clarity was needed around how data requirements might change according to the size of population.

The clinical experts stated that knowledge is improving about which patients will have side effects from CAR T-cell therapies. Knowledge on predictors of response (effect modifiers) was less developed, although the panel thought that the possibility of evidence review groups having access to individual patient data (IPD) during any assessment could be an important step to help identify possible effect modifiers and to assess the reliability of submitted evidence. For this assessment it was

suggested that there might be interest in whether relapsed patients responded better than refractory patients.

The clinical experts were also asked about the potential variability in efficacy and safety profiles of these types of intervention due to manufacturing variability and heterogeneity in patient response. It was considered that any differences in efficacy and safety due to variability in the manufacturing process are likely to be largest early on, but will be optimised with time. Variability of efficacy and safety due to individual patient heterogeneity is however likely to remain.

In response to a question from the panel about the success rates when manufacturing individual treatments, the clinicians said that although the success rates for 'expanding' CAR T-cells is high for B-ALL patients, it is difficult to tell which patients' (cells) can be successfully expanded (i.e. successful manufacture of the bespoke treatment). They stated that patients may die before the cell therapies can be produced and administered. It was noted that it will therefore be **very** important that trials report data relating to the full 'intention-to-treat' (ITT) population, including those patients for whom CAR T-cell expansion was not successful. If any patients required re-treatment this should also be clearly reported.

There were serious concerns from the panel regarding the level of uncertainty in the evidence base, in particular that it was based on single-arm trials with possibly large unknown bias. There were concerns from the panel that certain efficacy estimates, particularly for the minimum data set, might be too optimistic and questions were asked whether any such biases could be quantified and adjusted for; it would be useful to see the impact of more pessimistic efficacy estimates on cost-effectiveness results. There were concerns around the long-term benefit of the therapy and whether the estimate of overall survival in the minimum data set really could be carried into the mature data set.

Another issue which was raised was the panel being provided with knowledge on what further research had been mandated by the EMA (e.g. for conditional approvals). Understanding this may be key to knowing how much present uncertainty, and at what cost, can be accepted. The difficulties of decommissioning services once treatments are approved were also raised as potential problems.

Cost-effectiveness issues

A key consideration regarding the cost-effectiveness results and implications for the 'hypothetical' decisions was whether the panel considered that existing criteria considered within the TA process in relation to End of Life (EoL) and 1.5% discounting (applied to costs and health outcomes) could be applied. The panel accepted that, based on the patient numbers, current prognosis and the likely treatment benefit, CAR T-cell therapy for relapsed/refractory ALL would be likely to meet existing

criteria for EoL. However, the panel also noted that the existing criteria might need to be reconsidered more generally for therapies with curative potential. It was argued by one panel member that the EoL criteria were developed to cover scenarios where people with conditions such as cancer with a short life expectancy, were given some extension, but whose life expectancy was still short. It was suggested that different QALY weights might need to be considered over a longer period of projected survival benefits for therapies which have curative potential.

The use of the 1.5% discounting was also discussed by the panel. While it was noted that the existing criteria had been developed in response to a similar decision context, the panel were also aware that the criteria had only been applied in 1 previous appraisal (the TA for which it was developed). The lack of precedent was noted and the panel concluded that its application could generate significant debate in future appraisals. Hence, no conclusion was reached during the panel meeting about its application to CAR T-cell therapy. The use of stepped-discounting recommended by the Treasury discount was discussed by the panel but was considered to be more relevant for interventions which might have important intra-generational impacts (e.g. immunisation) as opposed to longer-term inter-generational effects.

In addition to the concerns noted by the panel in relation to the possible bias and additional uncertainty arising from comparisons based on single-arm trials, the panel also raised questions regarding whether there were wider structural uncertainties relevant to regenerative medicines and cell-based therapies that were not fully captured within the analyses presented. The panel concluded that identifiable sources of potential bias and appropriately reflecting structural sources of uncertainty would be an important consideration in future appraisals and manufacturers would need to clearly report how these had been addressed within their submission.

The panel discussed the sequence of assessments presented in the cost-effectiveness section and the exploratory approaches to quantifying decision uncertainty based on an assessment of the scale of the consequences associated with each decision, using population NHE. The panel agreed that these exploratory approaches provided a clearer and potentially important distinction between the different evidence sets and the impact of alternative pricing schemes. The panel also acknowledged that such assessments provided important information which could help inform their deliberations. However, the panel further noted that while such assessments were helpful and represented a useful starting point for deliberations, they were not necessarily sufficient for informing their final decisions. In particular, the panel expressed difficulty in determining how to interpret the numbers presented without a formal reference point to establish whether the consequences were sufficiently high to impact on their decisions and/or potential research recommendations.

The panel acknowledged that the estimates of the consequences represented a theoretical upper bound to the value of further research. However, the panel concluded that it would be important to further explore these consequences, both in terms of the underlying distribution (as opposed to the expected mean value of the consequences) as well as needing to decompose the overall estimate in relation to specific sources of uncertainty. This latter aspect was considered particularly important in determining the extent to which particular sources of uncertainty could be resolved by additional research, the type of research which might be most appropriate and finally whether this research would be feasible following a positive approval. The panel were also aware of the relevance of existing published work[88] and ongoing work by the NICE DSU that would be important to consider in any review of potential process or methods.

Prior to a more detailed discussion of the specific decision scenarios, the panel outlined a number of more general considerations related to the cost-effectiveness evidence and results:

- In discussing the appropriate cost-effectiveness threshold for the purpose of NICE decision making, the panel was clear that £50,000 per QALY (assuming the EoL criteria applied) represented an absolute upper bound to the range that NICE would consider acceptable. The panel concluded that other considerations (e.g. innovation) would not be applied in conjunction with the higher threshold considered in an EoL appraisal. Furthermore, the panel also considered that the upper end of the range was unlikely to be considered appropriate in the presence of significant evidential uncertainties.
- The panel concluded that if the hypothetical price of CAR T-cell therapy had been set using the conventional cost-effectiveness threshold range (£20,000 - £30,000 per QALY) that this could have mitigated some of these uncertainties, increasing the likelihood of a positive recommendation.
- The panel appreciated that there was a difference between the deterministic and probabilistic estimates of the ICER due to the non-linearity between the parameter inputs and the model outputs (i.e. mean costs, QALYs and ICER). The panel also noted that for some analyses, these differences resulted in ICER estimates which could have a material impact on their decisions (i.e. situations in which the deterministic and probabilistic estimates lay either side of the cost-effectiveness threshold). The panel concluded that the probabilistic estimates were the more appropriate basis for informing their decisions.
- The panel raised issues regarding the possible nature and magnitude of any irrecoverable costs that might be incurred by the NHS and the implications for their decisions. The panel concluded that an ‘exit strategy’ for the NHS would be a key consideration for interventions

which appear highly promising but where significant uncertainties and irrecoverable costs may exist.

- The panel acknowledged that the different pricing schemes had important impacts both in terms of the ICER but also in terms of the allocation of any risk between the NHS and manufacturers. The concept of the 'leasing approach' was identified as a potentially important option and there was consensus amongst the panel that this warranted further exploration by NICE and manufacturers (e.g. logistics, costs and overall feasibility).
- The panel recognised the various issues and challenges likely to be faced by the manufacturers of regenerative medicines and cell-based therapies. The panel also noted that many of the issues identified did not appear specific to these types of therapies and that many of the issues and implications identified were also apparent in appraisals of more conventional products. However, the panel acknowledged that that the challenges may be faced more routinely for regenerative medicines and cell-therapy manufacturers and that the resulting levels of uncertainty (and the potential scale of the consequences) may exceed that for which existing committees might conclude could be appropriately dealt with by existing processes and the current methods guide.

Panel discussion of scenarios

Following a general discussion of clinical and cost-effectiveness issues, the panel were presented with a series of 'decision scenarios' based on the results reported in Section 9. For each TPP, the scenarios started with the minimum evidence set and a fixed acquisition cost for CAR T-cell therapy (Scenario 1). Scenario 2 explored the impact of alternative pricing approaches based on the same minimum evidence set. Scenarios 3 and 4 were based on the results from the intermediate and mature evidence sets, assuming a fixed acquisition cost.

For each scenario the panel were presented with a summary of the deterministic ICER and the probabilistic, population level results including an estimate of the ICER, incremental NHE (expressed in monetary and QALY terms), the probability cost-effective and an assessment of the scale of the consequences of decision uncertainty (again expressed in monetary and QALY terms).

A summary of the panel considerations is provided below.

- For Scenario 1, the panel understood that the deterministic estimate of the ICER for CAR T-cell therapy was close to the £50,000 upper bound of the ICER range considered acceptable currently when the EoL criteria is met. However, the panel concluded that the probabilistic estimates of the ICER were more appropriate given the model non-linearity. Since the probabilistic ICER in the base-case of both TPPs exceeded the upper bound of the ICER range, the panel concluded that

CAR T-cell therapy would be unlikely to represent an efficient use of NHS resources in Scenario 1. Although other aspects of innovation were discussed, the panel concluded that while these were important considerations for CAR T-cell therapy, additional weight should not be incorporated over and above that which had already been permitted when applying the EoL criteria.

- For Scenario 2, the panel acknowledged the different impacts of the alternative pricing schemes on the ICER, the scale of the consequences of decision uncertainty as well as the apportionment of any risk between the NHS and a manufacturer. The panel noted that the lifetime leasing scheme resulted in a significant reduction in the scale of decision uncertainty compared to a fixed acquisition cost. The panel also acknowledged that the leasing scheme could also provide an important exit strategy for the NHS given the high uncertainties which were evident. There was consensus amongst panel members that innovative financing schemes could be an important consideration in future appraisals.
- The panel also noted that there were important differences in the scale of the consequences of decision uncertainty across the separate TPPs, with significantly higher consequences reported in the 'Curative Intent' TPP. The panel understood that the use of an external surrogate relationship between MRD, HSCT and remission status in the 'Bridge to HSCT' TPP had an important impact on reducing the scale of the decision consequences over the modelled time horizon.
- The panel found it difficult to determine the policy significance of the estimates reported for decision uncertainty without further analyses and an appropriate reference point. However, the panel also acknowledged in the 'Bridge to HSCT' TPP that the magnitude of the incremental NHE (i.e. the NHE that might be gained from immediate approval) significantly exceeded the scale of the consequences of decision uncertainty for the pricing schemes based on a pay for performance approach based on achieving remission. The panel understood that the higher incremental NHE reported in these scenarios (and reduction in the consequences of decision uncertainty) was driven by the lower ICER due to the direct or indirect impact on the acquisition cost of CAR T-cell therapy and that this had an important impact on the scale of the consequences of decision uncertainty.
- Faced with the high levels of uncertainty, the panel concluded that schemes which brought the ICER significantly lower than the upper bound of the EoL range and closer to the more conventional ICER range (£20,000-£30,000 per QALY) would increase the likelihood of approval.
- The panel also acknowledged that the reduction in the consequences of decision uncertainty in the leasing and the payment for remission schemes arose due to the risks being shared between the NHS and manufacturers. Although the ICER of the lifetime leasing method exceeded the upper bound of the EoL range, the panel concluded that they may have looked more favourably on a

combined scheme involving a fixed price discount and a leasing element in the 'Bridge to HSCT' TPP. However, in the absence of being provided with a formal assessment of this scheme, the panel felt it was not possible to make a clear recommendation.

- The panel were less clear on potential recommendations across the different pricing schemes for the 'Curative Intent' TPP. Although the panel acknowledged that the consequences of decision uncertainty were reduced by each of the alternative pricing approaches, the panel remained concerned at the scale of the consequences which remained. Again, the panel concluded that they may have looked more favourably on a combined scheme involving a fixed price discount and a leasing element but noted that they had not been presented with results from such a scenario.
- The panel were also aware that different prices were assumed across the separate TPPs reflecting the different effectiveness estimates reported in the different studies used in each. The panel indicated that if the same price that was used in the 'Bridge to HSCT' TPP had been applied to the 'Curative Intent' TPP, this would have potentially significantly improved the ICER and lowered the consequences of decision uncertainty.
- Faced with higher consequences in the 'Curative Intent' TPP, the panel concluded that the combination of using the same price in the 'Bridge to HSCT' and a leasing scheme would potentially improve the ICER and lower the consequences of decision uncertainty to a level which could potentially be acceptable. Again, in the absence of being provided with a formal assessment of such a scheme, the panel felt it was not possible to make a clear recommendation.
- The panel discussed the additional evidence sets that had been generated for each TPP (Scenarios 3 and 4). The panel acknowledged that these estimates were generated using a series of assumptions and hence remained subject to various additional uncertainties. However, the panel understood the principles which were being considered and that there were important differences across the evidence sets for the separate TPPs. The panel understood that the difference across the TPPs was primarily due to the use of an external surrogate relationship being used in 'Bridge to HSCT' TPP. The panel acknowledged that greater uncertainty could arise in situations where a robust surrogate relationship had not been demonstrated and that ensuring evidence is sufficiently robust (i.e. in terms of precision and/or maturity) for decision making would be an important consideration. The panel noted that the consequences of decision uncertainty in the intermediate evidence set for the 'Curative Intent' TPPs were significantly reduced compared to the minimum evidence set and were closer to the scale of those reported for the minimum evidence set for the 'Bridge to HSCT' TPP where a surrogate relationship had been assumed.
- The panel understood that the scale of consequences was further reduced in the mature evidence sets due to increased precision (compared to the intermediate data set) and maturity (compared to

the minimum evidence set) and that this was most evident in the 'Curative Intent' TPP because additional surrogate evidence had not been included.

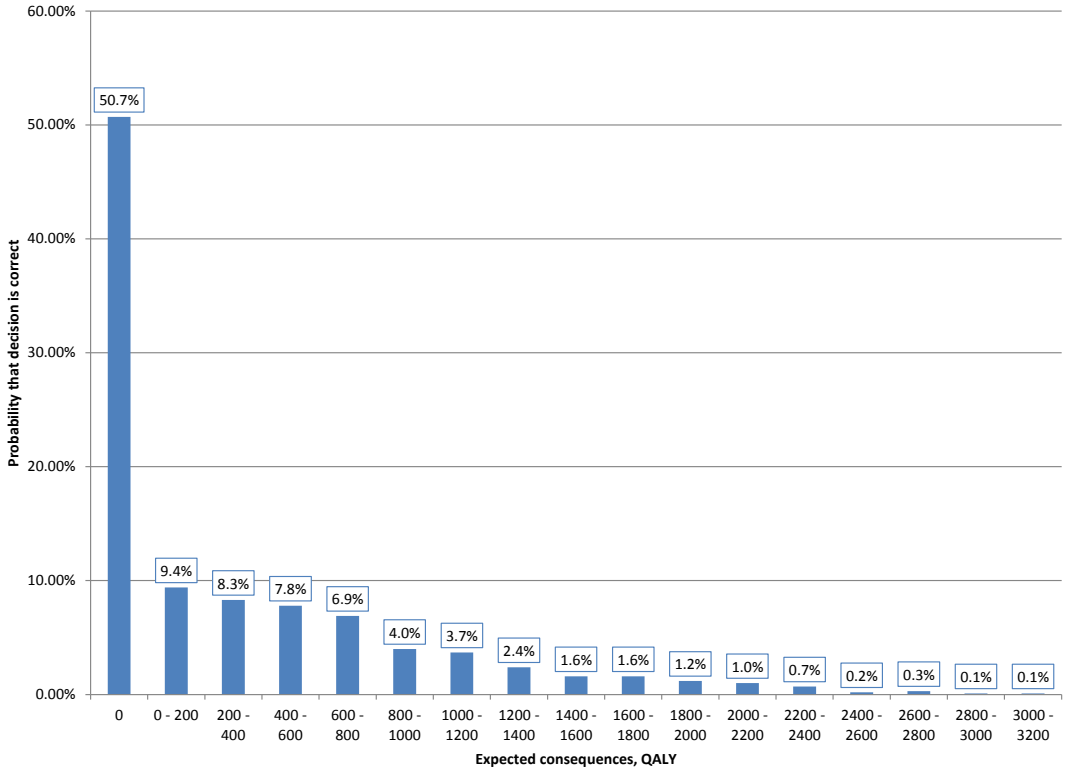
- The panel acknowledged the challenges and difficulties of generating mature evidence at the point a product is launched. However, the panel considered that the principles outlined through the different assessments would be important in informing future deliberations. In particular, the panel noted that a comparison of the magnitude of the incremental NHE and the consequences of decision uncertainty provided an important starting point for deliberations in considering the scale of the NHE that could be achieved by immediate approval and that which might be achieved by further research.
- The panel noted that further assessments could be helpful to further inform: (i) whether a positive approval decision might alter incentives to undertake the type of research necessary to resolve the main sources of uncertainty, and (ii) the full opportunity costs of approval and rejection decisions. The panel concluded that further information concerning the distribution of the consequences and further exploration of the main sources of these consequences would provide important additional information.

Additional exploratory analyses undertaken after Panel discussion

Following the panel meeting, a series of additional exploratory analyses were undertaken to capture some of the specific requests and considerations that were identified during the panel discussions. These analyses are not intended to be comprehensive but rather to reflect on some of the main points raised and to consider any further implications.

Information on the distribution of consequences is shown in Figure 30 based on the minimum evidence set of the Curative Intent TPP. The most common outcome (50.7%) is for CAR T-cell therapy to be cost-effective at a willingness to pay threshold of £50,000. Consequently, there are no negative consequences to the NHS in these instances. However, in 49.3% of iterations (1-probability of CAR T-cell therapy being cost-effective), the decision to recommend CAR T-cell therapy may be incorrect (at a willingness to pay threshold of £50,000). The consequences of a making an incorrect decision is expressed in terms of the NHE's forgone. In this analysis, most of the negative consequences are less than 1,000 QALYs (probability of 36.4%). The probability that the negative consequences exceed 1,000 and 3,000 QALYs is 12.9%, and 0.1%, respectively.

Figure 30: Distribution of the consequences of decision uncertainty



The panel were also interested in exploring the impact of a number of alternative pricing schemes on the cost-effectiveness of CAR T-cell therapy and associated decision uncertainty. These schemes included the application of the 'Bridge to HSCT' fixed acquisition cost to the 'Curative Intent' TPP, as well as considering the impact of a lifetime leasing approach with or without an additional 10% price discount.

Applying both a lifetime leasing method and a 10% discount to the cost of CAR T-cells to the minimum evidence set analysis for the 'Curative Intent' TPP improved the ICER (£45,502 per QALY), resulting in a large decrease in the consequences of decision uncertainty and increase in the probability of cost-effectiveness, as shown in Table 42.

Applying the 'Bridge to HSCT' fixed acquisition cost of CAR T-cell therapy (£356,100) to the minimum evidence set analysis for the 'Curative Intent' TPP significantly improved the cost-effectiveness of curative CAR T-cell therapy, resulting in an ICER of £34,337 per QALY, as shown in Table 42. With improved cost-effectiveness, the expected consequences of decision uncertainty is also improved; decreasing from 304 QALYs (£15m) in the base case to 73.1 QALYs (£3.7m). Applying a lifetime leasing method resulted in further reductions to the consequences of decision uncertainty to 2.3 QALYs (£0.11m). By applying an additional 10% discount alongside the leasing

and 'Bridge to HSCT' acquisition cost, it was possible to eliminate the potential consequences of decision uncertainty (at a willingness to pay threshold of £50,000).

These additional analyses further reinforce the importance of considering the implications both for the ICER as well as the scale of the consequences of decision uncertainty. A key finding from these additional analyses is that the consequences of decision uncertainty related to the minimum evidence set can be significantly lowered by reductions in price or the application of alternative pricing schemes. Indeed the additional exploratory analysis reveal that the scale of the consequences might be reduced to a similar or even lower magnitude than that which could be resolved through the provision of further evidence alone. Furthermore, by reducing the opportunity costs of early approval, increased flexibility in pricing and pricing approaches would allow more patients receive early access to potentially innovative regenerative medicines and cell-based therapies.

Table 42: Additional exploratory analyses: Curative intent TPP (minimum evidence set)

Pricing scenario	Population-level				Cost-effectiveness threshold of £50,000 per QALY gained			
	Treatment	E[Costs]	E[QALYs]	ICER	E[NHE], QALY (£)	Incremental NHE, QALY (£)	Probability cost-effective	Consequences of decision uncertainty, QALY (£)
Base case	CAR T-cell therapy	£183,931,590	3501.50	£50,906	-177.13 (-£8,935,381)	-56.48 (-£2,902,629)	50.7%	304.6 (£15,229,876)
	Standard of care	£25,270,727	384.76		-120.66 (-£6,032,752)			
Lifetime leasing and 10% discount ((£2,955) per month)	CAR T-cell therapy	£164,420,596	3458.93	£45,502	170.52 (£8,525,896)	275.00 (£13,750,033)	87.2%	27.2 (£1,358,584)
	Standard of care	£25,321,756	401.95		-104.48 (-£5,224,137)			
Same pricing as bridging TPP(fixed cost of £356,100)	CAR T-cell therapy	£129,435,001	3446.20	£34,337	857.50 (£42,874,913)	951.11 (£47,555,583)	85.6%	73.1 (£3,655,992)
	Standard of care	£25,178,368	409.95		-93.61 (-£4,680,670)			
Same total cost as bridging TPP with lifetime leasing (£2211.42 per month*)	CAR T-cell therapy	£129,689,785	3532.92	£33,277	939.12 (£46,956,030)	1050.02 (£52,500,851)	99.4%	2.3 (£112,597)
	Standard of care	£25,219,874	393.50		-110.90 (-£5,544,821)			
Same total cost as bridging TPP with lifetime leasing and 10% discount (£1990.28 per month*)	CAR T-cell therapy	£117,750,114	3509.04	£29,713	1154.04 (£57,701,888)	1262.40 (£63,120,093)	100.0%	0 (£0)
	Standard of care	£25,302,238	397.68		-108.36 (-£5,418,205)			

11 Discussion

11.1 Implications for NICE technology appraisal processes

Modifications (which may sometimes be informed by methods research) might be considered to update the methods guidance provided to manufacturers and ERGs in the following areas:

Use of surrogate endpoints

The choice of surrogate endpoints used by manufacturers in their submissions must be researched, explicit and justified. Ideally, a systematic review should be performed to evaluate the strength of the association between the surrogate and the patient-relevant outcome and the evidence on surrogate validation should be presented according to an explicit hierarchy.

Pivotal study design and the use of historical control datasets

For manufacturer submissions, consideration should be given to benefits of having recommendations and/or minimum reporting requirements on the methods used to obtain and analyse single-arm trial data when they are compared with historical control data. Where single-arm trial data form the main basis of an assessment, a clear rationale should be given for the type of comparisons made (implicit or explicit) and for the choice of the historical control data which were selected. For example the gold standard for historical data might be matched data obtained from a patient database (rather than relying on published studies, which might not fit the trial population being studied well enough). Manufacturers should also consider the evidence on the number of study sites when designing trials (multi-centre trials are likely to produce more reliable and generalizable results than single-centre trials).

ERG's might benefit from using checklists to help when appraising how historical control data were identified and analysed by manufacturers.

Efficacy estimates

Submission of IPD might be beneficial for ERGs, especially where datasets are small. Use of multivariate meta-analysis can lead to reduced uncertainty around the effectiveness parameter. By allowing all the relevant data to be incorporated in estimating clinical effectiveness outcomes - including data from surrogate outcomes - multivariate meta-analysis can improve the estimation of health utilities through mapping methods.

Manufacturers should report the data for the full trial population. i.e. all eligible patients including patients who died before they could receive treatment and patients for which a bespoke (autologous) treatment could not be produced.

The role of any further mandatory trial evidence

Manufacturers should provide details of mandated further studies (e.g. those in relation to conditional approvals or approvals made via the EMA's adaptive pathways approach). Future reports from the ADAPT SMART project should provide details about how the use of development plans across target populations being agreed up-front with EMA is working. Guidance may be needed regarding methodological approaches to utilising 'confirmatory' trial data in a related indication to update the decision NICE made for the original indication.

Consideration is also needed regarding the precise role NICE will play in EMA adaptive pathways processes. For example, what will be the mechanisms by which the EMA updates NICE with new efficacy and safety data for conditionally-approved ATMPs (in a timely way), and how will NICE deal with the new data (process-wise)?

Extrapolation approaches

Given the inevitable uncertainties which are likely to exist regarding the longer-term benefits of regenerative medicines and cell-based therapies, further methodological research could be usefully undertaken to help inform how these uncertainties might be appropriately quantified in a transparent manner to inform cost-effectiveness analyses. Further research may be particularly helpful to determine the appropriateness of alternative survival modelling approaches to regenerative medicines and cell-based therapies, including more flexible survival models and cure fraction models.

The level of data maturity is an important factor in deriving robust survival projections that are required for cost-effectiveness assessments. When follow-up is immature, a single 'best-fitting' survival distribution may not adequately characterise uncertainties over the longer-term extrapolation period. Although the robustness of the ICER estimates to alternative distributions can be explored through separate sensitivity analyses or scenarios, the transparency of the process may be impacted if the weighting of these is not explicitly considered in subsequent policy decisions. The feasibility and appropriateness of model-averaging approaches may also need to be more formally considered. The advantage of these approaches is that the parametric uncertainty associated within each distribution and the uncertainty (or weights) surrounding the choice of preferred method can be more explicitly characterised. However, given the potential complexity in both undertaking these analyses and communicating the results, efforts will need to be made to ensure models are developed to ensure that informal judgements can be explicitly incorporated in a timely and transparent manner.

Irrecoverable costs and possible learning curve effects

Given the complexity of the overall process of care that may be required for regenerative medicines and cell-based therapies, manufacturers will need to clearly report the resource and cost assumptions

of the different process to determine whether the full costs to the NHS have been included and any aspects where uncertainties may exist. Issues of irrecoverable costs may need to be more formally considered, particularly if a new technology could impose additional infrastructure requirements on the health system. If reimbursement decisions about the technology change before the end of the lifetime of the equipment (e.g. approval is withdrawn), then these costs may not be recovered and hence need to be explicitly considered.

The existence and possible impact of learning curves may also be as an important issue for clinical and cost-effectiveness assessments. Although the existence of learning curves has received attention in the clinical literature, the relevance of recent work in this area in the context of assessing the cost-effectiveness of medical devices should be considered.[92]

Quantification of decision uncertainty

Presentation of the scale of the consequences of decision uncertainty using population NHE may provide an important additional approach to quantifying decision uncertainty to the assessments already routinely specified with the exiting TA methods guide. The implications of existing research[88]and ongoing research by the DSU will also need to be considered by NICE to determine whether further changes to their processes or methods may be helpful for informing the nature of any additional assessments that may be required.

Such assessments could provide a more transparent and explicit basis for discussions between manufacturers, NICE and other relevant stakeholders in terms of how the existing uncertainties that exist might be appropriately managed, ensuring risks and benefits are more appropriately shared. Broader consideration will also need to be given to approaches which may extend beyond NICE's existing remit e.g. alternative payment schemes. Consequently other bodies and manufacturers themselves may also have an important role in identifying more innovative approaches to seeking reimbursement which recognise the inherent uncertainties and lead to a more efficient sharing of associated risk.

Existing criteria

NICE's existing processes also make separate provision for specific disease and technology characteristics which may be relevant to many regenerative medicines and cell therapies. Although NICE's current end of life criteria allows the Committee to explore a QALY weighting that is different from that of the reference case, the appropriateness of this criteria may need to be considered in relation to treatments which have curative potential. Further methodological research may also be important to determine whether an alternative weighting approach might be more appropriate for curative therapies. Existing research has identified a potential disconnect between individual and

societal preferences concerning valuation of treatment versus preventive interventions. Further research more specifically focused on the concept of cure may provide important additional insights.

Although the NICE methods guide permits the use of a non-reference-case discount to be applied in specific contexts, it remains unclear whether regenerative medicines and cell therapies would meet the existing criteria (e.g. uncertainties over the projected benefits and/or potentially significant irrecoverable costs). Consequently, NICE may need to provide additional guidance to ensure that manufacturers understand the likelihood of meeting these criteria.

12 References

1. Mason C, Dunnill P. A brief definition of regenerative medicine. *Regenerative Medicine* 2008;**3**:1-5.

2. House of Lords Science and Technology Committee. *Regenerative medicine report*. London: The Stationery Office Limited; 2013.

3. Regenerative Medicine Expert Group. *Building on your own potential: a UK pathway for regenerative medicine*: Department of Health; 2015.

4. European Medicines Agency. Committee for Medicinal Products for Human Use (CHMP). *Guideline on safety and efficacy follow-up risk management of advanced therapy medicinal products* London; 2008.

5. Edwards S, Lilford R, Braunholtz D, Jackson J, Hewison J, Thornton J. Ethical issues in the design and conduct of randomised controlled trials: a review. *Health Technol Assess* 1999;**2**.

6. Ashcroft R, Chadwick D, Clark S, Edwards R, Frith L, Hutton J. Implications of socio-cultural contexts for the ethics of clinical trials: a review. *Health Technol Assess* 1997;**1**.

7. Gupta S, Faughnan ME, Tomlinson GA, Bayoumi AM. A framework for applying unfamiliar trial designs in studies of rare diseases. *J Clin Epidemiol* 2011;**64**:1085-94.

8. European Medicines Agency. Committee for Medicinal Products for Human Use (CHMP). *Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design* London: EMA; 2007.

9. U.S. Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research (CDER). Center for Biologics Evaluation and Research (CBER). *Adaptive design clinical trials for drugs and biologics. Draft guidance*; 2010.

10. Tsimberidou A, Braiteh F, Stewart D, Kurzrock R. Ultimate fate of oncology drugs approved by the US Food and Drug Administration without a randomized trial. *J Clin Oncol* 2009;**27**:6243-50.

11. Baird LG, Banken R, Eichler HG, Kristensen FB, Lee DK, Lim JC, et al. Accelerated access to innovative medicines for patients in need. *Clin Pharmacol Ther* 2014;**96**:559-71.

12. European Medicines Agency. Human Medicines Research & Development Support. *Questions and answers following the initial experience of the Adaptive Licensing Pilot project. EMA/417706/2014*; 2014.

13. Lucas F. Medicines Adaptive Pathways to Patients (MAPPs) – opportunities and challenges in Europe. *Value & Outcomes Spotlight* 2015:11-3.

14. *ADAPTSMART. Accelerated Development of Appropriate Patient Therapies: a sustainable, multi-stakeholder approach from research to treatment-outcomes*. [cited 2015 11th September].

15. Commission Expert Group on Safe and Timely Access to Medicines for Patients ("STAMP"). Regulatory tools for early access: accelerated assessment procedure. In: *2nd Meeting of the STAMP Expert Group on 6 May*; 2015; Brussels. European Commission; 2015.

16. Van Bokkelen G, Morsy M, Kobayashi T-h. Demographic Transition, Health Care Challenges, and the Impact of Emerging International Regulatory Trends With Relevance to Regenerative Medicine. *Current Stem Cell Reports* 2015;**1**:102-9.

17. Sipp D. Conditional Approval: Japan Lowers the Bar for Regenerative Medicine Products *Cell Stem Cell* 2015;**16**:353-6.

18. Reuters. *German regulator puts UniQure gene therapy appraisal on hold*; 2015.

19. Benson K, Hartz AJ. A comparison of observational studies and randomized controlled trials. *N Engl J Med* 2000;**342**:1878-86.

20. Beynon R, Savovic J, Harris R, Altman D, Sterne J, Deeks J. Quantification of bias in the results of non-randomised studies compared with randomised studies. Poster presentation at the 16th Cochrane Colloquium: Evidence in the era of globalisation; 2008 Oct 3-7; Freiburg, Germany [abstract]. *Zeitschrift fur Evidenz, Fortbildung und Qualitat im Gesundheitswesen [serial online]* 2008;102 Suppl. 6:69. [cited

21. Dahabreh IJ, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *Eur Heart J* 2012;**33**:1893-901.

22. Hartz A, Bentler S, Charlton M, Lanska D, Butani Y, Soomro GM, et al. Assessing observational studies of medical treatments. *Emerg Themes Epidemiol* 2005;**2**:8.

23. Ioannidis JPA, Haidich A-B, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;**286**:821-30.

24. MacLehose R, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AMS. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess* 2000;**4**:1-162.

25. Shepherd J, Bagnall A-M, Colquitt J, Jacqueline D, Duffy S, Sowden A. *Sometimes similar, sometimes different: a systematic review of meta-analyses of randomised and non-randomised policy intervention studies*: Wessex Insitute for Health Research and Development. University of Southampton. Centre for Reviews and Dissemionation. University of York; 2006.

26. Concato J, Shah N, Horwitz RI. Randomized controlled trails observational studies and the hierarchy of research designs. *N Engl J Med* 2000;**342**:1887-92.

27. Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess* 1998;**2**:1-136.

28. Sacks H, Chalmers TC, Smith H, Jr. Randomized versus historical controls for clinical trials. *Am J Med* 1982;**72**:233-40.

29. Algra AM, Rothwell PM. Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomised trials. *Lancet Oncol* 2012;**13**:518-27.

30. Golder S, Loke YK, Bland M. Comparison of Pooled Risk Estimates for Adverse Effects from Different Observational Study Designs: Methodological Overview. *Plos One* 2013;**8**.

31. Lonjon G, Boutron I, Trinquart L, Ahmad N, Aim F, Nizard R, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. *Ann Surg* 2014;**259**:18-25. Available from: <Go to ISI>://MEDLINE:24096758

32. Abraham NS, Durairaj R, Young JM, Young CJ, Solomon MJ. How does an historic control study of a surgical procedure compare with the "gold standard"? *Dis Colon Rectum* 2006;**49**:1141-8.

33. Paulus JK, Dahabreh IJ, Balk EM, Avendano EE, Lau J, Ip S. Opportunities and challenges in using studies without a control group in comparative effectiveness reviews. *Res Synth Methods* 2014;**5**:152-61.

34. Vickers A, Ballen V, Scher H. Setting the bar in phase II trials: the use of historical data for determining "go/no go" decision for definitive phase III testing. *Clin Cancer Res* 2007;**13**:972-6.

35. Torgerson DJ, Torgerson C. *Designing randomised trials in health, education and the social sciences : an introduction*. Basingstoke, Hampshire: Palgrave Macmillan; 2008.

36. Sargent D, Taylor J. Current issues in oncology drug development, with a focus on Phase II trials. *J Biopharm Stat* 2009;**19**:556-62.

37. Korn E, Liu P, Lee S, Chapman J, Niedzwiecki D, Suman V, et al. Meta-analysis of phase II cooperative group trials in metastatic stage IV melanoma to determine progression-free and overall survival benchmarks for future phase II trials. *J Clin Oncol* 2008;**26**:527-34.

38. Philip P, Chansky K, LeBlanc M, Rubinstein L, Seymour L, Ivy S, et al. Historical controls for metastatic pancreatic cancer: benchmarks for planning and analyzing single-arm phase II trials. *Clin Cancer Res* 2014;**20**:4176-85.

39. Tang H, Foster NR, Grothey A, Ansell SM, Sargent DJ. Excessive false-positive errors in single-arm phase II trails: a simulation-based analysis. *J Clin Oncol* 2009;**25 Suppl. 15**:abstr 6512.

40. Pond GR, Abbasia S. Quantitative evaluation of single-arm versus randomized phase II cancer clinical trials. *Clin Trial* 2011;**8**:260-9.

41. Sambucini V. Comparison of single-arm vs. randomized phase II clinical trials: a Bayesian approach. *J Biopharm Stat* 2015;**25**:474-89.

42. Monzon J, Hay A, McDonald G, Pater J, Meyer R, Chen E, et al. Correlation of single arm versus randomised phase 2 oncology trial characteristics with phase 3 outcome. *Eur J Cancer* Forthcoming 2015.

43. Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol* 2012;**12**:9.

44. Unverzagta S, Prondzinsky R, Peinemannc F. Single-center trials tend to provide larger treatment effects than multicenter trials: a systematic review. *J Clin Epidemiol* 2013;**66**:1271-80.

45. Bellomo R, Warrillow SJ, Reade MC. Why we should be wary of single-center trials. *Crit Care Med* 2009;**37**:3114-9.

46. Sinhaa IP, Sinhab SK. Single-center trials in neonatology: issues to consider. *Semin Fetal Neonatal Med* 2015 [Forthcoming].

47. EUnetHTA. *Endpoints used in relative effectiveness assessment of pharmaceuticals: surrogate endpoints*; 2013.

48. Elston J, Taylor RS. Use of surrogate outcomes in cost-effectiveness models: a review of United Kingdom health technology assessment reports. *Int J Technol Assess Health Care* 2009;**25**:6-13.

49. Taylor RS, Elston J. The use of surrogate outcomes in model-based cost-effectiveness analyses: a survey of UK Health Technology Assessment reports. *Health Technol Assess* 2009;**13**:88.

50. Wilson MK, Karakasis K, Oza AM. Outcomes and endpoints in trials of cancer treatment: the past, present, and future. *Lancet Oncol* 2015;**16**:e32-e42.

51. Aziz A, Kempkensteffen C, May M, Lebentrau S, Burger M, Chun FK, et al. Prognostic, predictive and potential surrogate markers in castration-resistant prostate cancer. *Expert Rev Anticancer Ther* 2015;**15**:649-66.

52. Temple R. A regulatory authority's opinion about surrogate endpoints. *Clinical Measurement in Drug Evaluation* 1995;**3**:21.

53. Ciani O, Taylor RS. Surrogate, friend or foe? The need for case studies of the use of surrogate outcomes in cost-effectiveness analyses. *Health Econ* 2013;**22**:251-2.

54. Davis S, Tappenden P, Cantrell A. *A review of studies examining the relationship between progression-free survival and overall survival in advanced or metastatic cancer*. Sheffield: Decision Support Unit, ScHARR, University of Sheffield; 2012.

55. Gotzsche PC, Liberati A, Torri V, Rossetti L. Beware of surrogate outcome measures. *Int J Technol Assess Health Care* 1996;**12**:238-46.

56. Katz R. Biomarkers and surrogate markers: an FDA perspective. *NeuroRx: the American Society for Experimental NeuroTherapeutics* 2004;**1**:189–95.

57. Lerche la Cour.J, Brok.J, Gotzshe.P. Inconsistent reporting of surrogate outcomes in randomised clinical trials: cohort study. *BMJ* 2010;**341**.

58. Ciani O, Buyse M, Garside R, Pavey T, Stein K, Sterne JA, et al. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: meta-epidemiological study. *BMJ* 2013;**346**:f457.

59. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998;**54**:1014-29.

60. Bucher HC, Guyatt GH, Cook DJ, Holbrook A, McAlister FA. Users' guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. *JAMA* 1999;**282**:771–8.

61. Schievink B, Lambers Heerspink H, Leufkens H, De Zeeuw D, Hoekman J. The use of surrogate endpoints in regulating medicines for cardio-renal disease: opinions of stakeholders. *PLoS One* 2014;**9**:e108722.

62. Fleming TR, De Mets DL. Surrogate end points in clinical trials: are we being misled? *Ann Int Med* 1996;**125**:605-13

63. Lassere MN, Johnso KR, Boers M, Tugwell P, Brooks P, Simon L, et al. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. *J Rheumatol* 2007;**34**:607–15.

64. Berger VW. Does the Prentice criterion validate surrogate endpoints? *Stat Med* 2004;**23**:1571-8.

65. Baker SG, Kramer BS. Surrogate endpoint analysis: an exercise in extrapolation. *J Natl Cancer Inst* 2013;**105**:316-20.

66. Prasad V, Kim C, Burotto M, Vandross A. The strength of association between surrogate end points and survival in oncology: a systematic review of trial-level meta-analyses. *JAMA Intern Med* 2015;**175**:1389-98.

67. Ciani O, Davis S, Tappenden P, Garside R, Stein K, Cantrell A, et al. Validation of surrogate endpoints in advanced solid tumors: systematic review of statistical methods, results, and implications for policy makers. *Int J Technol Assess Health Care* 2014;**30**:312-24.

68. Ciani O, Buyse M, Garside R, Peters J, Saad ED, Stein K, et al. Meta-analyses of randomized controlled trials show suboptimal validity of surrogate outcomes for overall survival in advanced colorectal cancer. *J Clin Epidemiol* 2015;**68**:833-42.

69. Buckley S, Appelbaum F, Walter R. Prognostic and therapeutic implications of minimal residual disease at the time of transplantation in acute leukemia. *Bone Marrow Transplant* 2013;**48**:630-41.

70. Elorza I, Palacio C, Dapena JL, Gallur L, de Toledo JS, de Heredia CD. Relationship between minimal residual disease measured by multiparametric flow cytometry prior to allogeneic hematopoietic stem cell transplantation and outcome in children with acute lymphoblastic leukemia. *Haematologica* 2010;**95**:936-41.

71. Leung W, Pui CH, Coustan-Smith E, Yang J, Pei DQ, Gan K, et al. Detectable minimal residual disease before hematopoietic cell transplantation is prognostic but does not preclude cure for children with very-high-risk leukemia. *Blood* 2012;**120**:468-72.

72. Borowitz MJ, Devidas M, Hunger SP, Bowman WP, Carroll AJ, Carroll WL, et al. Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia and its relationship to other prognostic factors: a Children's Oncology Group study. *Blood* 2008;**111**:5477-85.

73. Eckert C, Hagedorn N, Sramkova L, Mann G, Panzer-Grumayer R, Peters C, et al. Monitoring minimal residual disease in children with high-risk relapses of acute lymphoblastic leukemia: prognostic relevance of early and late assessment. *Leukemia* 2015:1-8.

74. U.S. Food and Drug Administration. *Minimal Residual Disease (MRD) as a surrogate endpoint in Acute Lymphoblastic Leukemia (ALL) workshop. FDA briefing document*. Maryland: U.S. Food and Drug Administration; 2012.

75. Parker C, Waters R, Leighton C, Hancock J, Sutton R, Moorman AV, et al. Effect of mitoxantrone on outcome of children with first relapse of acute lymphoblastic leukaemia (ALL R3): An open-label randomised trial. *Lancet* 2010;**376**:2009-17.

76. Bujkiewicz S, Thompson JR, Sutton AJ, Cooper NJ, Harrison MJ, Symmons DP, et al. Use of Bayesian multivariate meta-analysis to estimate the HAQ for mapping onto the EQ-5D questionnaire in rheumatoid arthritis. *Value Health* 2014;**17**:109-15.

77. Sleijfer. S, A. W. The Challenge of Choosing Appropriate End Points in Single-Arm Phase II Studies of Rare Diseases. *J Clin Oncol* 2012.

78. National Institute for Health and Care Excellence. *Knee cartilage defects - autologous chondrocyte implantation [ID686]*. National Institute for Health and Care Excellence; [cited 6th May 2015].

79. National Institute for Health and Care Excellence. Autologous cartilage transplantation for full thickness cartilage defects in knee joints. *NICE technology appraisal guidance* 2000;**TA16**.

80. National Institute for Health and Care Excellence. The use of autologous chondrocyte implantation for the treatment of cartilage defects in the knee joints. *NICE technology appraisal guidance* 2005;**TA89**.

81. National Institute for Health and Care Excellence. *Sipuleucel-T for treating asymptomatic or minimally symptomatic metastatic hormone-relapsed prostate cancer. NICE technology appraisal guidance 332*. London: National Institute for Health and Care Excellence; 2015.

82. National institute for Health and Care Excellence. Guide to the methods of technology appraisal. *NICE: London* 2004.

83. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal. *NICE article* 2013;**PMG9**.

84. Mooney GH. *Economics, medicine and health care*: Pearson Education; 2003.

85. Sculpher M, Drummond M, Buxton M. The iterative use of economic evaluation as part of the process of health technology assessment. *J Health Serv Res Policy* 1997;**2**:26-30.

86. Fenwick E, Claxton K, Sculpher M, Briggs A. *Improving the efficiency and relevance of health technology assessment: the role of iterative decision analytic modelling*: University of York, Centre for Health Economics; 2000.

87. Towse A. *Regenerative medicine: a European HTA perspective*. In: ISPOR 17th Annual European Congress. Amsterdam; 2014.

88. Claxton K, Palmer S, Longworth L, Bojke L, Griffin S, McKenna C, et al. Informing a decision framework for when NICE should recommend the use of health technologies only in the context of an

appropriately designed programme of evidence development. Executive summary. *Health Technol Assess* 2012;**16**:1-8.

89. Faulkner E. *What Value Do We Place on a Cure? Value Demonstration Challenges Associated with Innovator and Regenerative Therapies in the EU, North America and Asia*. In: ISPOR 17th Annual European Congress. Amsterdam; 2014.

90. Husereau D. *What value do we place on cure?* In: ISPOR 17th Annual European Congress. Amsterdam; 2014.

91. Husereau D. How do we value a cure? *Expert Rev Pharmacoecon Outcomes Res* 2015;**15**:551-5.

92. McKenna C, Claxton K, Palmer S, Sculpher M, Epstein D, Iglesias C, et al. Evaluation of uncertainty and value of further research for devices – key methodological issues. *MedtecHTA, Milan* 2014;**WP5; 5.1**.

93. Walker S, Sculpher M, Claxton K, Palmer S. Coverage with evidence development, only in research, risk sharing, or patient access scheme? A framework for coverage decisions. *Value Health* 2012;**15**:570-9.

94. Gottlieb S, Carino T. Establishing new payment provisions for the high cost of curing disease. *AEI Research* 2014.

95. Edlin R, Hall P, Wallner K, McCabe C. Sharing risk between payer and provider by leasing health technologies: an affordable and effective reimbursement strategy for innovative technologies? *Value Health* 2014;**17**:438-44.

96. Westra TA, Parouty M, Brouwer WB, Beutels PH, Rogoza RM, Rozenbaum MH, et al. On discounting of health gains from human papillomavirus vaccination: effects of different approaches. *Value Health* 2012;**15**:562-7.

97. Jit M, Mibei W. Discounting in the evaluation of the cost-effectiveness of a vaccination programme: A critical review. *Vaccine* 2015;**33**:3788-94.

98. Mortimer D. Modelling downstream effects in the presence of technological change. *Pharmacoeconomics* 2008;**26**:991-1003.

99. Levine BL. Performance-enhancing drugs: design and production of redirected chimeric antigen receptor (CAR) T cells. *Cancer Gene Ther* 2015;**22**:79-84.

100. *Genetically Modified T Cell Therapies for Cancer - Basic Facts*. CATAPULT Cell Therapy; [cited 2015 14th September].

101. Hedden J. *CAR-T cell therapy CTL019 gets FDA breakthrough therapy status*. Datamonitor Healthcare; [cited 2015 14th September].

102. Maude SL, Teachey DT, Porter DL, Grupp SA. CD19-targeted chimeric antigen receptor T cell therapy for acute lymphoblastic leukemia. *Blood* 2015.

103. Maus MV, Grupp SA, Porter DL, June CH. Antibody-modified T cells: CARs take the front seat for hematologic malignancies. *Blood Cancer J* 2014;**123**:2625-35.

104. American Cancer Society. *How is childhood leukemia classified?* [cited 2015 7th September].

105. Cancer Research UK. *Acute lymphoblastic leukaemia (ALL) statistics*. [cited 2015 7th September].

106. Ko RH, Ji L, Barnette P, Bostrom B, Hutchinson R, Raetz E, et al. Outcome of patients treated for relapsed or refractory acute lymphoblastic leukemia: a Therapeutic Advances in Childhood Leukemia Consortium study. *J Clin Oncol* 2010;**28**:648-54.

107. Bassan R, Spinelli O. Minimal residual disease monitoring in adult ALL to determine therapy. *Curr Hematol Malig Rep* 2015;**10**:86-95.

108. Karen Seiter K, Besa EC. *Acute Lymphoblastic Leukemia (ALL) Guidelines* In: National Comprehensive Cancer Network Guidelines; 2014.

109. Kebriaei P, Poon M. Future of therapy in acute lymphoblastic leukemia (ALL)--potential role of immune-based therapies. *Curr Hematol Malig Rep* 2015;**10**:76-85.

110. Fielding AK, Richards SM, Chopra R, Lazarus HM, Litzow MR, Buck G, et al. Outcome of 609 adults after relapse of acute lymphoblastic leukemia (ALL); anMRCUKALL12/ECOG 2993 study. *Blood* 2007;**109**:944-50.

111. Fuster JL. Current approach to relapsed acute lymphoblastic leukemia in children. *World Journal of Hematology* 2014;**3**:49-70.

112. Saarinen-Pihkala U, Heilmann C, Winiarski J, Glomstein A, Abrahamsson J, Arvidson J, et al. Pathways through relapses and deaths of children with acute lymphoblastic leukemia: role of allogeneic stem-cell transplantation in Nordic data. *J Clin Oncol* 2006;**24**:5750-62.

113. Reismuller B, Peters C, Dworzak MN, Potschger U, Urban C, Meister B, et al. Outcome of children and adolescents with a second or third relapse of acute lymphoblastic leukemia (all): A population-based analysis of the austrian all-BFM (berlin-frankfurt-munster) study group. *J Pediatr Hematol Oncol* 2013;**35**:e200-e4.

114. von Stackelberg A, Völzke E, Kühl J, Seeger K, Schrauder A, Escherich G, et al. Outcome of children and adolescents with relapsed acute lymphoblastic leukaemia and non-response to salvage protocol therapy: a retrospective analysis of the ALL-REZ BFM Study Group. *Eur J Cancer* 2011;**47**:90-7.

115. Socié G, Veum Stone J, Wingard JR, Weisdorf D, Henslee-Downey PJ, Bredeson C, et al. Long-term survival and late deaths after allogeneic bone marrow transplantation. *N Engl J Med* 1999;**341**:14-21.

116. Bar M, Wood BL, Radich JP, Doney KC, Woolfrey AE, Delaney C, et al. Impact of minimal residual disease, detected by flow cytometry, on outcome of myeloablative hematopoietic cell transplantation for acute lymphoblastic leukemia. *Leukemia research and treatment* 2014;**2014**.

117. Lee DW, Kochenderfer JN, Stetler-Stevenson M, Cui YK, Delbrook C, Feldman SA, et al. T cells expressing CD19 chimeric antigen receptors for acute lymphoblastic leukaemia in children and young adults: a phase 1 dose-escalation trial. *Lancet* 2015;**385**:517-28.

118. Maude SL, Frey N, Shaw PA, Aplenc R, Barrett DM, Bunin NJ, et al. Chimeric antigen receptor T cells for sustained remissions in leukemia. *N Engl J Med* 2014;**371**:1507-17.

119. Davila ML, Riviere I, Wang X, Bartido S, Park J, Curran K, et al. Efficacy and toxicity management of 19-28z CAR T Cell therapy in B cell acute lymphoblastic leukemia. *Sci Transl Med* 2014;**6**:224ra25.

120. Grupp SA, Maude SL, Shaw P, Aplenc R, Barrett DM, Callahan C, et al. *380 T Cells Engineered with a Chimeric Antigen Receptor (CAR) Targeting CD19 (CTL019) Have Long Term Persistence and Induce Durable Remissions in Children with Relapsed, Refractory ALL*. In: 56th ASH Annual Meeting and Exposition. San Fansisco, CA; 2014.

121. Park JH, Riviere I, Wang X, Bernal Y, Purdon T, Halton E, et al. *Efficacy and safety of CD19-targeted 19-28z CAR modified T cells in adult patients with relapsed or refractory B-ALL*. In: ASCO Annual Meeting. Chicagi, Illinois; 2015.

122. Grupp SA. *CAR targeting CTL019 in Children with Relapsed, Refractory ALL*. In; 2014.

123. European Medicines Agency. *EVOLTRA (Clofarabine): EMEA/H/C/000613/Article 46 - 035*;

124. All Wales Medicines Strategy Group. Final Appraisal Report: Clofarabine (Evoltra®). 2007;Advice No: 0107 – June 2007.

125. *Historical data analysis of hematological remission and survival in adults with R/R acute lymphoblastic leukemia. NCT02003612*. U.S. National Institutes of Health; [cited 2015 22nd September].

126. Amgen Inc. *Blinatumomab (AMG 103). Background information for the Pediatric Subcommittee of the Oncologic Drugs Advisory Committee meeting 04 December 2012,*: U.S. Food and Drug Administration; 2012.

127. O'Brien S, Thomas D, Ravandi F, Faderl S, Cortes J, Borthakur G, et al. Outcome of adults with acute lymphocytic leukemia after second salvage therapy. *Cancer* 2008;**113**:3186-91.

128. Food and Drug Administration. Oncologic Drugs Advisory Committee. *Marqibo® (vincristine sulfate liposomes injection) for the treatment of advanced relapsed and/or refractory Philadelphia chromosome negative (ph-) adult acute lymphoblastic leukemia* Silver Spring, MD; 2012.

129. Goekbuget N, Topp MS, Zugmaier G, Viardot A, Stelljes M, Neumann S, et al. Anti-CD19 bite blinatumomab induces high complete remission rate in adult patients with relapsed bprecursor all: Updated results of an ongoing phase II trial. *Ann Oncol* 2012;**23**:i20-i1.

130. Kantarjian HM, Thomas D, Ravandi F, Faderl S, Jabbour E, Garcia-Manero G, et al. Defining the course and prognosis of adults with acute lymphocytic leukemia in first salvage after induction failure or short first remission duration. *Cancer* 2010;**116**:5568-74.

131. Jeha S, Gaynon PS, Razzouk BI, Franklin J, Kadota R, Shen V, et al. Phase II study of clofarabine in pediatric patients with refractory or relapsed acute lymphoblastic leukemia. *J Clin Oncol* 2006;**24**:1917-23.

132. Davila ML, Bouhassira DC, Park JH, Curran KJ, Smith EL, Pegram HJ, et al. Chimeric antigen receptors for the adoptive T cell therapy of hematologic malignancies. *Int J Hematol* 2014;**99**:361-71.

133. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* 2013.

134. Costa V, McGregor M, Laneuville P, Brophy JM. The cost-effectiveness of stem cell transplantations from unrelated donors in adult patients with acute leukemia. *Value Health* 2007;**10**:247-55.

135. Lis J, Kawalec P, Glasek M. Economic evaluation of acute lymphoblastic leukaemia treatment with clofarabine (Evoltra®) combined with chemotherapy for children and adolescents in Poland. *JHPOR* 2012;**2**:20-33.

136. National Institute for Health and Care Excellence. Leukaemia (acute lymphoblastic) - dasatinib. *NICE technology appraisal guidance* 2008;**ID386**.

137. All Wales Medicines Strategy Group. Final Appraisal Report: Nelarabine (Atriance®). 2009;**Advice No: 0909 – June 2009**.

138. Campana D, Leung W. Clinical significance of minimal residual disease in patients with acute leukaemia undergoing haematopoietic stem cell transplantation. *Br J Haematol* 2013;**162**:147-61.

139. Mody R, Li S, Dover DC, Sallan S, Leisenring W, Oeffinger KC, et al. Twenty-five–year follow-up among survivors of childhood acute lymphoblastic leukemia: a report from the Childhood Cancer Survivor Study. *Blood* 2008;**111**:5515-23.

140. MacArthur AC, Spinelli JJ, Rogers PC, Goddard KJ, Abanto ZU, McBride ML. Mortality among 5-year survivors of cancer diagnosed during childhood or adolescence in British Columbia, Canada. *Pediatr Blood Cancer* 2007;**48**:460-7.

141. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med* 2002;**21**:2175-97.

142. Jackson CH, Thompson SG, Sharples LD. Accounting for uncertainty in health economic decision models by using model averaging. *J R Stat Soc Ser A Stat Soc* 2009;**172**:383-404.

143. Reaves A. Can CAR-T Therapy Live Up To The Biotech IPO Hype? 2015.

144. Royal Surrey NHS Trust. FLAG-IDA dosing guide. 2015.

145. Tavil B, Aytac S, Balci Y, Unal S, Kuskonmaz B, Yetgin S, et al. Fludarabine, cytarabine, granulocyte colony-stimulating factor, and idarubicin (FLAG-IDA) for the treatment of children with poor-prognosis acute leukemia: the Hacettepe experience. *Pediatr Hematol Oncol* 2010;**27**:517-28.

146. UK Stem Cell Strategy Oversight Committee. Unrelated Donor Stem Cell Transplantation in the UK: Effective Affordable Sustainable. *NHS: Blood and Transplant* 2014.

147. London Specialised Commissioning Group. Adult Blood and Bone Marrow Transplant Services: Costing Model Guidelines. 2009.

148. Department of Health. NHS reference costs 2013 to 2014. 2015.

149. van Agthoven M, Groot MT, Verdonck LF, Lowenberg B, Schattenberg AV, Oudshoorn M, et al. Cost analysis of HLA-identical sibling and voluntary unrelated allogeneic bone marrow and peripheral blood stem cell transplantation in adults with acute myelocytic leukaemia or acute lymphoblastic leukaemia. *Bone Marrow Transplant* 2002;**30**:243-51.

150. Kelly MJ, Pauker SG, Parsons SK. Using nonrandomized studies to inform complex clinical decisions: the thorny issue of cranial radiation therapy for T-cell acute lymphoblastic leukemia. *Pediatr Blood Cancer* 2015;**62**:790-7.

151. van Litsenburg RR, Kunst A, Huisman J, Ket JC, Kaspers GJ, Gemke RJ. Health status utilities in pediatrics: a systematic review of acute lymphoblastic leukemia. *Med Decis Making* 2014;**34**:21-32.

152. Sung L, Buckstein R, Doyle JJ, Crump M, Detsky AS. Treatment options for patients with acute myeloid leukemia with a matched sibling donor: a decision analysis. *Cancer* 2003;**97**:592-600.

153. National Institute for Health and Care Excellence. NICE DSU TECHNICAL SUPPORT DOCUMENT 14: SURVIVAL ANALYSIS FOR ECONOMIC EVALUATIONS ALONGSIDE CLINICAL TRIALS - EXTRAPOLATION WITH PATIENT-LEVEL DATA. *NICE DSU Report* 2011.

154. Lambert PC, Thompson JR, Weston CL, Dickman PW. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics* 2007;**8**:576-94.

155. Bujkiewicz S, Jones HE, Lai MC, Cooper NJ, Hawkins N, Squires H, et al. Development of a transparent interactive decision interrogator to facilitate the decision-making process in health care. *Value Health* 2011;**14**:768-76.

156. Verde PE, Ohmann C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Research Synthesis Methods* 2015;**6**:45-62.

157. Doi SA. Evidence synthesis for medical decision making and the appropriate use of quality scores. *Clin Med Res* 2014;**12**:40-6.

158. Droitcour J, Silberman G, Chelimsky E. A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *Int J Technol Assess Health Care* 1993;**9**:440-9.

159. Jackson C, Best N, Richardson S. Improving ecological inference using individual-level data. *Stat Med* 2006;**25**:2136-59.

160. Jackson C, Best N, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *J R Stat Soc a Stat* 2008;**171**:159-78.

161. Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Stat Med* 2000;**19**:3359-76.

162. Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC. Models for potentially biased evidence in meta-analysis using empirically based priors. *J R Stat Soc Ser A* 2009;**172**:119-36.

163. Soares MO, Dumville JC, Ades AE, Welton NJ. Treatment comparisons for decision making: facing the problems of sparse and few data. *J R Stat Soc a Stat* 2014;**177**:259-79.

164. Spiegelhalter DJ, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Stat Med* 2003;**22**:3687-709.

165. Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. *J R Stat Soc Ser A* 2009;**172**:21-47.

166. McNamee R. Regression modelling and other methods to control confounding. *Occup Environ Med* 2005;**62**:500-6.

167. Harrell FE. *Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis*. New York ; London: Springer; 2001.

168. Bland JM, Altman DG. Matching .9. *BMJ* 1994;**309**:1128-.

169. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivar Behav Res* 2011;**46**:399-424.

170. Austin PC. Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The international journal of biostatistics* 2009;**5**:Article 13.

171. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med* 2007;**26**:734-53.

172. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Stat Med* 2006;**25**:2084-106.

173. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 2004;**23**:2937-60.

174. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol* 2003;**158**:280-7.

175. Crosby DA, Dowsett CJ, Gennetian LA, Huston AC. A tale of two methods: comparing regression and instrumental variables estimates of the effects of preschool child care type on the subsequent externalizing behavior of children in low-income families. *Dev Psychol* 2010;**46**:1030-48.

176. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol* 2006;**163**:262-70.

177. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. *Int J Epidemiol* 2008;**37**:1142-7.

178. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol* 2005;**58**:550-9.

179. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA* 2007;**297**:278-85. A

180. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol* 2006;**59**:437-47.

181. Laborde-Castérot H, Agrinier N, Thilly N. Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review. *J Clin Epidemiol* 2015;**68**:1232-40.

182. Biondi-Zoccai G, Romagnoli E, Agostoni P, Capodanno D, Castagno D, D'Ascenzo F, et al. Are propensity scores really superior to standard multivariable analysis? *Contemp Clin Trials* 2011;**32**:731-40.

183. Hamre HJ, Glockmann A, Kienle GS, Kiene H. Combined bias suppression in single-arm therapy studies. *J Eval Clin Pract* 2008;**14**:923-9.

184. Herson J. The use of surrogate endpoints in clinical trials (an introduction to a series of four papers). *Stat Med* 1989;**8**:403-4.

185. Ellenberg SS, Hamilton JM. Surrogate endpoints in clinical trials: cancer. *Stat Med* 1989;**8**:405-13.

186. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989;**8**:431-40.

187. The Meta-Analysis Group in Cancer. Modulation of fluorouracil by leucovorin in patients with advanced colorectal cancer: an updated meta-analysis. *J Clin Oncol* 2004;**22**:3766-75.

188. Holloway RG, Dick AW. Clinical trial end points. On the road to nowhere? *Neurology* 2002;**58**:679-86.

189. Freedman LS, Graubard BL, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med* 1992;**11**:167-78.

190. Zee J, Xie SX. Assessing treatment effects with surrogate survival outcomes using an internal validation subsample. *Clin Trial* 2015;**12**:333-41.

191. de Gruttola V, Fleming T, Lin DY, Coombs R. Perspective: validating surrogate markers - are we being naive? *J Infect Dis* 1997;**175**:237-46.

192. Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliairy endnpoints in clinical trials, with potential applications in cancer and AIDS research. *Stat Med* 1994;**13**:955-68.

193. Ciani O, Hoyle M, Pavey T, Cooper C, Garside R, Rudin C, et al. Complete cytogenetic response and major molecular response as surrogate outcomes for overall survival in first-line treatment of chronic myelogenous leukemia: a case study for technology appraisal on the basis of surrogate outcomes evidence. *Value Health* 2013;**16**:1081-90.

194. Hawkins N, Richardson G, Sutton AJ, Cooper NJ, Griffiths C, Rogers A, et al. Surrogates, meta-analysis and cost-effectiveness modelling: a combined analytic approach. *Health Econ* 2012;**21**:742-56.

195. Office for National Statistics. Cancer Registration Statistics, England, 2013. 2015.

196. American Cancer Society. *What are the key statistics about acute lymphocytic leukemia?* 2015. [cited 21/09/2015].

13 Appendices

13.1 Appendix 1: Regenerative medicines licensed by the EMA

Table 43 Regenerative medicines which are (or have been) licensed by the EMA

Glybera (Alipogene tiparvovec) *EMA Assessment (CAT and CHMP) 2012* *EMA marketing authorisation under exceptional circumstances*	
Nature of the Disease	
Indication	The indication initially applied for was: "Glybera is indicated for the long term correction of lipoprotein lipase deficiency, to control or abolish symptoms and prevent complications in adult patients clinically diagnosed with lipoprotein lipase deficiency (LPLD)". The indication for which a licence was granted is more restricted: "Glybera is indicated for adult patients diagnosed with familial lipoprotein lipase deficiency (LPLD) and suffering from at least one pancreatitis episode despite dietary fat restriction. The diagnosis of LPLD has to be confirmed by genetic testing. The indication is restricted to patients with detectable levels of LPL protein."
Orphan status?	Yes
Is this a rare condition?	The calculated prevalence of this condition was reported to be 0.02 per 10,000.
What is the natural history of the disease without this treatment/ with current treatment?	LPLD is a rare autosomal recessive inherited condition caused by homozygosity or compound heterozygosity for mutations in the LPL gene. The condition may only become evident after several episodes of pancreatitis in adolescence or adulthood. Laboratory investigation reveals genuine lactescent plasma (lipemia) due to the increased CM concentrations. The symptom severity is proportional to the degree of chylomicronemia and the most severe complication associated with LPLD is pancreatitis. Pancreatitis in an LPLD subject may lead to admission to an intensive care unit. In severe cases, patients may eventually develop chronic pancreatitis, ultimately resulting in endocrine and exocrine pancreatic insufficiency. Treatment of LPLD patients currently consists of severe reductions in dietary fat to less than 20% of caloric intake. Compliance with this dietary regimen is very difficult, and even with good compliance, the diet is often ineffective at reducing chylomicronemia and triglyceride levels. Currently no triglyceride-lowering drug is available. Enzyme replacement therapy is not expected to be effective, due to the short intravascular half-life of the LPL protein.
Nature of the medicine	
How does it work?	Glybera is a replication-deficient adeno-associated viral vector designed to deliver and express the human LPL gene variant LPLS447X. Transduction of part of the skeletal muscle mass is expected to restore a level of LPL activity which is sufficient to hydrolyse the triglyceride-rich lipoproteins, and influence lipid homoeostasis, and thus lead to clinical improvement or stabilisation.
Is it claiming to meet an otherwise unmet need?	Yes, the therapeutic aim of Glybera was to control symptoms of LPLD, and to prevent complications in adult patients clinically diagnosed with LPLD.

How is it given?	A sterile solution for injection presented as single use vials. Each vial contains 3 x 1012 genomic copies (gc) of alipogene tiparvovec (AAV1-LPLS447X) in 1ml of a phosphate based formulation buffer containing 5% sucrose. Glybera is to be administered **once** at multiple sites intramuscularly at a dose of 1 x 1012 gc per kg body weight. Note Glybera is intended as a single procedure but with multiple injections (up to 60 injection sites) administered under regional or spinal anaesthesia. All 27 patients reported adverse events related to the injection procedure.
Are there any comparator treatments?	Reducing chylomicronemia and triglyceride levels by reducing dietary fat to less than 20% of caloric intake.
Is there any mention of the intervention evolving over time?	The applicant uses two different company codes to differentiate between the current production system, AMT-011, versus the previous production system, which is referred to as AMT-010. There were changes during the development phases but CHMP felt issues relating to these had been resolved and 'consistency of product quality throughout development has been shown.'
Is there any mention of persistence of the treatment within the patient	**Negative persistence** - It is considered that although recombinant adeno-associated virus (rAAV) has potential integration risk, the risk of a consequent cancer is minimal. In the context of treating patients with this disease, these data suggest an acceptable safety profile. Overall, the CAT and CHMP agreed that the data do not substantiate a concern for tumourigenicity. **Positive persistence** – The post treatment observation period was insufficient to conclude on a rate of change of pancreatitis event long term. The totality of evidence derived from all studies combined suggested AMT-011 may temporarily reduce mean fasting TG levels but the proposed single treatment was insufficient to provide a durable and measurable effect.
Trial Design	
Trial description	

Study number	**Dose (gc/kg)**	**Number of patients**	**Duration of monitoring**	**Duration of follow-up**	**Status**
PREPARATION-01	None	18	13 – 78 weeks	-	Completed
AMT-010-01	1 x 1011 3 x 1011	4 4	12 weeks	5 years	Active phase completed, follow-up ongoing
PREPARATION-02	None	22	2 – 83 weeks	-	Completed
AMT-011-01	3 x 1011 1 x 1012	6 8	12 weeks	5 years	Active phase completed, follow-up ongoing
AMT-011-02	1 x 1012	5	18 weeks (incl. 4 weeks run-in)	1 year	Completed

There were two observational preparation studies to collect baseline data (no treatment control).

Glybera was studied in three uncontrolled, open-label interventional studies 9 CT-AMT-010-01, CT-AMT-011-01, and CT-AMT-011-02) with

	a combined total n=27. Three different dose regimens were evaluated in CT-AMT-010-01, CT-AMT-011-01, and CT-AMT-011-02. CT-AMT-011-02 was a safety and efficacy trial, initially planned as a controlled study it was subsequently amended to an uncontrolled study due to difficulties in identifying patients with high baseline risk of pancreatitis. It should be noted that a different Glybera product was used in the AMT-010 and AMT-011 trials due to a change in the manufacturing process. The first cohort in the AMT-011 trials (n=2 subjects) was administered 3 x 1011 gc/kg of AMT-011 to serve as a bridging arm to gauge similarity of the safety and efficacy of AMT-011 relative to AMT-010. CT-AMT-011-01 and CT-AMT-011-02 included an immunosuppressive regimen. CT-AMT-011-01 included a combination of cyclosporine A (3 mg/kg/day) and mycophenolate mofetil (2 g/day) which was given over 12 weeks. CT-AMT-011-02 had a modified regimen the same as CT-AMT-011-01 but also included a single bolus of methylprednisolone (single IV bolus 1mg/kg) half an hour before AMT-011 administration. Efficacy was assessed over 12 weeks, with long term follow up planned for 5 years. The analysis of pancreatitis events was attempted post-hoc by examining the number of events or admissions to ICU retrospectively as this was not a pre-specified analysis.
Trial population (adults/children/all?); any further specifics of disease not covered in 'indication'?	**Preparation 01;** 18 LPL-deficient patients ≥18 yrs with type I hyperchylomicronaemia, post-heparin LPL activity <25% of normal level, plasma concentrations of TG >95th percentile for age and gender. Seventeen subjects completed the study, 1 subject died of a cardiac arrest. **AMT-010-01**; 8/18 Patients from Preparation 01 cohort with confirmed homozygotic and compound heterozygotic LPL gene mutations. **Preparation-02** ; 22 subjects with LPLD, lipoprotein lipase activity ≤ 20 % of normal, LPL mass > 5 % of normal and fasting plasma TG concentrations > 10 mmol/l. Twenty subjects completed the study, 2 subjects withdrew. **AMT-011-01;** 15/22 subjects from Preparation-02, 1 subject was withdrawn thus 14 subjects entered the study long term. Follow up extended up to 5 years. **AMT-011-02;** 5 pts enrolled to examine pp-CM metabolism, fasting TG, serum LPL activity, pancreatitis. 1 patient only provided data.
Trial size/ Total trial population?	Combined total n=27
Length of follow up?	See table above
Control/comparator used?	The two observational studies (Preparation-01 and Preparation-02) that included patients receiving only diet reduction and no active treatment, acted as the control for the active treatment studies. NB some patients (not all) from Prep 1 and 2 went into the active treatment studies.
How is the control/comparator constructed?	See previous section
Outcomes	
Response outcome 1	**Prep studies -** Fasting plasma TG levels and disease complications in LPL deficient subjects on a low-fat diet. **Active treatment studies –** Across the three studies a measure of the reduction in fasting plasma triglyceride levels was a primary and secondary outcome: a reduction to < 10 mmol/l or to 40% of starting level. Decrease fasting plasma TG.
Response outcome 2	**Prep studies -** To record the incidence of pancreatic events in the context of the safety evaluation. **Active treatment studies -** A reduction in frequency and/or severity of clinical signs and symptoms related to LPL deficiency (i.e. eruptive

	xanthomas, lipaemia retinalis, pancreatitis, episodes of abdominal pain, plasma lactescence, lack of energy/fatigue and QoL and diabetes management). The incidence of pancreatitis was the most clinically meaningful endpoint.
Response outcome 3	Other measures of the effect of active treatment, e.g. clearance of chylomicrons and other determinants of the biological activity of lipoprotein lipase (LPLS447X) transgene product.
Adverse events	Overall, Glybera was well tolerated by all patients during initial 12 week observational period and during long-term phase of observation (up to 3 years with AMT-010 01). All reactions were self-limiting and mild in nature. There were no obvious serious adverse events seemingly related to Glybera.
Surrogate or intermediate clinical outcome? Yes/No	Yes. The effect on lipid profiles, such as a reduction in fasting triglycerides to <10 mmol/l, a >40% reduction in fasting triglycerides are surrogate markers of lipoprotein lipase activity related clinical benefit. A reduction in post-prandial chylomicronemia has been proposed as an alternative surrogate marker and subject to clinical validation a reduction in post-prandial CM could be accepted as a surrogate marker for efficacy.
Real clinical outcome? Yes/No	Yes **(**a reduction of pancreatitis events was suggested using retrospective data)
Summary of efficacy evidence	
Overall evidence base provided	**AMT-011-02**; is the only study yielding data allowing the possibility to make a link between surrogate and clinical endpoints (pp-CM metabolism, fasting TG, serum LPL activity, pancreatitis). Only one patient out of 5 responded to the treatment. The presented dataset in relation to the restricted indication includes 12 out of 27 patients treated with Glybera, aged 40-70 years of age and diagnosed with LPLD condition relatively late in life. The reduction in post-prandial chylomicronemia as an alternative surrogate marker for efficacy, although not at present validated, was considered biologically plausible and acceptable. The data on pancreatitis remain very limited and in a very small number of patients (12 patients) with limitations acknowledged in the statistical analysis. In summary, the evidence generated by the reduction of pancreatitis events and severity of attacks, although hampered by statistical limitations and by fluctuations in the occurrence of pancreatitis, suggested that Glybera leads to a clinically relevant reduction of pancreatitis risk at least in some patients. This is also supported by the reduction in hospital admissions and ICU stay. Of particular note is the fact that while about half of 17 patients required an ICU stay due to pancreatitis before treatment, no ICU stay was recorded in the same patients after treatment, as compared to non-treated patients.
Estimate of effect HrQoL?	The reduction in SF36 scores (those from both the physical functioning and mental domains) in 3 out of 5 patients from CT-AMT-011-02 study at week 14 following treatment was of major concern. The applicant explained the QoL reduction by adverse events and immunosuppression. However the data on Quality of Life from later time-points (up to week 52) and from all other studies conducted with Glybera are not available.
Other issues	
Any issues of scale-up for the product?	1. During the development of the AMT-011 process a number of changes have been made during scale up. 2. A two-tiered system has been established for commercial DS production based on a Master and Working Cell Bank (MCB, WCB) and Master and Working Viral Seed Stock (MSV, WSV).
Is further evidence requested for approval?	1. The MAH shall set up a long term surveillance programme/ disease register before launch of the product in each country to collect information on the epidemiology of the disease and the demographics, safety, and the effectiveness outcomes of patients treated with Glybera. The patients enrolled in clinical studies (CT-AMT-010 -10, CT-AMT 011-01, CT-AMT 011-02) should be followed up in the LPLD registry. 2. Assessment of postprandial chylomicron metabolism in at least 12 patients before and 12 months after treatment with Glybera to be chosen in addition to the patients included in study AMT.011.02; and eight healthy subjects in the second cohort. Assessment of immune response at baseline, 6 months and 12 months in at least 12 newly treated patients. The study should start by July 2013 and should enrol at least 4 patients per year.

	3. Re-evaluation of immune responses from all patients enrolled in study CT-AMT-011-01 by using a validated assay method should also be provided. 4. To improve the virus safety profile of the product 5. To complete the validation of the residual infectious baculovirus assay
Notes	Given the rarity of LPLD (prevalence in the EU: 2:1000000), the uncontrolled study design applied in all 3 clinical trials using subjects as their own control was accepted and in line with the scientific advice given. Development of studies was hampered by difficulties in recruitment of sufficient numbers of patients. The SAG considered that it is not possible to exclude completely the hypothesis that the reduction in the incidence of pancreatitis in some patients is due to the inherent temporal rarity of pancreatitis events. Issues inherent to retrospective data assessment in comparison to prospective data were highlighted by the CAT. Across the three active treatment trials the primary and secondary outcomes were not the same

MACI - matrix applied characterised autologous cultured chondrocyte implant *EMA Assessment 2013 (CHMP and CAT), NICE MTA 2014* *EMA marketing authorisation in April 2013 which was subsequently suspended in September 2014* *(an authorised manufacturing site no longer existed)*	
Nature of the Disease	
Indication	MACI is to be used in skeletally mature patients for the repair of symptomatic cartilage defects of the knee (grade III and IV of the arthroscopic staging of osteochondral lesions as described by the Modified Outerbridge Scale).
Orphan status	No
Is this a rare condition?	No. Cartilage injuries were observed 5-11% of diagnostic knee arthroscopies in predominantly young adult populations with knee pain.
What is the natural history of the disease without this treatment/ with current treatment	Cartilage defects of the knee occur along a spectrum of disease and severity. Larger, more chronic lesions are often symptomatic, may contribute to joint misalignment and can cause disabling symptoms such as pain, catching, locking, and swelling. Focal chondral lesions that are left untreated may progress to debilitating joint pain, dysfunction and degenerative arthritis.
Nature of the medicine	
How does it work?	It is the first advanced-therapy medicine to be combined with a medical device, in this case where the cells are embedded in a biodegradable matrix. It attempts to generate hyaline or hyaline-like cartilage. ACI requires two surgical procedures, first to harvest autologous chondrocytes, which are then grown extra-corporeally, and then to transplant the cultivated cells back into the lesions. The benefit of ACI over other restoration techniques is that larger lesions can be treated.
Is it claiming to meet an otherwise unmet need?	No, other treatment options (such as microfracture) exist and clinical practice varies.
How is it given?	Autologous chondrocytes are seeded onto a collagen membrane of porcine origin, which is secured into the lesion with fibrin glue. At implantation, the membrane is trimmed to the correct size and shape, and implanted cell-side down into the base of the defect; the implant is secured in place using fibrin sealant. The recommended dose of MACI implant is 500,000 to 1 million cells per cm^2 of defect. The dose is the same for all patients, regardless of age.
Are there any comparator treatments?	Repair techniques such as microfracture aim at marrow stimulation and induce the formation of fibrocartilage repair tissue to treat patients with focal chondral defects in the knee. These techniques penetrate the subchondral bone and cause release of marrow components into the defect site. The reparative response produced from these procedures is one that may generate primarily fibro-cartilage. Single-stage restoration techniques such as osteochondral autograft, mosaicplasty, and osteochondral allograft attempt to replace the cartilage defect with host or donor articular cartilage.
Is there any mention of the intervention evolving over time?	Yes - MACI is a third generation ACI product. ERG report (for NICE): "There is a general problem when long-term results are needed but the technology continues to evolve."

Is there any mention of persistence of the treatment within the patient	In concept, the MACI implant would contribute to the repair of articular cartilage defects through proliferation of seeded chondrocytes, resulting in synthesis of hyaline-like repair tissue.	
Trial Design		
Trial description	The clinical data consist of the pivotal trial "SUMMIT" (MACI00206) supported by several clinical studies reported from the literature.	The Academia studies were small-scale, non-randomised prospective studies
	SUMMIT – Superiority of MACI Versus Microfracture Treatment Trial Multi-centre, randomised, open-label parallel-group trial. The aim of this trial was to demonstrate the superiority of MACI implant versus arthroscopic microfracture for the treatment of symptomatic articular cartilage defects of the femoral condyle, including the trochlea.	
Trial population (adults/children/all?); any further specifics of disease not covered in 'indication'?	**MACI00206** Male and female patients between the ages of 18 and 55 years (inclusive), with at least 1 symptomatic outerbridge grade III or IV focal cartilage defect on the medial femoral condyle (MFC), lateral femoral condyle (LFC) and/or trochlea (defect size equal to or greater than 3.0 cm^2 irrespective of location.	
Trial size/ Total trial population?	144 patients 72 patients MACI/ 72 patients microfracture	
Length of follow up?	2 year follow up data already collected from the MACI00206 study (5 year follow up planned)	
Control/comparator used?	Microfracture treatment	
How is the control/comparator constructed?	RCT	
Outcomes		
Response outcome 1	Co-primary endpoint of KOOS (Knee Injury and Osteoarthritis Outcome Score) for Pain and Function (sports and recreational activities).	
Response outcome 2	Secondary endpoints Histology of cartilage forming (Histological evaluation of structural repair of evaluable biopsies harvested from the core of the index lesion during arthroscopy).	
Response outcome 3	MRI of cartilage - MRI assessments of structural repair parameters.	
Adverse events	Most AEs were thought to be surgery related, rather than product related.	
Surrogate or intermediate clinical outcome?	Yes - Structural and functional repair of cartilage defects as measured by MRI or histology scoring	
Real clinical outcome?	Yes - Knee Injury and Osteoarthritis Outcome Score (KOOS)	
Summary of evidence		
Overall evidence base provided	A clinically and statistically significant difference in the improvement from baseline to Week 104 was seen for the co-primary endpoint of	

	KOOS (Knee Injury and Osteoarthritis Outcome Score) for Pain and Function in patients treated with MACI over the comparator (p=0.001). Significantly more patients treated with MACI (87.50%) met the responder analysis criteria than patients treated with microfracture (68.06%), which is considered clinically relevant. The primary efficacy endpoint was corroborated by several other patient reported outcome measures and a responder analysis of the primary efficacy measures demonstrated superior clinical efficacy for patients treated with MACI compared to microfracture.
Estimate of HrQoL?	Knee-related quality of life is one of the 5 key dimensions of KOOS, although the NICE report highlighted the 'lack of good quality of life data'.
Other issues	
Any issues of scale-up for the product?	The manufacture of the product is patient-specific (autologous). Production will be centralised at one site.
Is further evidence requested for EMA/FDA approval?	1. As part of the ongoing monitoring of MACI, the Agency requested the 5-year follow-up data from the main clinical study, which will provide information on the sustainability of the cartilage repair and maintenance of effect of MACI compared to microfracture over time, as well as the long-term safety of the medicine. 2. Periodic safety updates report for this product within 6 months following authorisation and a risk management plan. 3. Education pack for surgeons
Any additional information provided?	1. The applicant sought advice from the EMA/CHMP regarding the design of the trial. 2. MACI has now been recommended for licensing as the first advanced-therapy medicine to be combined with a medical device. 3. Surgical skill was identified as being an important issue; MACI will likely only be available from a few specialised centres. The marketing authorisation for MACI was suspended in September 2014 as an authorised manufacturing site no longer existed (the developer closed the site).

ChondroCelect – Characterised viable autologous cartilage cells expanded ex vivo expressing specific marker proteins *EMA Assessment 2009, NICE MTA 2014* *EMA marketing authorisation*	
Nature of the Disease	
Indication	The indication for ChondroCelect is repair of single symptomatic cartilaginous defects of the femoral condyle of the knee (International Cartilage Repair Society grade III or IV) in adults.
Orphan status	No
Is this a rare condition?	No
What is the natural history of the disease without this treatment/ with current treatment?	The healing capacity of articular cartilage is poor and damaged articular cartilage is thought to be a precursor to the development of osteoarthritis. Damaged articular cartilage can result in pain, loss of joint function and disability. An early intervention on symptomatic cartilage lesions may prevent or delay irreversible changes in the joint surface. Currently, there is no uniform approach to managing significant knee cartilage defects.
Nature of the medicine	
How does it work?	ChondroCelect is a suspension of approximately 10,000 autologous cartilage cells per microlitre of medium for autologous use. The cells have been obtained by ex vivo expansion of chondrocytes isolated from a biopsy of the articular cartilage from the patient's knee. The active substance is a centrifuged pellet of 4 to 12 million cells that were expanded ex vivo, harvested and washed. The expansion process is designed to preserve the integrity and function of the cells and particularly to maintain the cells' ability to produce hyaline cartilage.
Is it claiming to meet an otherwise unmet need?	No. Other treatment options exist, and clinical practice varies.
How is it given?	In the first step a cartilage biopsy is obtained arthroscopically from healthy articular cartilage from a lesser weight bearing area of the patient's knee, approximately 4 weeks prior to implantation. Chondrocytes are isolated from the biopsy by enzymatic digestion, expanded in vitro, characterised and delivered as a suspension of 1 x 104 cells/µl for implantation in the same patient. During the second step of the procedure the expanded chondrocyte suspension is implanted during open-knee surgery.
Are there any comparator treatments?	Repair techniques such as microfracture aim at marrow stimulation and induce the formation of fibrocartilage repair tissue to treat patients with focal chondral defects in the knee. These techniques penetrate the subchondral bone and cause release of marrow components into the defect site. The reparative response produced from these procedures is one that may generate primarily fibro-cartilage. Single-stage restoration techniques such as osteochondral autograft, mosaicplasty, and osteochondral allograft attempt to replace the cartilage defect with host or donor articular cartilage.
Is there any mention of the intervention evolving over time?	Yes - ChondroCelect is a third generation ACI (autologous chondrocyte implantation) product.
Is there any mention of persistence of the treatment within the patient	Implanted cells become a structural part of newly formed cartilage.
Trial Design	
Trial description	Study TIG/ACT/01/2000 is a phase III, multicentre, randomized, controlled trial to compare ChondroCelect to microfracture in the repair of symptomatic single cartilaginous lesions of the femoral condyles of the knee.

	Supportive study: Prospective, long-term follow-up study of patients in the Belgian Armed Forces treated with ChondroCelect (TIG/ACT/02)
Trial population (adults/children/all?); any further specifics of disease not covered in 'indication'?	TIG/ACT/01/2000 Patients aged between 18 and 50 years, who had a single symptomatic cartilage lesion between 1-5cm2 of the femoral condyles met the inclusion criteria. TIG/ACT/02; This study is a prospective, non-comparative, open-label study of 2 to 5 years' duration in 20 patients with single and multiple symptomatic cartilage defects, in any location of the knee, who underwent CCI using ChondroCelect.
Trial size/ Total trial population?	TIG/ACT/01/2000; 118 participants, n=57 ChondroCelect & n=61 microfracture TIG/ACT/02; Of all reported lesions, 80% were reported to be of ICRS Grade III or IV. Of 24 femoral lesions reported in 19 patients, 21 were treated with CCI.
Length of follow up?	TIG/ACT/01/2000 = 12 months extended to 36 months for AEs TIG/ACT/02 = 5 years
Control/comparator used?	Microfracture is considered an effective standard treatment for smaller femoral cartilage lesions according to currently available literature data, and is an acceptable control therapy.
How is the control/comparator constructed? Source of comparative data? Confounding?	RCT
Outcomes	
Response outcome 1	Knee Injury and Osteoarthritis Outcome Score (KOOS)
Response outcome 2	Structural repair
Adverse events	The overall safety summary showed that the main difference in treatment related adverse events compared to microfracture was related to the open knee surgery (arthrotomy) which caused an increase in joint swelling and possible joint effusion. Cartilage hypertrophy can be reduced by using a biomembrane to cover the lesion, and will therefore not pose a major safety concern in future applications of ChondroCelect. However, a higher number of patients in the microfracture arm have a treatment failure and require a subsequent surgical intervention. Therefore the short and long term complication rate is not higher for ChondroCelect compared to microfracture.
Surrogate or intermediate clinical outcome? Yes/No	Yes, structural repair (histological analysis).
Real clinical outcome? Yes/No	Yes (KOOS)
Summary of efficacy evidence	
Overall evidence base provided	The mean change in overall KOOS from baseline to the average of 12 to 18 months was slightly higher for patients in the ChondroCelect group than for patients in the microfracture group. The results fulfil the predefined criteria for non-inferiority and changes are clinically relevant. Results of the histological analysis of structural repair at 12 months favoured ChondroCelect and the difference was statistically significant for both qualitative and quantitative analysis. It was, however, acknowledged that this end point was not in compliance with GCP as it was developed during the conduct of the study as the original a priori determined primary efficacy point was considered as invalid.
Estimate/ measure of effect (e.g. HrQoL)	Knee-related quality of life is one of the 5 key dimensions of KOOS, although the NICE report highlighted the 'lack of good quality of life data'.
Other issues	
Any issues of scale-up for the	The manufacture of the product is patient-specific (autologous). Production will be centralised at one site.

product?	
Is further evidence requested for approval?	The GCP inspection highlighted the amount of missing data on the structural endpoint and the change to the ICRSII read-out in the pivotal study as major concerns. The CAT, considered the following particular causes for concern: • There were deficiencies in the conduct of the pre-authorisation studies and uncertainties related to the result of the submitted single pivotal trial. • There is unknown long-term durability of the product efficacy. • Benefit/risk of the product is significantly influenced by the level of compliance with the defined procedures throughout the treatment with ChondroCelect, from biopsy harvest to receiving correct physiotherapy. The CAT also considered that performing of post-authorisation studies will need to be a part of the Pharmacovigilance plan and Efficacy follow-up plan presented in the Risk Management Plan. The CHMP agreed with the above.

Holoclar (Ex vivo expanded autologous human corneal epithelial cells containing stem cells) *EMA Assessment 2014 (CHMP ,CAT, COMP)* *EMA conditional marketing authorisation*	
Nature of the Disease	
Indication	Corneal lesions, with associated (limbal) stem cell deficiency, due to ocular burns. The clinical spectrum of limbal stem cell deficiency (LSCD) includes pain, photophobia, inflammation, corneal neovascularisation, and eventually, the reduction or complete loss of visual acuity.
Orphan status	Designated as an orphan medicinal product (2008) in the following indications: Corneal lesions, with associated (limbal) stem cell deficiency due to ocular burns.
Is this a rare condition?	The condition is considered to be rare with an estimated prevalence of 0.34 per 10,000.
What is the natural history of the disease without this treatment/ with current treatment	If left untreated, the condition may progress to a stage whereby persistent epithelial defects present with an associated high risk for the development of bacterial keratitis, corneal perforation and blindness.
Nature of the medicine	
How does it work?	Holoclar is specifically 'Ex-vivo expanded autologous human corneal epithelial cells containing stem cells' and replaces damaged corneal epithelium cells and creates a reservoir of limbic stem cells (LSC) in LSC deficient areas of the cornea for continuous regeneration. Transparent circular sheet of living tissue containing autologous human corneal epithelial cells, limbal stem cells and derived transient amplifying cells.
Is it claiming to meet an otherwise unmet need?	Yes the product claims to respond to an unmet medical need by providing a new active substance to treat patients with irreversible and extensive damage as a result of an ocular burn. At the time of application, no medicinal products had been approved in the European Union/European Economic Area (EU/EEA) for this indication and there was no gold-standard in treatment.
How is it given?	Single topical placement without systemic effect
Are there any comparator treatments?	Limbal allografts which have associated risk of rejection and which require long-term systemic immunosuppression. Non-expanded limbal autografts from the healthy fellow eye which may lead to iatrogenic induction of LSCD in the donor eye
Is there any mention of the intervention evolving over time?'	No
Is there any mention of persistence of the treatment within the patient (keyword search)?	Negative effects of persistence - Possible risks include systemic distribution of cells derived from Holoclar that are tumour forming, accelerated immune response or transmission of adventitious agents. The cells are not expected to migrate beyond the ocular surface, or to produce systemic effects. Tumourigenicity was investigated in-vivo and results suggested a low risk. Positive effects of persistence - Some information on the potential for biodistribution was derived from a historical dataset. Data from a histological and morphological evaluation of corneal material collected from 26 patients who had undergone perforating keratoplasty post limbal stem cell transplantation with Holoclar.Available long-term follow up data up to 10 years after ACLSCT, though limited, supported persistence of treatment success beyond 12 months. Additional long-term efficacy data will be collected in the margins of a post-authorisation safety study to confirm this outcome.
Trial Design	
Trial description	Multi-centre retrospective observational case series.

	Primary efficacy/safety study and supportive study.		
	HLSTM01 **1998-2007**	**HLSTM02** **1998-2007**	**HLSTM04** **2008-2013**
Trial population (adults/children/all?); any further specifics of disease not covered in 'indication'?	Male or female with moderate/severe LSCD Median age 49, mostly adults	Male or female Median age 43.5, mostly adults	15 patients treated from 2008 (from additional centres not originally provided as a part of HLSTM01)
Trial size/ Total trial population?	104 patients with moderate/severe LSCD 2 centres	29 patients with moderate/severe LSCD 7 centres	15 patients with moderate/severe LSCD 3 centres
Length of follow up?	12months post intervention assessment, max 10 year follow up. 28% pts 1-2 years, 22% pts 2 – 3 years, 12% pts for 5 years or more to a maximum of 10 years post transplantation. After year 5, only 5 patients had long term follow-up, of which 4 were reported as continued treatment success.		
Control/comparator used? (i.e. what were the results compared to for validation such as improvement compared to no treatment or historical trials using a different treatment)	Patients acted as their own controls – outcomes were compared with baseline data.		
How is the control/comparator constructed? Source of comparative data? Confounding?	See above. The assumption that the condition would not heal was accepted – so any healing could be ascribed to Holoclar.		
Outcome			
Response outcome 1	Successful transplant at 12 months based on the co-presence of clinical signs (i) a superficial corneal neovascularisation classified as 'NONE' or 'MILD' and (ii) epithelial defects classified as 'NONE' or 'TRACE'		
Response outcome 2	Symptomatic relief (pain, burning, photophobia)		
Response outcome 3	Improvement in visual acuity or visual stabilisation at month 12 verses baseline		
Surrogate or intermediate clinical outcome?	Yes - Corneal epithelial integrity and Absence of significant corneal neovascularisation		
Real clinical outcome?	Yes - Improved Visual acuity		
Adverse events	Eye-related disorders were the most commonly observed adverse events occurring in 57% of the safety population. Overall the rate of serious adverse events was low. Out of a total of 11 SAEs, three were judged as related to administration of Holoclar.		
Summary of efficacy evidence			
Overall evidence base provided	1. Significant decrease ($p<0.001$) in ocular symptoms (reduction from 40 pts to 12pts with ocular symptoms) 2. No change in inflammation by the 12 month endpoint, 32 pre-surgical pts to 33 post-surgical pts. 3. Superficial corneal neovascularisation (CNV) evaluated before and after the transplantation, 73.1% of patients showed an improvement at 12 months post transplantation, and a significant decrease in CNV from baseline to 12 months post-surgery ($p<0.001$) 4. 83.6% showed a reduced epithelial defect of none or trace. 5. An improvement in visual acuity was noted in 51 (49%) of patients.		

	6. The majority of patients required only one transplant for a successful outcome 7. Persistent success of limbal cell transplantation after keratoplasty, 14 patients had one prior transplantation & 2 patients had 2 previous transplants. 8. 57 pts had keratoplasty subsequent to the Holoclar graft and success was achieved in 24 pts. 9. The results of impression cytology in a subset of patients in the pivotal studies, showed an increase of the percentage of keratinocytes and a decrease of the percentage of conjunctival cells after Holoclar treatment, thus providing evidence that Holoclar enables corneal type epithelialisation of the ocular surface and exerts a regenerative effect.
Estimate of HrQoL	Not assessed
Other issues	
Any issues of scale-up for the product?	Clinical success depends on factors unique to cell therapies, including manufacturing procedures, clinical and pharmacologic standardization of protocols, and regulation. Manufacture of the active substance is patient-specific and the manufacturing process is state-of-the-art and highly complex. As such the applicant implemented a training program for surgeons to ensure collection of seed material, and a structured approach to manufacturing comprising many monitored stages and sub-stage in-process controls (IPCs). The applicant was also required to provide further evidence on stability of the product (integrity and viability) and transport information.
Is further evidence requested for EMA/FDA approval?	A multinational, multicentre, prospective, open-label, uncontrolled interventional study to assess the efficacy and safety of autologous cultivated limbal stem cells grafting for restoration of corneal epithelium in patients with limbal stem cell deficiency due to ocular burns was required by December 2020. A major objection was raised with regard to the proliferation of irradiated cells and further validation was requested. Evidence was provided in the form of a demonstration of several methods to show the irradiated cells do not proliferate. Paediatric application was deferred at time of submission pending further measures.
Notes	At the time of application more than 200 patients had already been treated with Holoclar in clinical practice since 1998, however many clinics declined the request to provide data. The assessors considered this may introduce bias but felt that the supporting literature and the similarity of the findings reported to those in the published articles provided some confidence in the numbers and therefore they were happy to allow the data. Supportive data were also considered by the CAT from published articles and this appears to have a strong influence on the decision making process, although only provided supportive information. As the condition was considered to have a low incidence the small sample size was considered to be acceptable. The CAT noted that at baseline the majority of patients already presented with no or only trace epithelial defects and as such already presented with a successful treatment outcome. However they considered that LSCD is a condition with impaired ability to maintain or restore an intact corneal epithelium so defects over the follow-up period were considered clinically relevant. The fact that the studies were uncontrolled and not randomised further added to the uncertainties of the validity of the dataset, but was considered inevitable due to the lack of a suitable comparator considering that there is neither an approved treatment for LSCD nor an ubiquitous accepted standard of care. Since this condition would not heal spontaneously, the single arm, uncontrolled design was considered acceptable by the CAT

Provenge (sipuleucel-T / Autologous peripheral blood mononuclear cells activated with PAP-GM-CSF) *EMA assessment 2013 (CAT and CHMP), NICE STA 2014, FDA assessment 2009* *EMA marketing authorisation in June 2013 which was withdrawn in May 2015 at the request of the manufacturer for commercial reasons*			
Nature of the Disease			
Indication	Asymptomatic or minimally symptomatic metastatic (non-visceral) hormone-relapsed prostate cancer in men for whom chemotherapy is not yet clinically indicated.		
Orphan status?	No		
Is this a rare condition?	Hormone refractory metastatic prostate cancer affects around 5000 patients/year in the UK		
What is the natural history of the disease without this treatment/ with current treatment?	Asymptomatic patients have a median overall survival of 18-24 months. Patients with symptomatic disease have a median OS of 9-16 months.		
Nature of the medicine			
How does it work?	Sipuleucel-T is an autologous active cellular immunotherapy product designed to stimulate an antigen (CD59) immune response to prostate cancer. Patients' peripheral blood mononuclear cells are incubated with a recombinant fusion protein, the prostate protein prostatic acid phosphatase (PAP).		
Is it claiming to meet an otherwise unmet need?	No		
How is it given?	Following blood sampling leukapheresis is performed (day 1) after which Sipuleucel-T is manufactured at a central facility (days 2-3) and then infused into the patient (day 3 or 4). This process happens three times, at approximately two-week intervals.		
Are there any comparator treatments?	Best supportive care (radiotherapy, bisphosphonates, steroid, analgesics, active surveillance), abiraterone acetate		
Is there any mention of the intervention evolving over time?	No		
Is there any mention of persistence of the treatment within the patient?	No. The achievement and maintenance of the antigen response was assessed – maximum duration tested was 26 weeks in one trial. There was no clear indication of whether or not persistence was required for benefit. No adverse effects related to persistence of antigen response were mentioned.		
Trial Design			
Trial description	**D9902B IMPACT**	**D9902A**	**D9901**
	Multi-centre RCT (with cross-over allowed after progression) using a 2:1 ratio (favouring allocation to Sipuleucel-T)	As for IMPACT trial	As for IMPACT trial
Trial population (adults/children/all?); any further specifics of disease not covered in 'indication'?	Asymptomatic or minimally symptomatic metastatic hormone-relapsed prostate cancer	Asymptomatic metastatic hormone-relapsed prostate cancer	Asymptomatic metastatic hormone-relapsed prostate cancer
Trial size/ Total trial population?	512	98	127
Length of follow up?	3 years (Follow-up was planned to continue until the number of events (deaths) reached	3 years	3 years

	that required by the analysis plan)		
Control/comparator used?	Placebo, consisting of one third of the patient's cells being re-infused, but the cells have not been activated with the fusion protein; the remaining two-thirds were cryopreserved.	As for IMPACT study	As for IMPACT study
How is the control/comparator constructed? Source of comparative data? Confounding?	Following confirmation of disease progression, placebo patients could receive activated cells (i.e. very similar to Sipuleucel-T) derived from their cryopreserved cells. Open-label phase.	As for IMPACT study	As for IMPACT study
Outcomes			
Response outcome 1	Overall survival	Time to disease progression	Time to disease progression
Response outcome 2	Time to objective disease progression	Overall survival	Time to onset of disease-related pain
Response outcome 3	Safety	Time to objective disease progression	Grade 3 AEs
Surrogate or intermediate clinical outcome? Yes/No	Time to progression. Antigen response was also measured. Note this did not correlate with OS results		
Real clinical outcome? Yes/No	Overall survival		
Adverse events?	Overall, the leukapheresis procedure and Provenge infusions were well tolerated. The main risks identified were acute infusion reactions, toxicities (e.g., citrate toxicity) associated with the leukapheresis procedure and infections (principally associated with catheters). Treatment with Provenge may lead to unwanted long term immunological effects in the body system. This potential risk is adequately addressed in the risk management plan. Additional data will become available to further characterise the long term safety profile of Provenge through registries.		
Summary of efficacy evidence			
Overall evidence base provided	For the IMPACT trial: overall survival was significantly improved with Sipuleucel-T, HR 0.8 (95% CI 0.61 to 0.98, p=0.03) but there was no difference in time to objective disease progression HR 0.95, 95% CI 0.77 to 1.17, p=0.63). Two trials reported a significant advantage in OS favouring Sipuleucel-T, although no significant differences in time to disease progression was seen in any of the three trials. The RCTs had low risk of bias, but only up to the point of disease progression, after which crossover from placebo to active was permitted. No analyses was performed to adjust for cross-over. Also the lack of significant effect on PfS may also have been due to a delay in effect.		
Estimate of HrQoL	Not assessed		
Other issues			
Any issues of scale-up for the product?	Yes, patient cells must be transferred from their local hospital to a central manufacturing facility, and then back again to the local hospital. The final product has a short shelf life.		
Is further evidence requested for approval?	Periodic safety update reports		
Notes	13 members of the CHMP did not agree with the CHMPs recommendation and the granting of a marketing authorisation. The objections were based around whether the differences in OS resulted from a true and clinically relevant effect of Sipuleucel-T. The effect was neither supported by PFS nor time to progression results. Importantly, in case of disagreement between these outcomes the efficacy evidence should be particularly convincing and ideally corroborated by other secondary endpoints, which was not the case. There was a lower proportion of patients treated with docetaxel in the placebo group and also delayed treatment with docetaxel in the placebo group (in the pivotal trial) which		

	may have had an effect on OS. Confounding may also have been caused by the cross-over from placebo to an active treatment (Sipuleucel-T prepared from cryopreserved cells); as stated above, no analyses were performed to adjust for cross-over, and therefore the treatment effect may have been underestimated. 'Lack of consistency' between time to disease progression and overall survival, and possible confounding of the OS results by non-randomised post-progression, post-blinding treatment was also noted in the ERG report for NICE. A possible reason for the lack of association between OS and time to progression was that current clinical metrics of progression assessed in bone are inadequate. Also, immune responses to may require time to develop and the lack of differences in progression could result from such a delayed anti-tumour response. FDA analyses of docetaxel treatment following randomisation did not provide evidence that the survival difference between the two arms was attributable to the post-treatment of docetaxel. The FDA statistician based these analyses on the following assumption which was thought very likely to be true: more patients with good prognosis were in the placebo arm compared to the Sipuleucel-T arm in the subgroup receiving docetaxel. This implies that more patients with poor prognosis were in the placebo arm in the other subgroup in which patients did not receive docetaxel, since overall the two treatments were comparable. In May 2015 the EU marketing authorisation for Provenge was withdrawn at the request of the manufacturer (Dendreon) for commercial reasons.

ReCell Spray-on Skin system *NICE Medical Technologies Evaluation Programme (MTEP) 2014* *Authorisation granted in 2005 under medical devices Directive 93/42/EEC*		
Nature of the Disease		
Indication	Adults or children treated in burns units or centres for: 1) partial thickness burns including scalds caused by hot water where mesh grafting is not required 2) large area burns; full thickness or deep partial thickness burns including where mesh grafting is required	
Orphan status	No	
Is this a rare condition?	No	
What is the natural history of the disease without this treatment/ with current treatment	The treatment of burns can be considered in 2 phases: acute and reconstructive. The acute phase is the initial management of the injury with the intention that burn wound healing will occur with minimal scarring and physical limitation. The reconstructive phase aims to improve the functional or visual impact of scarring, usually by surgical means, and may be done months or years after the initial injury. Full-thickness burns more than 1 cm in diameter need skin grafts because the regenerative components of the skin have been lost. Healing can occur only from the edges of the wound; without a graft the skin contracts, leading to a poor cosmetic outcome and reduced mobility. Deep dermal burns are unlikely to heal within 3 weeks and will therefore often need grafting.	
Nature of the medicine		
How does it work?	ReCell is a stand-alone autologous cell harvesting device that enables a thin split-thickness skin biopsy to be processed to produce a mixed cell population for immediate delivery onto a prepared wound surface.	
Is it claiming to meet an otherwise unmet need?	No	
How is it given?	The ReCell device allows a small, thin split thickness shave biopsy to be physically and enzymatically broken down, yielding a viable suspension of mixed keratinocytes, fibroblasts and melanocytes that can be immediately sprayed or dripped on to the de-epithelialised area. The process is rapid – around 30 minutes – and does not require specialist skills or facilities to carry out. A cell suspension derived from a 1 square cm biopsy is sufficient to treat an area of around 80 sq cm, making it particularly valuable for patients with limited available healthy donor sites.	
Are there any comparator treatments?	a) Partial thickness burns: Biosynthetic dressings, or Standard dressings b) Large area burns: Skin mesh graft alone, or Skin mesh graft plus biosynthetic dressing	
Is there any mention of the intervention evolving over time?	No	
Is there any mention of persistence of the treatment within the patient?	No	
Trial Design		
Trial description	Eleven studies were included in the submission to NICE: 3 RCTs and 8 observational studies. Two of the RCTs were pilot studies with very small samples (13 and 14 each). All but one of the observational studies were also small (range 5 to 40 patients) case series. The two main studies are summarised below.	
	Gravante et al 2007	**Park et al 2013**
	Single-centre RCT	Retrospective cohort study (3 groups)
Trial population (adults/children/all?); any further specifics of disease not covered in 'indication'?	Adults with deep partial thickness burns (<320 cm^2)	Burns treated with skin grafting or replacement All ages

Trial size/ Total trial population?	82	767
Length of follow up?	6 months	NR
Control/comparator used?	RCT: Split thickness skin grafting	RCT: ReCell Spray-On Skin system plus standard skin graft, and standard skin graft alone
How is the control/comparator constructed?	RCT	Both the intervention and the two comparators used historical data. Multiple regression was used although gender and type of burn agent were not included in the model input variables. Burn depth is greater in patients treated with standard skin graft than in patients treated with ReCell alone, although burn depth was controlled for in the multiple regression.
Outcomes		
Response outcome 1	Time to complete epithelialisation	Wound infection
Response outcome 2	Aesthetic and functional quality of the scar	Graft loss
Response outcome 3	Wound infections	
Response outcome 4	Post-operative pain	
Surrogate or intermediate clinical outcome? Yes/No	No	No
Real clinical outcome? Yes/No	Yes	Yes
Adverse events	None reported	None reported
Summary of efficacy evidence		
Overall evidence base provided	The one RCT found ReCell and SSG to be comparable in terms of wound healing time and long-term aesthetics, but Recell was significantly less painful and the mean size of donor site was significantly smaller. These results were reflected in the one large cohort study, which also found no difference in terms of wound infection. The remaining evidence was supportive, indicating a range of patients who can be treated with ReCell. EAC concluded that ReCell may be a clinically suitable alternative to the use of split thickness skin grafts in mid-deep partial thickness burns. There was no clinical evidence examining the use of ReCell in partial thickness burns which are considered not to require skin grafting. There was also no evidence that demonstrated improved outcomes for the use of ReCell plus split thickness skin graft compared with split thickness skin graft alone.	
Estimate of effect HrQoL	Not reported	
Other issues		
Any issues of scale-up for the product?	No	
Is further evidence requested for approval?	NICE concluded that, "The ReCell Spray-On Skin system shows potential to improve healing in acute burns. However, there is insufficient evidence on its use in clinical practice, particularly in relation to which patients might benefit most from its use, to support the case for its routine adoption in the NHS." NICE recommended research to address uncertainties about the claimed patient and system benefits of the ReCell Spray-On Skin system. Clinical outcomes should include time to 95% healing, length of hospital stay, cosmetic appearance of the scar and function of the burned area, compared with standard care.	

Notes	Note within the NICE assessment The claimed benefits of ReCell in the case for adoption presented by the sponsor were:
	• A reduction in the size and depth of the skin graft donor site.
	• Shorter healing time, fewer complications and reduced morbidity at the donor site.
	• Shorter healing time at the recipient site, leading to:
	  • improved aesthetic results for burn wounds, with a reduced likelihood of scarring
	  • reduced likelihood of later readmission to hospital for corrective surgery as a result of improved aesthetic results
	• Repopulation of melanocytes to reduce hypopigmentation and improve skin colour match in healed wounds.
	• Reduced frequency of dressing changes to weekly rather than daily, allowing for a shorter stay in hospital and outpatient management.
	• Reduced need for dressing changes under anaesthetic.
	• A reduction in the need for external technical laboratory support.
	This long list of potential benefits was not supported by robust evidence.

13.2 Appendix 2: Adjustment for bias in non-randomised studies

Adjustment for bias in NRS

Methods developed to adjust effect estimates obtained from NRS for potential biases have taken two broad approaches: to adjust either at the study level, or as part of the process of evidence synthesis. These are discussed separately below.

Adjusting for bias in the evidence synthesis process

The review of the literature on methods to adjust for bias in the evidence synthesis process identified 10 relevant studies.[156-165] These articles included two comprehensive reviews [156, 157]and also individual articles all of which were identified in the review articles. Many of these techniques described in Verde et al.[156] and Doi et al.[157] however have limited applicability to regenerative medicine (i.e. where only limited evidence from a small number of studies is available), as they require significant numbers of studies or/data from RCTs to be applied. A small number of these techniques can, however, be applied where only a single or small number of studies are available. These methods are outlined below.

Adjusting using external data

Welton et al. (2009)[162] present a Bayesian hierarchical model to model bias in RCTs that are at high risk of bias. The authors developed a mixed effects model where treatment effects are considered as fixed and bias effect as random. Estimates of bias in any given meta-analysis are given as a function of prior distribution, which is estimated from published meta-analysis of RCTs, and data from the current meta-analysis. Where a meta-analysis contains no information about the size and magnitude of the bias i.e. where there are only high risk studies the estimate of bias is based on the prior distribution alone. This method allows treatment effect estimates to include information from the high risk studies, accounting for the uncertainty in the magnitude of the bias in any particular meta-analysis. This technique was designed with adjustment of RCTs in mind but is extendable to the adjustment of NRS where by RCTs represent the low risk studies and NRS represent the high risk studies. An appropriate library of meta-analyses combining data from RCTs and NRS would, however, be necessary to apply this technique.

Elicitation

Turner et al.[165] recognizing the practical limitations of basing adjustment on external empirical data, propose an alternative approach in which the direction and magnitude of biases are elicited by reviewers. This method can deal with multiple sources and types of bias including both internal validity bias and external validity bias. In brief, Turner et al.[165] propose that authors design an

idealized study aimed at answering the specific question in mind. This study may not be plausible to carry out and is simply a tool for exploring bias in the completed studies. To identify the potential biases, the completed studies are compared to the idealized study considering a number of potential sources of bias. For each form of bias identified, assessors then elicit the likely magnitude and variance of the bias. These estimates of the magnitude and variance of the potential biases can then be used to adjust treatment effect estimates accounting for both the magnitude and uncertainty of any potential bias identified. External empirical evidence of bias can be included in the analysis rather than relying on eliciting values, but it is assumed that this data will be largely unavailable.

Adjusting for bias at the study level

There are a number of established statistical methods for analysing NRS that attempt to minimise the potential bias from confounding. Each of these methods is briefly described below followed by a brief review of the literature discussing the efficacy of these methods.

Regression analysis

Confounding bias occurs in the context of estimating clinical effectiveness when individual patient characteristics such as age, sex and disease duration that influence efficacy outcomes are also correlated with treatment received. Regression analysis seeks to directly adjust for these potential confounding variables by building a statistical model[166] of the form:

Outcome variable = f (control variables + treatment decision)

Regression models therefore allow the estimation of the treatment effect conditional upon these confounding variables. There are many types of regression model. The choice of any particular model depends on the characteristics of the outcome variable (i.e. continuous or categorical) and on the way it is mathematically related to the explanatory variables. Typically for dichotomous outcomes, a logistic regression model is used. For continuous outcomes a linear regression model is used, and a proportional hazards regression (Cox regression) model is used for time-to-event data.

Theoretically regression models can be used to entirely eliminate bias due to confounding as long as the appropriate parameters are included with a regression equation. However, in reality confounding factors will either be unobserved, preventing their inclusion in the regression model, or a lack of understanding of the disease process will mean that we do not know to include them in the regression model. Where such unobserved confounders are not included in the regression model confounding bias can persist. Regression techniques can be used in conjunction with other methods of adjusting of confounding including propensity scoring and instrumental variables.[166] Regression models also

require a minimum number of participants per additional explanatory variable, with useful rule of thumb of at least 10 observations per explanatory variable.[167] This requirement may limit their application in to regenerative medicine where effectiveness estimates can be based on relatively small studies.

Stratification

Stratification involves the division of participants into subgroups with respect to categorical (or categorized quantitative) prognostic factors, for example classifying age into decades, or weight into quartiles. The intervention effect is then estimated in each stratum and a pooled estimate is calculated across strata. This procedure can be interpreted as a meta-analysis at the level of an individual study. Major limitations are that it is feasible and meaningful only when effects are consistent across strata, and that it can usually be employed only for few variables, as strata increase exponentially in keeping with the number of stratification factors.[166] As such, stratification is method can only minimise rather than completely remove the bias resulting from confounding.

Matching

Matching involves selecting participants with similar values for important prognostic factors to make the control and treatment groups more similar and that any differences between the treatment and control group cannot be as a result of differences in the matched variables. Matching can be carried out both prospectively or retrospectively. Matching prospectively can, however, cause significant recruitment problems. Matching retrospectively can also cause problems as it is not always possible to match individuals In large studies it is often easier to use an unmatched control group and use regression analysis to adjust for what we would have matched on.[168] Matching can however, be useful in small studies where there are insufficient participants to adjust for multiple variables at once.[168] As such matching may be potentially useful techniques for controlling for confounding when using historical controls with the small single arm studies that have typified regenerative medicine clinical evidence. While matching can be used to reduce confounding bias it is unlikely to completely account for all difference due to unobserved confounding.

Instrumental variables analysis

Instrumental variables techniques attempt to approximate the experimental approach by using an instrument variable or variables. A parameter is considered a valid instrument if it meets the following two conditions:

I. The instrument must be correlated with receiving of treatment (or exposure);

II. The instrumental variable must be independent (uncorrelated) with unobserved confounders.

Where a valid instrument exists the instrumental variables approach leads to unbiased estimates equivalent to those from a randomised study. Indeed, randomisation can be thought of as the perfect instrumental variable as it is by definition perfectly associated with treatment allocation and independent of unobserved heterogeneity. The problem with the instrumental variables approach, however, is that identification of a valid instrument is often difficult. Furthermore, while the first requirement of valid instrument is easily tested, the second requirement is essentially untestable and therefore we can never be certain that an instrument is valid. The application of an instrumental variables approach also leads to significant reduction in the power to detect a difference, particularly where the instrumental variable is poorly correlated with treatment allocation. This latter issue may be particular problematic in regenerative medicine where studies are often small with low power to detect differences between alternative treatments.

Propensity Scoring

Propensity scoring rather than being a single method is a suite of methods that consider confounding bias as a form of selection bias where treatment allocation is acknowledged to be non-random and that treatment selection is often influenced by a patient's characteristics.[169] All propensity scoring methods seek to model this process of treatment selection and estimate the propensity to receive treatment based on baseline patient characteristics. Conditional on the propensity score the distribution of base line characteristics will be similar in both the treatment and control groups. Therefore in patients with similar propensity scores patient characteristics will be the same independent of whether treatment was received. The propensity is typically estimated using a logistic regression model, though other methods have been applied. The estimated propensity score can be used to remove the effects of confounding in four different ways.[169] These are described very briefly below:

- Matching – The propensity score is used to match participants in the treatment and control groups who have similar values of the propensity score.
- Stratification – Subjects are ranked on the propensity score and stratified into groups, typically quintiles. Stratum-specific mean differences are then calculated and these differences are effectively meta-analysed to estimate an overall difference in means.

- Inverse probability of treatment weighting (IPTW) – This involves using the propensity score as weight such that an individual participant's weight is equal to the inverse of the probability of receiving the treatment.
- Covariate – The propensity score is added as covariate within a regression equation. The propensity score can be added either with or without additional explanatory variables.

A number of studies have sought to compare propensity scoring methods to ascertain which is the most effective at removing confounding bias[169-173] These studies have shown matching and IPTW to be more effective than stratification or covariate adjustment[170-173] The principal advantage of propensity scoring over other adjustment methods such as regression analysis is that it can be used even with small sample sizes and therefore may me particularly relevant to regenerative medicine. Propensity scoring also has a number of disadvantages . Firstly, propensity scoring only controls for differences in observed variables, and does nothing to remove bias resulting unobserved characteristics. Secondly, including variables that affect whether a treatment is received but not the outcome of interest increases the variance of the estimated treatment effect without a concomitant reduction in bias. This is problematic as sometimes it can be difficult to establish which variables will only impact on which treatment is received. [169]

Effectiveness of adjustment methods

Our review identified a total of nine studies: eight studies[174-181] compared the results of regression analysis, instrumental variables and propensity scoring, and a further paper was identified that discussed the relative merits of the alternative methods of adjustment.[182] Two of these studies were systematic reviews: Shah et al.[178] reviewed comparisons of Propensity scoring versus regression methods. Shah et al.[178]and [181] reviewed comparisons of Propensity scoring versus instrumental variable analyses.

Propensity scoring versus regression methods

Six studies[174, 176-180] compared the different adjustment methods compared propensity scoring with regression methods. The conclusions from these studies were inconsistent. Two studies[178, 180]concluded that estimates obtained from regression methods are similar to those obtained using propensity scoring. Two studies[176, 179], however, also came to the opposite conclusion that estimates obtained from regression and propensity scoring differ significantly. One simulation study [177]comparing the two methods considered propensity scoring to be the superior method while another Cepeda et al[174]found that propensity scoring is superior when the number of events per confounder is low. The disparate results of these studies means conclusions regarding the relative performance is difficult to make, but the conclusion of Kurth et al.[176] makes an important observation that potentially

explains these different results. Kurth et al.[176] notes that each method of adjustment answers subtly different questions as they make different assumptions. This inevitably means that different methods of adjustment will yield different results. Kurth et al.[176] advise that researchers need to consider carefully the population for which an overall treatment estimate is most appropriate.

Regression analysis versus instrumental variable analysis

Only two studies compared regression analysis with instrumental variable methods.[175, 179] Crosby et al.[175] found that results from regression analysis and instrumental variable methods differed somewhat and suggested that instrumental variables are potentially superior. Stukel et al.[179]compared all three methods of adjustment and concluded that instrumental variables may lead to less biased estimates of treatment effects. Although the evidence on instrumental variables is limited it nevertheless suggests it may offer advantages over other methods and may produce the least biased estimates.

Propensity scoring versus instrumental variable analyses

A recently published systematic review found 55 comparisons (37 studies) of propensity scoring with instrumental variable analyses. [181] The review found there to be a slight/fair agreement between the methods [Cohen's kappa coefficient = 0.21 (95% CI 0.00-0.41). In 23 cases (42%) results were nonsignificnat using one method whilst being significant with the other; using instrumental variable methods results were non-significant in most cases (87%). The study authors recommended caution when interpreting the results of these analyses and that further research is needed to clarify the roles for these methods.

In addition to the seven empirical studies identified a discussion paper by Biondi-Zoccai et al.[182]provides a useful overview of the alternative methods of adjustment and their relative methods. Biondi-Zoccai et al.,[182]concluded that there is no clearly superior method noting that "both standard multivariable methods and propensity scores have key limitations, and none is able to take into account unknown confounders." Biondi-Zoccai et al.,[182]however, go on to suggest that propensity scoring methods may have advantages over regression methods where the sample size is small and that while instrumental variables methods are not without their limitations, they are the only methods that allow for unobserved confounding to be adjusted for.

Adjustment methods applied specifically to single-arm trials

The objective of this analysis[183] was to improve the methods to minimize bias in single-arm studies. Four bias factors were suppressed stepwise: attrition bias (by replacing missing values with the baseline value carried forward), bias from natural recovery (by sample restriction to patients with disease duration of 12 months), regression to the mean due to symptom driven self-selection (by

replacing baseline scores with scores three months before enrolment) and bias from adjunctive therapies (by sample restriction to patients not using adjunctive therapies). In the cohort analysed, these four bias factors could together explain a maximum of 37% of the 0- to 6-month improvement of disease score. However, this method has not been widely tested on other cohorts.

Table 44 Methods and results from studies comparing RCTs with NRSs

Study name	Sampling methods	Selection criteria	Number of studies included	Outcomes	Conclusions
Abraham (2006)[32]	A case control and RCT of the effectiveness of laposcopic surgery were carried out and the results compared	NA	1 RCT and 1 NRS. No topic areas were in oncology.	Direction of measured effects Statistical significance of effects Magnitude of measured effects	The results of a surgical historic control trial compared favourably with those of a randomized, controlled trial conducted under similar circumstances in determining the direction of measured effects but tended to yield larger estimates of effect magnitudes.
Algra (2012)[29]	PubMed and the National Library of Medicine were searched.	Papers were eligible for inclusion if they reported results of case–control and cohort studies of use of aspirin or NSAIDs and risk of cancer	12 Oncology areas: 6 RCTS and 195 NRS studies	Subjective assessment of similarity Correlation between estimated effect sizes	Results of methodologically rigorous NRS are consistent with those obtained from randomised controlled trials, but sensitivity is particularly dependent on appropriately detailed recording and analysis of aspirin use.
Benson (2000)[19]	Observational studies published between 1985 and 1998 were searched for in Medline and the Cochrane Database of Systematic Reviews. These were matched to RCTs investigating the same interventions by searching Medline.	NRSs were included if they met the following criteria: • Did not use an experimental design • Included a control group • Treatment was provided by a physician • Assessed the difference between two treatments. No restriction were applied to included RCTs other than that they were relevant to one of the included NRSs.	19 topic areas were included. 53 RCTs and 83 NRSs. 1 topic area was in oncology.	Overlap in confidence intervals. Subjective assessment of similarity of odds ratios.	There was little evidence that effect estimates differed systematically in NRS and RCTs. The authors noted there may be clinically important differences and that their data set was relatively small.

Study name	Sampling methods	Selection criteria	Number of studies included	Outcomes	Conclusions
Beynon (2008)[20]	Randomly selected RCTs from the Cochrane Central Register of Controlled Trials, followed by searches for NRSs addressing the same topic.	RCT or NRS reporting all-cause mortality.	6 topic areas were included. 54 RCTs and 27 NRSs. It was not reported whether topic areas included oncology.	Ratio of odds ratios	Suggest that NRSs overestimated treatment effects by 10% on average, compared with RCTs. However, these are only preliminary results.
Britton (1998)[27]	Searched for studies comparing results from NRS and RCTs in four areas: coronary artery bypass grafting, calcium antagonists, stroke units and malaria vaccines	• The results of the RCT must be compared with a non-randomised study, or the results of several RCTs combined compared with several nonrandomised studies combined. • The intervention must be the same and in similar settings. • The control arms of the studies must receive similar therapy. • There must be comparable outcome measures, preferably valid and reliable.	3 topic areas were included. 29 RCTs and 5 NRSs. No topic areas were in oncology.	Subjective assessment of differences	No evidence from stroke units or calcium antagonists to support using adjustment of observational data to close the gap on RCT data. Differences are probably due to patient characteristics.
Concato (2000)[26]	Searched 5 major journals in Medline between 1991 to 1995.	Meta-analyses of RCTs or NRSs. Excluded studies with historical controls and those that did not report point estimates.	5 topic areas. 55 RCTs and 44 NRSs. 1 topic area was oncology.	Subjective comparison of point estimates and range of estimates obtained from study types.	The results of NRSs are not systematically larger than those obtained from RCTs.
Dahabreh (2012)[21]	Medline search for NRS and studies in acute coronary syndromes. The search was limited to top 8 journals Cardiac and cardiovascular systems and 4 in Medicine, general and internal as defined by Thompson Reuters. RCTs were identified using	Any NRS that used propensity scoring to estimate the treatment efficacy of therapeutic interventions administered to patients with in acute coronary syndromes. RCTs were matched on the basis of interventions, patient populations, and type of mortality outcomes investigated in the NRS.	17 topic areas were included. 63 RCTS and 21 NRS.	Proportion of studies in which ratio of NRS and RCTS treatment effect were lower than 0.70 or greater than 1.43. Number of comparisons in which difference between RCTs and NRSs were statistically significantly different.	For the treatment of ACS, observational studies using propensity scoring methods produce treatment effect estimates that are of more extreme magnitude compared with those from RCTs, although the differences are rarely statistically significant

Study name	Sampling methods	Selection criteria	Number of studies included	Outcomes	Conclusions
	searches of Medline, Cochrane Database of Systematic Reviews and relevant guidelines.			How often the direction of the treatment effect estimated from the NRS and RCT evidence was the same.	
Golder (2011)[30]	Searched multiple databases, including: Cochrane methodology register, DARE and Web of Knowledge for methodological studies relating to the incorporation of adverse effect in to systematic reviews	Any meta-analysis including RCTs and NRS aimed at quantifying relative adverse effects of a health care intervention.	58 meta-analyses in 19 topic areas. 311 RCTS and 222 NRS.	Descriptive summary of overlap in confidence intervals, direction of results and statistical significance of results.	Empirical evidence from this overview indicates that there is no difference on average in the risk estimate of adverse effects of an intervention derived from meta-analyses of RCTs and meta-analyses of observational studies. Some indication that case-control studies gave higher estimates of harm compared to RCTs.
Hartz (2005)[22]	Used data obtained from previous studies.	Included meta-analyses from two previous comparisons of RCTs and NRSs that contained at least 4 observational studies.	10 topic areas. 62 RCTs and 113 NRSs. No topic areas were in oncology.	Number of comparisons in which difference between RCTs and NRSs were statistically significantly different. Comparison of failure rates in intervention and control group. Reporting characteristics and efforts to address confounding in observational studies.	Poor methodological reporting in NRSs prevents conclusions about relative size of effect estimates from being drawn.
Lonjon (2014)[31]	A systematic search of Medline and Pubmed for NRS. Sensitive searches of	Prospective NRS using propensity scoring to evaluate a surgical procedure.	Evidence evaluating 31 clinical questions were included.	Ratio of odds ratios	There was no statistically significant difference in treatment

Study name	Sampling methods	Selection criteria	Number of studies included	Outcomes	Conclusions
	Pubmed were then carried out to identify relevant RCTs. Searches for RCTs were limited to 5 years before and after the oldest and most recent NRS was published.		94 RCTS and 70 NRS.		effect between NRSs with PS analysis and RCTs. Prospective NRSs with suitable and careful PS analysis can be relied upon as evidence when RCTs are not possible.
Ioannidis (2001)[23]	Searched Medline and the Cochrane Library as well as previous studies and personal data.	Meta-analyses of RCTs and NRSs in which the one of the primary outcomes was of binary form and was analysed in the meta-analysis.	45 topic areas. 240 RCTs and 168 NRSs. 5 topics areas in oncology.	Correlation of summary effects obtained from randomised and non-randomised evidence. Proportion of summary effects obtained from NRSs that were larger than those obtained from RCTs. Number of comparisons in which difference between RCTs and NRSs were statistically significant.	Despite good correlation between estimates obtained from RCTs and NRSs, NRSs on average tended to produce larger estimates of effectiveness.
MacLehose (2001)[24]	The Cochrane Library, DARE, and the Science Citation Index were searched. Additionally, references of relevant papers identified were searched and experts were consulted.	Studies reporting estimates of effect from both RCT, quasi experimental and NRSs. This could be from a single study or a pooled analysis from multiple studies	14 topic areas. The number of RCTs and NRSs was not reported. 5 topics areas were in oncology.	Ratio of relative risks. Ratio of risk differences. Comparison of number of events intervention and control group.	Concluded that where quality of NRS is high the disparity between outcomes between RCT and QEO is small. However, the authors caution about generalising findings.
Sacks (1982)[28]	Medline was searched for RCTs and NRS addressing the same topic.	Studies reporting estimates of effect from both RCT, quasi experimental and NRSs.	56 RCTs and 50 NRS It was not reported whether topic areas included oncology.	Magnitude of differences Performance of control group	The data suggest that biases in patient selection may irretrievably weight the outcome of case control studies in favour of new

Study name	Sampling methods	Selection criteria	Number of studies included	Outcomes	Conclusions
					therapies.
Shepard (2006)[25]	A “comprehensive” search of a number of databases was completed. No further information was provided.	Systematic reviews published between 1999 and 2004; evaluated a policy intervention; included both RCTs and NRSs; and, quantitatively synthesised evidence.	16 meta-analyses from one topic area were included. The number of RCTs and NRSs included was not reported. No topic areas were in oncology.	Proportion of reviews in which authors graded the results of RCTs and NRSs "similar", "not similar" or "mixed".	Suggested there may some evidence of differences in results from RCTs and NRSs. However, noted that the lack of consistent criteria to evaluate such differences and lack of exploration into possible other explanations for any differences means it is not possible to draw any strong conclusions.

NRS – non-randomised study; RCT – randomised controlled trial

13.3 Appendix 3: Studies comparing bias adjustment methods

Table 45 Studies comparing bias adjustment methods

Study name	Objective	Methods compared	Summary of findings
Biondi-Zoccai (2011)[182]	Discussion piece comparing relative merits of alternative methods of adjusting for confounding bias in NRS.	Regression analysis, Propensity scoring and instrumental variables	Propensity scoring may have advantages over other methods of adjustment, but all methods have important limitations.
Cepeda (2011)[174]	Simulations study comparing logistic regression with propensity scores in terms of bias, precision, empirical coverage probability, empirical power, and robustness	Propensity scoring and logistic regression	That logistic regression is superior to propensity scoring when the number of events is greater than 8 per confounder
Crosby (2010)[175]	To assess the potential usefulness of instrumental Variables and OLS regression for addressing biases that can confound causal inferences in child care research	Regression analysis and instrumental variables	Note some discrepancies in results obtained using regression analysis and instrumental variables. Suggest that instrumental variables may be superior to regression analysis as a method of accounting for confounding bias.
Kurth (2006)[176]	To assess the utility of different techniques to adjust for confounding.	Propensity scoring and logistic regression	That different methods to control for confounding yielded extremely different treatment effect estimates. This disparity is suggest to be a result of each analyses answering a different

			question implicit or explicit to that methods of adjustment.
[Laborde-Casterot 2015] [181]	Systematic review of studies comparing the performance of propensity scoring with instrumental variable analyses	Propensity scoring with instrumental variable analyses	There was slight/fair agreement between the methods [Cohen's kappa coefficient = 0.21 (95% CI 0.00-0.41). In 42% of cases the two methods produced different results in terms of results significant / non-significant; using instrumental variable methods results were non-significant in 87% of cases.
Martens (2008)[177]	Simulation study comparing the treatment effect estimates from propensity scoring and logistic regression.	Propensity scoring and logistic regression	On average estimates from propensity scoring are closer to true marginal treatment effect than those generated by logistic regression.
Shah (2005)[178]	Systematic review: to determine whether adjusting for confounder bias in observational studies using propensity scores gives different results than using traditional regression modelling.	Propensity scoring and standard regression analysis	Observational studies had similar results whether using traditional regression or propensity scores to adjust for con- founding. Propensity scoring produced modestly more conservative estimates of effect on average.
Stukel (2007)[179]	To compare 4 analytic methods for removing the effects of selection bias in observational studies.	Regression analysis, Propensity scoring (via matching and covariate adjustment) and instrumental variables	Estimates of the observational association of cardiac catheterization with long-term AMI mortality are highly sensitive to analytic method. Compared with standard modelling, instrumental variable

			analysis may produce less biased estimates of treatment effects,
Sturmer (2006)[180]	Top examine the use of propensity scoring methods and whether results obtained using propensity scoring differed substantially from those obtained using standard regression techniques.	Propensity scoring and standard regression analysis	Little evidence that these methods yield substantially different estimates compared with conventional multivariable methods.

13.4 Appendix 4: Studies on surrogate endpoints

Table 46 Review and data extraction of the literature on the use of surrogate measures as clinical endpoints in therapeutic trials

Author(s) & publication	Description/ Aim	Summary / Findings
Health Technology Assessment and Regulation		
Davis *et al* (2012)[54]	The aim of the review is to examine the relationship between progression-free survival (PFS)/Time to progression (TTP) and OS and the evidence available to support surrogate endpoints for OS in advanced cancer.	PFS or TTP are sometimes regarded as valid surrogate outcomes when establishing the clinical benefit of a treatment in the absence of a mature dataset, but an estimate of OS is still needed within the economic analysis. The relationship between surrogate and OS can be used to populate the economic model as an alternative to OS from trial data. Unfortunately when comparing studies the lack of standardised methodology or approach made it difficult to establish a relationship. Some correlation was found but the size of the effect and statistical significance varied considerably. These authors support Taylor and Elston[49] in recommending that any cost-effectiveness analysis based on a surrogate relationship between PFS and OS should be supported with a transparent explanation of how the relationship is quantified in the model and should be accompanied by sensitivity analysis exploring the uncertainty associated with that relationship and a systematic review of papers examining the relationship between PFS and OS in the relevant setting. This would allow decision makers to judge the appropriateness of the model in light of the evidence available in that specific disease area.

Katz, R. (2004)[56]	In this article, the relevant FDA regulatory context is discussed, as well as the epistemological problems related to the interpretation of clinical trials in which un-validated surrogate markers are used as primary outcomes.	From a regulatory standpoint, the use of biomarkers and surrogates are supported when used in the appropriate context and they can be shown to offer a clinical benefit to some of the patients in an 'adequate and well-controlled trial'. In research where the treatment outcome is considered to be clinically important, and where they are few if any available alternative treatment options, accelerated approval on the basis of the drug product having an effect on a surrogate endpoint may be granted by the FDA. The surrogate endpoint is expected to be based on research evidence including epidemiologic, therapeutic and pathophysiologic findings (other than survival or irreversible morbidity). Important points to note with regard to the regulation of surrogate use in research include; i. FDA regulation require the effect on the surrogate to be only "reasonably likely" to predict clinical benefit. ii. The regulation applies to use of an "un-validated" surrogate marker in definitive effectiveness trials (to provide unequivocal evidence of a treatment's tangible benefit to the patient). iii. An un-validated surrogate is described as "reasonably likely" to provide a measure of clinical benefit in circumstances where evidence of a clinical effect is immature. iv. Only for proposed treatments for serious and life-threatening conditions that "….provide meaningful therapeutic benefit……over existing treatments…..," v. Ongoing research must be continued after marketing. If research is not continued or if continued but efforts to validate the surrogate fail the drug must be withdrawn.
Fleming & DeMets (1996)[62]	The most commonly used guidance on the validity of surrogate end points.	For the surrogate to be a reliable outcome measure it must be on the 'causal pathway' from the intervention to the clinical outcome, this is the "setting that provides the greatest potential for the surrogate end point to be valid". Reasons for failure when using surrogate outcomes include, i. The surrogate is not on the causal pathway of the disease process, ii. Of several causal pathways of disease, the intervention affects only the pathway mediated by the surrogate. iii. The surrogate is not in the pathway of the interventions effect or is insensitive to its effect. iv. The intervention has mechanisms of action independent of the disease process
Bucher *et al* (1999)[60]	How to use an article measuring the effect of an intervention on surrogate end points. The JAMA Evidence-Based Medicine Working Group thoughts on the validity of surrogate outcome measures.	For a surrogate to be valid there must be no important effects of that intervention on the outcome of interest that are not mediated through, or captured by, the surrogate. Reliance on a surrogate may be beneficial or harmful. Use of a surrogate may lead to the rapid and appropriate dissemination of new treatments, eg. The FDAs decision to grant approval on new antiretrovirus drugs for the treatment of HIV have subsequently led to effective treatments identified from RCTs. However reliance on surrogate endpoints may lead to morbidity and mortality, e.g. cardiac inotropes may improve short term cardiac hemodynamic function in patients with heart failure but RCTs have found excess mortality with a number of them including flosequinan which had to be withdrawn.

EUnetHTA (2013)[47]	Recommendations for endpoints used in relative effectiveness assessment (REA) of pharmaceuticals. The EUnetHTA summarised their findings into 8 recommendations for endpoints used in relative effectiveness assessment of pharmaceuticals	i. The REA of pharmaceuticals should be based whenever possible on final patient-relevant clinical endpoints (e.g. morbidity, overall mortality). ii. Biomarkers and intermediate endpoints will be considered as surrogate endpoints in REA if they can reliably substitute for a clinical endpoint and predict its clinical benefit. iii. Surrogate endpoints should be adequately validated and must have been demonstrated based on biological plausibility and empirical evidence. iv. Validation of a surrogate is normally undertaken in a specific population and for a specific drug intervention. Demonstration of surrogate validation both within and across drug classes should be thoroughly justified v. The availability of a sufficiently large safety database is particularly important and evidence on safety outcomes should always be reported. vi. The absence of data on clinical endpoints might be acceptable when a clinical endpoint is difficult or impossible to study (very rare or delayed) or target population is too small to obtain meaningful results on relevant clinical endpoints even after very long follow-up (very slowly progressive and/or rare diseases). However, these exceptions need to be carefully argued and agreed in advance. vii. Re-assessment requirements for further data should be clearly defined when a REA has been previously made based on surrogate endpoints for the first assessment. Further methodological research on the use of surrogate outcomes is needed to inform future approaches for the handling of surrogates.

Elston & Taylor (2009)[48]	Paper published prior to the following HTA Report. This paper presents the method used and the findings of the reports.	This paper specifically discusses the role of surrogate outcomes in cost-effectiveness models and is often cited as a key paper. They included the following recommendations: **RECOMMENDATION I** Ideally, the assessment of clinical effectiveness and cost-effectiveness of a health technology should be based on final patient-related outcomes (i.e., mortality, important clinical events, and health-related quality-of-life). To minimize the risk of bias, this evidence should be identified from a systematic review (and meta-analysis) of well-conducted randomized clinical trials. **RECOMMENDATION II** Where this is not possible and there is a requirement to use a surrogate outcome, the following should be undertaken: i. A review of the evidence for the validation of the surrogate/ final outcome relationship. To minimize the risk of bias, such a review should be systematic. ii. The evidence on surrogate validation should be presented according to an explicit hierarchy such as the following: ***Level 1***: evidence demonstrating treatment effects on the surrogate correspond to effects on the patient-related outcome (from clinical trials); ***Level 2***: evidence demonstrating a consistent association between surrogate outcome and final patient-related outcome (from epidemiological/observational studies); ***Level 3***: evidence of biological plausibility of relationship between surrogate and final patient-related outcome (from pathophysiologic studies and/or understanding of the disease process). iii. Consideration for undertaking a CEM analysis based on a surrogate outcome when there is Level 1 or 2 validation evidence. **RECOMMENDATION III** When a CEM analysis based on a surrogate outcome is undertaken: i. Provide a transparent explanation as to how the relationship of the surrogate and final outcome is quantified within the CEM. ii. Explicitly explore and discuss the uncertainty associated with use of the surrogate outcome iii. in the CEM, especially through sensitivity analysis. In accord with recent HTA methodological developments, such uncertainty may be quantified using probabilistic sensitivity analysis. iv. Make specific research recommendations regarding the need for future research on the surrogate/final outcome relationship. In accord with recent HTA methodological developments, the impact of the surrogate outcome on decision uncertainty may be quantified by a value of information analysis. v. Include the term "surrogate outcome" in the report executive summary/abstract to assist bibliographic identification.

Taylor & Elston (2009)[49]	Full HTA report to explore the use of surrogate outcomes in Health Technology Assessment (HTA) and provide a basis for guidance for their future use, validation and reporting.	This report focuses on the role of surrogate outcomes in cost-effectiveness models (CEMs) within UK HTA Programme reports. Reports were selected on the basis that they addressed a treatment effectiveness/efficacy question, that they included a CEM and that the CEM was primarily based on a surrogate outcome. Reports published in the UK HTA Programme monograph series in 2005 and 2006 formed the sampling frame for this study. Only one of the reports undertook a systematic review to specifically seek the evidence base for the association between surrogate and final outcomes. Furthermore, this was the only report to provide level 1 surrogate–final outcome validation evidence, i.e. RCT data showing a strong association between the change in surrogate outcome (BPAR) and the change in final outcome (graft survival) at an individual patient level. The outcome of the study was to make recommendations for the evaluation of surrogate endpoints in HTA. A key output from this work was the design of a schema used to evaluate the cost effectiveness ratio of a surrogate accessed via a meta-analysis HTA.
Key publications		
Aziz *et al* (2015)[51]	This review article focuses on the current evidence from clinical trials in the treatment of mCRPC and concerning the rationale and potential advantages to use prognostic and/or predictive markers in clinical routine and as surrogate endpoints (SEPs) in clinical trials.	….against the background of increasing possibilities of therapy sequencing in clinical practice and in the clinical trial landscape in castration-resistant prostate cancer, does an isolated evaluation of overall survival reliably mirror the benefit attributable to a single compound? …..Suitable parameters serving as surrogates for intermediate and long-term endpoints and reflecting individual benefit, respectively, need to be identified and proven.
Herson (1989)[184]	An introduction to a series of papers which were presented at a meeting in 1989 to address the interest and controversy on surrogate endpoints	"Long completion times are not only a component of overall cost, but also frequently result in the intervention under investigation being rendered obsolete by the time the trial terminates……..The use of surrogate endpoints constitutes an effort to control the cost and completion time for clinical trials."
Ellenberg & Hamilton (1989)[185]	A review of surrogate endpoints in clinical trials with a special focus on cancer	"A surrogate endpoint is usually proposed on the basis of a biologic rationale" "In cancer studies with survival time as the primary endpoint, surrogate endpoints frequently employed are tumour response, time to progression, or time to reappearance of disease"
Prentice (1989)[186]	Discusses the definition and operational criteria for using surrogate endpoints in clinical trials comparing two or more treatments or interventions.	In order that treatment comparisons based on a surrogate response variable have a meaningful implication for the corresponding true endpoint treatment comparison, a rather restrictive criterion is proposed for use of the adjective 'surrogate'. Specifically…..that a surrogate for a true endpoint yields a valid test of the null hypothesis of no association between treatment and true response.

Buyse & Molenberghs (1998)[59]	Statistical model to test prentice frameword for the validation of surrogate measures and specifically in binary datasets with a normal distribution.	In this study the authors examine the relationship between PFS and OS in a set of historical trials in clinical opthalmology. The retrospective trial data found 1,760 patients (57%) had progressed or died at 6 months, and 1,622 (52%) had died at 12 months. The rank correlation coefficient between PFS and OS was equal to 0.82 (95% CI, 0.82 to 0.83). The correlation coefficient between treatment effects on PFS and on OS ranged from 0.99 (95% CI, 0.94 to 1.04) when all trials were considered to 0.74 (95% CI, 0.44 to 1.04) after exclusion of one highly influential trial which exhibited extreme treatment benefits[187]. The authors present data which suggests additional measures are required to validate a and propose two such measures: the first relates the effect of treatment on the true endpoint to that on the surrogate at the population level; the second quantifies the association between the true and the surrogate endpoints after taking treatment into account at the individual level.
Holloway & Dick (2002)[188]	The authors hypothesize that lingering therapeutic uncertainty exists because many of the clinical trial end points have been surrogate outcome measures rather than end points with clear and convincing value to patients	Consequences of using surrogate outcomes that have not been validated include ambiguous evidence and wasted resources as well as patient harm and missed opportunities.
Lessere *et al* (2007)[63]	Review of the literature on biomarkers and surrogates to develop a hierarchical schema that systematically evaluates and ranks the surrogacy status of biomarkers and surrogates; and to obtain feedback from stakeholders.	A new quantitative surrogate validation level of evidence schema (was designed) that evaluates biomarkers along 4 domains: Target, Study Design, Statistical Strength, and Penalties. Scores derived from 3 domains the Target that the marker is being substituted for, the Design of the (best) evidence, and the Statistical strength are additive. Penalties are then applied if there is serious counterevidence. Most stakeholders agreed that this operationalization of the National Institutes of Health definitions of biomarker, surrogate endpoint, and clinical endpoint was useful.
Freedman *et al* (1992)[189]	In this paper the authors expand on the work of Prentice, with respect to the criterion for validation of intermediate variables or surrogate endpoints, by describing and discussing the statistical implementation of this criterion and by using the example of serum cholesterol as an intermediate endpoint for coronary heart disease (CHD).	The authors state a major obstacle in the study of the aetiology of chronic diseases and the development of effective prevention is the long latent period between the initiation of the disease and its diagnosis. Intermediate endpoints or surrogate endpoints are of interest in the study of several diseases as they can usually be observed prior to the clinical appearance of disease. In this paper an attempt is made to clarify the criteria that may be used to validate an intermediate endpoint. The authors found the original general criterion was difficult to test in practice and as such found the validation analysis would require some aspect of statistical modelling.

Zee & Xie (2015)[190]	In studies with surrogate outcomes available for all subjects and true outcomes available for only a subsample, survival analysis methods are needed that incorporate both endpoints in order to assess treatment effects. Our proposed method allows for real-time validation of surrogate outcomes and flexible censoring mechanisms	The proposed method is able to account for the uncertainty of surrogate outcomes using a validation subsample of true outcomes in estimating a binary covariate effect. The proposed estimator can outperform standard semiparametric survival analysis methods and can therefore save on costs of a trial or improve power in detecting treatment effects.
Wilson *et al* [50]	Review paper	Considers the issue of who defines what is a clinically meaningful outcome in cancer treatment; patients, clinicians or regulatory bodies. Also highlights the variation in opinion between these groups.
De Gruttola *et al* (1997)[191]	In this study the authors consider why surrogate endpoints can be unreliable and illustrate the importance of variability in evaluating the reliability of surrogates, with specific focus on HIV/AIDs treatment.	The variety of proposed metrics for evaluating the degree to which this criterion is met are subject to misinterpretation because of the multiplicity of mechanisms by which drugs operate. Without detailed understanding of these mechanisms, metrics of "surrogacy" are not directly interpretable. In order for a marker to be a valid surrogate by the 'Prentice' definition. It must capture all of a treatment's beneficial and harmful effects. Markers that truly capture all of a treatment's effects have never been found. While' 'partial surrogate markers" that capture some of a treatment's effect may provide insight into biologic mechanisms, analyses of the degree of surrogacy must be regarded with caution.
Fleming *et al* (1994)[192]	The applicability surrogate endpoint criteria, is discussed, with emphasis on cancer and AIDS research settings. Auxiliary endpoints are defined as response variables, or covariates, that can strengthen true endpoint analyses such response variables provide some additional information on true endpoint occurrence times for study subjects having censored values for such times.	Circumstances in which surrogate endpoint operational criteria are known to be met, or can be argued on theoretical and empirical bases to be approximately met, are likely to occur very rarely. Many informative intermediate response variables may exist even though individually or collectively they may not satisfy surrogate endpoint criteria. However there is potential for data on pertinent intermediate endpoints to play an auxiliary role in strengthening true endpoint analyses. The two approaches to the use of auxiliary data, respectively, involve an augmented scores method and an augmented likelihood method. The gains will be particularly evident when sufficient follow-up occurs to observe both auxiliary and true endpoints in one set of study subjects, while another set of subjects exists in which the auxiliary endpoint is observed and the true endpoint is censored.

Gotzsch *et al* (1996)[55]	Surrogate outcome measures may speed up clinical research if they can be measured earlier in a study than the primary outcome of interest. In this paper the authors review the justification for the use of surrogates and conclude that reliance on them may be harmful.	Surrogate outcomes can be any measurable event or value related to the disease and true outcome of interest; for example, they can be laboratory values, genetic tests, measures of morbidity, x-ray diagnoses, or absence from work. Surrogate in one trial may be the true outcome in another, depending on the purpose of the study. **Example,** Bone Mineral Content and Fractures. Loss of bone mineral content (BMC) is strongly associated with increased risk of fractures. Logically, one would expect that interventions increasing BMC would also be associated with decreased fracture rates. Fluoride intake is known to increase BMC, and many felt that long-term studies of its effect on fractures were not really necessary. However, such studies were performed, and although the short-term effect on BMC was confirmed, the incidence of fractures was larger in the fluoride group than in the placebo group. Apparently, the new bone was of poor quality. In contrast, intake of the biphosphonate, etidronate, resulted in a slightly decreased fracture rate in addition to the expected positive effect on BMC. BMC is therefore unreliable as a surrogate.
Fleming & De Mets (1996)[162]	This paper provided examples from several disease areas to illustrate how surrogate end points have been misleading about the actual effects that treatments have on the health of patients .	In theory, for a surrogate end point to be an effective substitute for the clinical outcome, effects of the intervention on the surrogate must reliably predict the overall effect on the clinical outcome. In practice, this requirement frequently fails. Surrogate end points can be useful in phase 2 screening trials for identifying whether a new intervention is biologically active and for guiding decisions about whether the intervention is promising enough to justify a large definitive trial with clinically meaningful outcomes. In definitive phase 3 trials, except for rare circumstances in which the validity of the surrogate end point has already been rigorously established, the primary end point should be the true clinical outcome.
Schievink *et al* (2014)[61]	Online questionnaire to inquire for conditions under which surrogate endpoints can be used, the validity of various cardio-renal biomarkers and new approaches for biomarker use.	Questionnaire of various stakeholder groups (regulatory agencies, pharmaceutical industry, academia, relevant public sector organisations) and medical specialties (cardiology or nephrology vs. other). Out of four proposed surrogates (blood pressure (BP), HbA1c, albuminuria, CRP) for cardiovascular outcomes or end-stage renal disease, only use of BP for cardiovascular outcomes was deemed moderately accurate (mean: 3.6, SD: 1.1). Specialists in cardiology or nephrology tended to be more positive about the use of surrogate endpoints
Lerche la Cour *et al* (2010)[57]	To assess if authors of randomised clinical trials convey the fact that they have used surrogate outcomes and discussed their validity.	Of 626 published randomised clinical trials, 109 (17%) used a surrogate as a primary outcome. Of these trials, 62 (57%, 95% confidence interval 47% to 67%) clearly reported that the primary outcome was a surrogate. Only 38 (35%, 26% to 45%) also discussed the validity of the surrogate. Given the shortcomings of surrogates it is surprising that they are used as primary outcomes in about one fifth of published randomised clinical trials. One reason may be the involvement of for profit organisations in many trials. These organisations have an interest in using surrogate outcomes, as it shortens the trial, makes it less costly, and speeds up the implementation of new interventions.

[1] Repeated in table as paper covers many subjects

Bujkiewicz *et al* (2014)[76]	The aim of this study was to illustrate the potential effect of reduced uncertainty around the clinical outcome on the utility when estimating it from a multivariate meta-analysis.	In the areas of highest priority in health care, decisions are required to be made on a short time scale. Therefore, alternative clinical outcomes, including surrogate end points, are increasingly being considered for use in evidence synthesis as part of economic evaluation. Bayesian multivariate meta-analysis was used to synthesize data on correlated outcomes in rheumatoid arthritis. Estimates of Health Assessment Questionnaire were mapped onto the health-related quality-of-life measure EuroQol five-dimensional questionnaire, and the effect was compared with mapping the Health Assessment Questionnaire obtained from the univariate approach. The use of multivariate meta-analysis can lead to reduced uncertainty around the effectiveness parameter. By allowing all the relevant data to be incorporated in estimating clinical effectiveness outcomes, multivariate meta-analysis can improve the estimation of health utilities through mapping methods. While reduced uncertainty may have an effect on decisions based on economic evaluation of new health technologies, the use of short-term surrogate end points can allow for early decisions. More research is needed to determine the circumstances under which uncertainty is reduced.
Ciani *et al* (2015)[68]	This study aimed to to quantify and compare the treatment effects on three surrogate end points, progression-free survival (PFS), time to progression (TTP), and tumor response rate (TR) vs. overall survival (OS) based on a meta-analysis of randomized controlled trials (RCTs) of drug interventions in advanced colorectal cancer (aCRC).	A systematic search was performed of RCTs of pharmacologic therapies in aCRC between 2003 and 2013, 101 RCTs were included. Univariate and multivariate random-effects meta-analyses were used to estimate pooled summary treatment effects. The ratio of hazard ratios (HRs)/odds ratios (ORs) and difference in medians were used to quantify the degree of difference in treatment effects on the surrogate end points and OS. The finding reported a larger treatment effect for the surrogates than for OS. Compared with OS, treatment effects were on average 13% higher when HRs were measured and 3% to 45% higher when ORs were considered; differences in median PFS/TTP were higher than on OS by an average of 0.5 month. None of the end points in this study were found to achieve the level of evidence (ie, mean R2 trialO0.60) that has been set to select high or excellent correlation levels by common surrogate evaluation tools. Previous surrogacy relationships observed between PFS and TTP vs. OS in selected settings may not apply across other classes or lines of therapy.
Ciani *et al* (2013)[58]	The aim was similar to the above trial in that the group worked to quantify and compare the treatment effect and risk of bias of trials reporting biomarkers or intermediate outcomes (surrogate outcomes) versus trials using final patient relevant primary outcomes and a meta-epidemiological methodology.	84 trials using surrogate outcomes and 101 using patient relevant outcomes were considered for analyses. Their risk of bias did not differ. Primary analysis showed trials reporting surrogate endpoints to have larger treatment effects (odds ratio 0.51, 95% confidence interval 0.42 to 0.60) than trials reporting patient relevant outcomes (0.76, 0.70 to 0.82), with an unadjusted ratio of odds ratios of 1.47 (1.07 to 2.01) and adjusted ratio of odds ratios of 1.46 (1.05 to 2.04). This result was consistent across sensitivity and secondary analyses. Trials reporting surrogate primary outcomes are more likely to report larger treatment effects than trials reporting final patient relevant primary outcomes. This finding was not explained by differences in the risk of bias or characteristics of the two groups of trials.

Ciani *et al* (2013)[193]	This case study aimed to illustrate the validation of CCyR and MMR as surrogate outcomes for overall survival in CML and how this evidence was used to inform National Institute for Health and Care Excellence's recommendation on the public funding of these first-line treatments for CML.	This case study illustrates the consideration of surrogate outcome evidence in health technology assessment. Although it is often recommended that the acceptance of surrogate out comes be based on randomized controlled trial data demonstrating an association between the treatment effect on both the surrogate outcome and the final outcome, this case study shows that policymakers may be willing to accept a lower level of evidence (i.e., observational association).
Ciani & Taylor (2013)[53]	Letter to the editor commenting on analytical approaches discussed by Hawkins et al (2012)[194]on the use of surrogates in health technology assessments (HTA) and cost-effectiveness models (CEM).	Presents opinion on the thrtee main issues raised by Hawkins et al. on best practice for the use of surrogate outcomes in HTA and CEMs. i. The need for a new definition of surrogate outcome which not only fits the regulatory and licensing need but also the evaluation of treatment for HTA. ii. To recognise the need for pragmatic, high level evidence, preferably from meta-analyses and regression modelling using both surrogate and final outcomes. iii. The need for systematic evidence which proves a link between the surrogate and final outcome measures allowing a calculation of 'incremental cost-effectiveness ratio' (ICER).
Ciani *et al* (2014)[67]	The authors state it is essential that candidate surrogate endpoints be properly validated. However, believe there is no consensus on statistical methods for such validation and on how the evidence thus derived should be applied by policy makers. This study proposed a method for validation.	A review current statistical approaches to surrogate-endpoint validation based on meta-analysis in various advanced-tumor settings was performed. The authors assessed the suitability of two surrogates (progression-free survival [PFS] and time-to-progression [TTP]) using three current validation frameworks: Elston and Taylor's framework, the German Institute of Quality and Efficiency in Health Care's (IQWiG) framework and the Biomarker-Surrogacy Evaluation Schema (BSES3). Findings suggest, the strength of the association between the two surrogates and OS was generally low. The level of evidence (observation-level versus treatment-level) available varied considerably by cancer type, by evaluation tools and was not always consistent even within one specific cancer type. This study emphasizes the challenges of surrogate-endpoint validation and the importance of building consensus on the development of evaluation frameworks.

13.5 Appendix 5: Licensed treatments for relapsed/refractory B-ALL

Table 47 Licensed treatments for relapsed/refractory B-ALL

Clofarabine **EMA, All Wales Medicine Strategy Group (AWMSG)-Evoltra; FDA - Clolar**	
Nature of the disease and medicine	
Indication(s)?	Relapsed or refractory paediatric ALL patients after receiving at least 2 prior regimens and where there is no other treatment option anticipated to result in a durable response.
How does it work?	Clofarabine is a purine nucleoside anti-metabolite (affects DNA elongation, synthesis, repair).
Is it claiming to meet an otherwise unmet need?	Yes, (indicated in patients where no other durable treatment options exist)
How is it given?	Intravenous infusion for 5 consecutive days every 2 to 6 weeks. Dose for paediatrcs is 52 mg/m^2 over 2 hours.
Are there any comparator treatments?	Not at the time of evaluation (other than palliative care).
Is there any mention of the intervention evolving over time?	No (NOT an RM)
Is there any mention of persistence of the treatment within the patient	No (NOT an RM)
Trial Design	Only a single efficacy trial is available
Trial description	CLO 212: multi-centre single arm phase II trial
Trial population (adults/children/all?); any further specifics of disease not covered in 'indication'?	Paediatric patients: age $\geq$1 tp $\leq$21 years
Trial size/ Total trial population?	61 patients
Length of follow up?	Data cut-off point was 2 years after the start of recruitment.
Control/comparator used	Results for clofarabine were compared with rates expected by expert clinical evaluation. No suitable published studies were available to provide appropriate comparator data.
How is the control/comparator constructed? Source of comparative data	Median survival of 9 to 10 weeks was estimated (using German and Dutch cancer registries).

Outcomes	
Response outcome 1	Overall remission rate (incorporates complete remission (CP) and complete remission without platelet recovery (CRp))
Response outcome 2	Partial response
Response outcome 3	Duration of remission
Response outcome 4	Overall survival
Adverse events	Nausea and vomiting in around two-thirds of patients and febrile neutropenia in around a third. 2 Patients stopped treatment due to an SAE, although 4 deaths were considered to be related to clofarabine.
Surrogate or intermediate clinical outcome?	Remission outcomes
Real clinical outcome?	Overall survival
Summary of evidence	The overall remission rate was 12/61 (20%). 10 of 61 patients (16%) went on to receive HSCT. Median survival (all patients) was 17.7 weeks. In patients who achieved a complete or partial response (18/61 (18%) median overall survival was 66.6 (95% CI: 42089) weeks The effect in terms of remission and facilitating HSCT is considered to be clinically significant and may have a significant impact on long-term treatment outcome. 8/18 responders received an HSCT.
Overall evidence base provided – Trial result summary	
Estimate of HrQoL	No information reported
Product info and registration	
Any issues of scale-up for the product?	No - not an RM
Is further evidence requested for EMA/FDA approval? See final section of post authorisation of product	Specific risk minimisation activities were required. Prescribers were also encouraged to participate in a voluntary adverse event reporting system. In particular monitoring of SIRS was important.
Any additional information provided?	EMA review stated that given the efficacy seen early on in the clinical programme, studies using a placebo comparator were considered to be clinically unethical. Active comparator studies were not appropriate as there were no other recognised therapeutic options available. "The indication is encountered so rarely that the applicant cannot reasonably be expected to provide comprehensive data on clinical efficacy and safety." Marketing authorisation was therefore granted 'under exceptional circumstances'. AWMSG recommended use only if the intended use was as a bridge to HSCT (but should not be used with palliative intent).

Blincyto (blinatumomab) **FDA assessment**	
Nature of the Disease and medicine	
Indication(s)?	Philadelphia chromosome-negative relapsed or refractory B-cell precursor ALL
What is the natural history of the disease with current treatment	Five-year disease-free survival rates in patients with second and third remission are reported to be 27% and 15% respectively. 2% of the children who do not achieve remission have refractory disease; this has a worse prognosis than relapsed disease.
How does it work?	Blinatumomab is a monoclonal antibody (a type of protein) that has been designed to specifically recognise and attach to CD19 proteins and to :the 'T-cell-receptor/CD3 complex', which is responsible for the activation of some cells of the immune system (the body's natural defences) called T cells. By attaching to the cancer cells and the T-cell-receptor/CD3 complex blinatumomab is expected to stimulate the T cells to kill the cancer cells.
Is it claiming to meet an otherwise unmet need?	Clofarabine and marqibo already exist as current treatments although blinatumomab might be a significant alternative because it works in a different way to existing treatments
How is it given?	Intravenous infusion over 4 weeks, with 2 week interval between each treatment cycle
Are there any comparator treatments?	Yes, clofarabine and marqibo have been granted accelerated approval by the FDA for a similar indication prior to blinatumomab
Is there an issue of the intervention evolving over time?	No (not a RM)
Is there an issue of persistence?	No (not a RM)
Trial Design	Trial MT103-211 (with supporting data from MT103-206)
Trial description	Single-arm pivotal Phase II trial (MT103-211)
Trial population (adults/children/all?); any further specifics of disease not covered in 'indication'?	Adults, mean age 39yrs
Trial size/ Total trial population?	189 (MT103-211) + 36 (MT103-206)
Length of follow up?	24 months
Control/comparator used?	Historical controls
How is the control/comparator	Analysis of patient-level data from 694 historical controls: the CR+CRh rate was 24%

constructed?	
Outcomes	**Trial MT103-211 (with supporting data from MT103-206)**
Response outcome 1	Rate of complete remission (CR) + complete remission with partial haematological recovery (CRh)
Response outcome 2	Relapse free survival (RFS)
Response outcome 3	Overall survival (OS)
Response outcome 4	HSCT
Adverse events	Boxed warning for cytokine release syndrome and neurological toxicities (including seizures).
Surrogate or intermediate clinical outcome?	CR+CRh RFS
Real clinical outcome?	OS
Summary of evidence	
Overall evidence base provided	CR+CRh rate was 42% (95% CI 34 to 49). Median RFS 6.7 months (95% CI <0.1 to 16.5).
HrQoL measure	No data
Product info and registration	
Any issues of scale-up for the product?	No (not a RM)
Is further evidence requested for EMA/FDA approval?	A confirmatory phase III RCT, versus standard care chemotherapy in the same population, was ongoing at the time of submission. Randomisation method used will ensure a 2:1 treatment ratio (i.e. more patients will receive blinatumomab than will receive standard care). Overall survival is the primary endpoint. Four post marketing commitments to test the stability of the product once stored.
Any additional information provided?	

Marqibo (vincristine sulphate liposomes injection) FDA assessment	
Nature of the disease and medicine	
Indication(s)?	Adult ALL patients with Philadelphia chromosome negative (Ph-) 2nd or greater relapse or who are refractory to treatment.
How does it work?	Targeted delivery of vincristine is achieved through encapsulating it in nanoparticle liposomes. This allows increased vincristine doses to be achieved without the associated increases in toxicity (dose-limiting neuropathy).
Is it claiming to meet an otherwise unmet need?	Yes (no other durable treatment options existed at the time for this indication)
How is it given?	Intravenously, for one hour every week. Four doses = one course of treatment.
Are there any comparator treatments?	Not at the time of evaluation (other than palliative care).
Is there any mention of the intervention evolving over time?	No
Is there any mention of persistence of the treatment within the patient (keyword search)?	No
Trial Design	Only one trial using the correct dose HBS407 see below). Supporting evidence from a A phase I/II, multi-centre, dose escalation study (VSLI-06) was also submitted.
Trial description	HBS407: multi-centre, single arm, phase II trial (minimax 2 stage design used for sample size)
Trial population (adults/children/all?); any further specifics of disease not covered in 'indication'?	Adults only. All patients had previously been treated with standard vincristine.
Trial size/ Total trial population?	65 patients
Length of follow up?	Up to 5 years (planned)
Control/comparator used? (i.e. what were the results compared to for validation such as improvement compared to no treatment or historical trials using a different treatment)	Data from relevant patients included in a retrospective study were identified and used as a historical control group. Median OS was less than 3 months.

How is the control/comparator constructed? Source of comparative data? Confounding?	
Outcomes	
Response outcome 1	Rate of CR + CR with incomplete blood count recovery (CRi)
Response outcome 2	Duration of CR + CRi
Response outcome 3	OS
Adverse events	Most frequent were constipation (57%) and nausea (52%). Around a third of patients had a neuropathy AE ≥ grade 3.
Surrogate or intermediate clinical outcome?	CR, CRi
Real clinical outcome?	OS
Summary of evidence	
Overall evidence base provided – Trial result summary	10/65 patients (15%) achieved CR or CRi. 5 of the 8 FDA-confirmed CR+CRi patients had duration of response of < 1 month (median duration of response for these 8 patients was 28 days). Five patients who lived for a year or more were considered potential long-term survivors; 2 of the 5 did not respond to Marqibo. 12 patients received a stem cell transplant; 7 patients did not achieve CR or CRi with Marqibo but 6 of these received other chemotherapy and had subsequent SCT.
Estimate/ measure of effect (HrQoL)	No information
Product info and registration	
Any issues of scale-up for the product?	No
Is further evidence requested for EMA/FDA approval? See final section of post authorisation of product	Post-approval confirmatory commitment study: a multi-centre, open-label, RCT of standard vincristine versus Marqibo in adults >60 years with newly diagnosed, untreated PH- ALL. Proposed sample size of 348.
Any additional information provided?	"Accelerated approval" regulations were used. Final vote was 'Yes' 7, 'No' 4, 'Abstain' 2. Members discussed the liposomal formulation of the product and its possible impact on the effectiveness of the drug; they consistently stated that the proposed phase 3 trial was critical in assessing the benefit of Marqibo. Some members indicated that the trial should be completed before approval, while several indicated that accelerated approval may be appropriate, but with the expectation that this approval would be withdrawn if the phase 3 trial failed to confirm clinical benefit. One member stated that the "yes" vote was more an indictment of the lack of

	other options than enthusiasm about Marqibo.

13.6 Appendix 6: Review of previous economic evaluations in ALL

Table 48 Systematic review of previous economic evaluations in ALL

Database: Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) <1946 to Present>
Search Strategy:
1 acute lymphoblastic leukaemia.ti,ab. (4080)
2 acute lymphoblastic leukemia.ti,ab. (19141)
3 Leukemia, Lymphocytic, Chronic, B-Cell/ (12539)
4 1 or 2 or 3 (35395)
5 "ALL R3".ti,ab. (7)
6 ALLR3.ti,ab. (2)
7 "ALL R2".ti,ab. (31)
8 ALLR2.ti,ab. (0)
9 5 or 6 or 7 or 8 (39)
10 4 or 9 (35431)
11 economics/ (26627)
12 exp "costs and cost analysis"/ or Cost Allocation/ or Cost-Benefit Analysis/ or Cost Control/ or Cost of Illness/ or Cost Sharing/ or Health Care Costs/ or Health Expenditures/ (188408)
13 economics, dental/ (1861)
14 exp "economics, hospital"/ or Hospital Charges/ or Hospital Costs/ (20315)
15 economics, medical/ (8619)
16 economics, nursing/ (3916)
17 economics, pharmaceutical/ (2575)
18 (economic$ or cost$ or price or prices or pricing or pharmacoeconomic$).tw. (536282)
19 (expenditure$ not energy).tw. (20070)
20 (value adj1 money).tw. (27)
21 budget$.tw. (20416)
22 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 (665721)
23 ((energy or oxygen) adj cost).ti,ab. (3059)
24 (metabolic adj cost).ti,ab. (925)
25 ((energy or oxygen) adj expenditure).ti,ab. (18354)
26 or/23-25 (21563)
27 22 not 26 (660845)
28 letter.pt. (882177)
29 editorial.pt. (379418)
30 historical article.pt. (317175)
31 28 or 29 or 30 (1563299)
32 27 not 31 (630599)
33 exp animals/ not humans/ (4056152)
34 32 not 33 (586163)
35 10 and 34 (489)

13.7 Appendix 7: Summary of patient characteristics in previously published multivariate prognostic models of ALL

Study	Trial	T or B-cell?	Sample size of interest (for prognostic model)	duration of follow up, median months	sex (%female)	children/adult	CNS disease	Proportion prior transplantation	number of relapses
Fielding et al 2007	UKALL12/ECOG2993	Both	609	49	37%	adults (15-60)	9%	0	1
Ko et al 2010	TACLT2005-002	Both	225	NR	41.30%	children(0-21)	8.30%	NR	1
Nguyen et al 2008	Children's Oncology Group clinical trials (10 trials)	Both	1961	51.7	44%	children	>20.9%	NR	1
Tavernier et al 2007	LALA-94 trial	Both	421	51.6	33%	adults(15-55)	15%	NR	1
Oriol et al 2009	four PETHEMA trials	Both	263	NR	43%	adults (15-70)	<7%	NR	1
Schrappe et al 2012	case based (14 cooperative study groups)	Both	1041	99.6	39%	children(0-18)	6%	NR	induction failure
Thomas et al 1999	MCACC cases	Both	314	NR	39%	adults	15%	NR	1

13.8 Appendix 8: Incidence of relevant population estimate

To estimate budget impact associated with CAR T-cell therapy it is necessary to estimate the incident population eligible for treatment per year. No observed estimates were available due to the small numbers of patients involved and the late stage of treatment, therefore an estimate was constructed based on a three step calculation:

1. Estimate of new ALL diagnosis per year in the UK in the age of interest;
2. Adjustment for B-cell ALL;
3. Adjustment for patients who have relapsed (with no further planned curative chemotherapy or haematopoietic stem cell transplant (HSCT)) or who are refractory to standard chemotherapy

The ONS publishes registrations of newly diagnosed cases of cancer, shown for ALL in England in Table 49.[195] The age of the population of relevance is assumed to be from birth to 30 inclusive, consistent with the definition of children and young adults in Lee et al.,[117] giving an annual incidence estimate of 460.

Table 49: Incidence of new ALL diagnosis in 2013[195]

	Under 1	1-4	5-9	10-14	15-19	20-24	25-29	**Total**
Male	6	87	60	36	33	10	14	**246**
Female	9	102	48	22	18	6	9	**214**
Total	**15**	**189**	**108**	**58**	**51**	**16**	**23**	***460***

Of these ALL incident cases an estimated 80-85% are B-cell ALL,[196] for simplicity we assume 82.5%, giving a B-cell ALL incidence of 379.5. Finally, Fuster[111] estimated that 20% of children experience relapse after current frontline therapy. In addition, Fuster finds that of this population 50% will not respond to salvage therapy or suffer a second relapse, giving a population incidence of relevance of 37.95 per annum in England.

13.9 Appendix 9: Full list of Advisory Group and NICE Panel members

Table 50: Members of the Advisory Group for the project

Andrew Stevens (Chair)	Professor Public Health, University of Birmingham
Natalie Mount	Chief Clinical Officer, Cell Therapy Catapult
Ian McKay	Senior Scientific Officer, Genomics Science and Emerging Therapies, Department of Health
Nick Crabb	Programme Director Scientific Affairs, NICE
Robert Hawkins	Professor of Medical Oncology, University of Manchester
Panos Kefalas	Head of Health Economics and Market Access, Cell Therapy Catapult
Matthew Taylor	Director York Health Economics Consortium, University of York
Philip Newsome	Professor of Experimental Hepatology, University of Birmingham
Chris Mason	Professor of Regenerative Medicine Bioprocessing, UCL
Angela Blake	Head of Health & Value, Pfizer UK
Andrew Webster	Director of the Science and Technology Studies Unit
Paul Catchpole	Director of Value and Access, ABPI
Michael-Hunt	Chief Financial Officer, ReNeuron
Siobhan Connor	Clinical Effectiveness Executive, BUPA
Holger Mueller	SVP, Commercial Operations, Cell Medica
Ahmed Syed	NHS England
Claude Schmitt	Head of Market Access, Rare Diseases, GSK
Angela Crossman	Global Market Access Director, Gene Therapy, GSK
Helen Tayton-Martin	Chief Operating Officer, Adaptimmune
Matthew Durdy	Chief Business Officer , Cell Therapy Catapult

Table 51: NICE Panel meeting participants

The meeting took place on the 29th October 2015 at NICE (Manchester)

Andrew Stevens (Chair of Panel)	Chair of Appraisal Committee C, Professor of Public Health, University of Birmingham
Peter Jackson	Consultant Physician & Hon. Reader in Clinical Pharmacology and Therapeutics Sheffield Teaching Hospitals
Gary McVeigh	Professor of Cardiovascular Medicine, Queen's University Belfast and Consultant Physician, Belfast Health and Social Care Trust
Peter Selby	Consultant Physician, Central Manchester University Hospitals NHS Foundation Trust
Jonathan Michaels	Hon. Professor of Clinical Decision Science, Sheffield
Mark Sculpher	Professor of Health Economics, University of York
Allan Wailoo	Professor of Health Economics & Director of NICE Decision Support Unit University of Sheffield
John Cairns	Professor of Health Economics Public Health and Policy, London School of Hygiene and Tropical Medicine
Norman Waugh	Professor in Public Health, Warwick Medical School
Paul Miller	Director, Payer Evidence, AstraZeneca
Chris O'Regan	Head of Health Technology and Outcomes Research, Merck Sharp & Dohme
Danielle Preedy	Assistant Director , NIHR Evaluation, Trials and Studies Coordinating Centre
David Chandler	Chief Executive, Nominated by Psoriasis and Psoriatic Arthritis Alliance