NICE National Institute for Health and Care Excellence

NICE real-world evidence framework

Corporate document

Published: 23 June 2022

www.nice.org.uk/corporate/ecd9

Contents

Overview

Key messages

- Real-world data can improve our understanding of health and social care delivery, patient health and experiences, and the effects of interventions on patient and system outcomes in routine settings.
- As described in the NICE strategy 2021 to 2026 we want to use real-world data to resolve gaps in knowledge and drive forward access to innovations for patients.
- We developed the real-world evidence framework to help deliver on this ambition. It does this by:
  - identifying when real-world data can be used to reduce uncertainties and improve guidance
  - clearly describing best practices for planning, conducting and reporting real-world evidence studies to improve the quality and transparency of evidence.
- The framework aims to improve the quality of real-world evidence informing our guidance. It does not set minimum acceptable standards for the quality of evidence. Users should refer to relevant NICE manuals for further information on how recommendations are made (see the section on NICE guidance).
- The framework is mainly targeted at those developing evidence to inform NICE guidance. It is also relevant to patients, those collecting data, and reviewers of evidence.
- Table 1 summarises key considerations for conducting real-world evidence studies. The following core principles should be followed to generate high-quality and trusted real-world evidence:
  - ensure data is of good provenance, relevant and of sufficient quality to answer the research question
  - generate evidence in a transparent way and with integrity from study planning through to study conduct and reporting

  - use analytical methods that minimise the risk of bias and characterise uncertainty.

- The framework provides in-depth guidance and tools to support the implementation of these core principles across different uses of real-world evidence. It is structured as follows:

  - the introduction section provides background on real-world data and real-world evidence, discusses its strengths and weaknesses, and summarises current and potential uses within NICE guidance

  - the section on study conduct describes NICE's expectations for planning, conducting and reporting real-world evidence studies, recognising that acceptability of evidence will depend on the type of evidence and other contextual factors

  - the section on assessing data suitability describes the information needed to assess data provenance and its quality and relevance for addressing specific research questions

  - the section on methods for real-world studies of comparative effects provides more specific recommendations for conducting non-randomised studies. These include traditional observational studies as well as clinical trials that use real-world data to form an external control.

- The framework is a living framework that will be updated periodically to reflect user feedback, learnings from implementation including exemplar case studies, developments in real-world evidence methodology, and to extend its scope to include additional guidance on priority topics.

- We encourage companies planning to use real-world data in their submissions to engage early with NICE's Advice service on how to make best use of real-world data as part of their evidence-generation plans.

Table 1

Summary of key considerations in planning, conducting and reporting real-world evidence studies

Stage of evidence generation	Key considerations
Planning	• Clearly define the research question including, as relevant, definitions of population eligibility criteria, interventions, outcomes and the target quantity of estimation • Plan the study in advance and make protocols (including a data analysis plan) publicly available • Choose data that is of good provenance and of sufficient quality and relevance to address the research question (see the section on assessing data suitability) • When planning primary data collection, consider how to implement this collection in a patient-centred manner while minimising the burden on patients and healthcare professionals • Use data in accordance with local law, data governance processes, codes of practice and the requirements of the data controller
Conduct	• Use a study design and statistical methods appropriate to the research question, considering the key risks of bias • Use sensitivity and/or bias analysis to assess the robustness of studies to key risks of bias and uncertain data curation or analytical decisions • Create and implement quality assurance standards and protocols to ensure the integrity and quality of the study

Stage of evidence generation	Key considerations
Reporting	• Report study design and analytical methods in sufficient detail to enable independent researchers to fully understand what was done and why, critically appraise the study and reproduce it • Reporting should also cover: — provenance, quality, and relevance of the data (see the section on assessing data suitability) — data curation — patient attrition from initial data to the final analyses — characteristics of patients (including missing data) and details of follow up overall and across key population groups — results for all planned and conducted analyses (clearly indicating any analyses that were not pre-planned) — assessment of risk of bias and generalisability to the target population in the NHS — limitations of the study and interpretation of what the results mean

Real-world data and its role in NICE guidance

- Real-world data refers to data relating to patient health or experience or care delivery collected outside of highly controlled clinical trials. It can come from many different sources including patient health records, administrative records, patient registries, surveys, observational cohort studies and digital health technologies.
- Real-world data is already widely used to inform NICE guidance to, for example:
  - characterise health conditions, interventions, care pathways and patient outcomes and experiences
  - design, populate and validate economic models (including estimates of resource use, quality of life, event rates, prevalence, incidence and long-term outcomes)

  - develop or validate digital health technologies (for example, digital technologies may use a clinical algorithm developed using real-world data)
  - identify, characterise and address health inequalities
  - understand the safety of medical technologies including medicines, devices and interventional procedures
  - assess the impact of interventions (including tests) on service delivery and decisions about care
  - assess the applicability of clinical trials to patients in the NHS.

- Real-world data that represents the population of interest is NICE's preferred source of evidence for most of these applications. Such data is regularly used for these purposes in NICE guidance, but its use could be more commonplace, especially of routinely collected data.
- Randomised controlled trials are the preferred source of evidence on the effects of interventions. Randomisation ensures that any differences in baseline characteristics between groups are because of chance. Blinding (if applied) prevents knowledge of treatment allocation from influencing behaviours. However, randomised trials are sometimes unavailable or are not directly relevant to decisions about patient care in the NHS.
- Randomised trials may not be available for several reasons, including:
  - randomisation is considered unethical or unfeasible (for instance, for some rare or severe diseases with unmet need)
  - technical challenges make randomisation impractical, which is most common for medical devices and interventional procedures
  - funding is not available for a trial (for example, when the intervention is already used in routine practice).
- Even if randomised evidence is available, it may not be sufficient for decision making in the NHS for several reasons including:
  - the comparator does not reflect standard of care in the NHS
  - relevant population groups are excluded

  - there are major differences in patient behaviours, care pathways or settings that differ from implementation in routine practice
  - follow up is limited
  - unvalidated surrogate outcomes are used
  - learning effects are present
  - trials were of poor quality.

- Non-randomised studies are already widely used to estimate the effects of medical devices and procedures and public health interventions, for which trials are less common. They are becoming more widely used in initial assessments of medicines, as more are granted regulatory approval based on uncontrolled single-arm trials. Finally, the increased focus on the lifecycle evaluation of technologies and lived experiences of patients relies on non-randomised studies after initial approvals. The most common non-randomised studies using real-world data to assess comparative effects are observational cohort studies and single-arm trials with real-world external control.
- Real-world data could be used more routinely to fill evidence gaps and speed up patient access. For this promise to be realised, real-world evidence studies must be performed transparently and with integrity, use fit-for-purpose data, and address the key risks of bias.
- We are communicating our view on best practices for the conduct of real-world evidence studies to ensure they are generated transparently and are of good quality. This is essential to improving trust in real-world evidence studies and their use in decision making.

Introduction to real-world evidence in NICE decision making

Background

Real-world data can improve our understanding of health and social care delivery, patient health and experiences, and the effects of interventions on patient and system outcomes in routine settings.

As described in the NICE strategy 2021 to 2026, we want to use real-world data to resolve gaps in knowledge and drive forward access to innovations for patients.

We developed the real-world evidence framework to help deliver on this ambition. It does this by:

- identifying when real-world data can be used to reduce uncertainties and improve guidance
- clearly describing best practices for planning, conducting and reporting real-world evidence studies to improve the quality and transparency of evidence.

The framework aims to improve the quality of real-world evidence informing our guidance. It does not set minimum acceptable standards for the quality of evidence. Users should refer to relevant NICE manuals for further information on how recommendations are made (see the section on uses of real-world data in NICE guidance).

The framework is mainly targeted at those developing evidence to inform NICE guidance. It is also relevant to patients, those collecting data, and reviewers of evidence.

What is real-world data?

We define real-world data as data relating to patient health or experience or care delivery collected outside the context of a highly controlled clinical trial. Real-world data can be routinely collected during the delivery of health or social care. It can also be collected prospectively, to address 1 or more specific research questions. Most real-world data

sources are observational (or non-interventional), that is, any interventions (or exposures) are not determined by a study protocol. Instead, medical interventions are decided by patients and healthcare professionals. And in public health or social care, interventions may be determined by individual behaviours, environmental exposures or policy makers.

Some interventional studies, such as pragmatic clinical trials, can also produce real-world evidence. Such trials may also make use of real-world data sources to design trials, recruit participants or collect outcome data. For more information, see the UK Medicines and Healthcare products Regulatory Agency's (MHRA) guideline on randomised controlled trials using real-world data to support regulatory decisions.

Table 2 describes common sources of non-interventional real-world data. These include original data collections (such as patient health records) and data curated from original sources (such as the data obtained from retrospective chart reviews). While each type of data source has some general strengths and weaknesses, the value for a given research question will depend on the characteristics of the specific data (for further information, see the section on assessing data suitability). Different sources of real-world data can be combined by linking or pooling to improve data quality and coverage, potentially allowing additional research questions to be answered.

Real-world data can be quantitative or qualitative. Common data types include patient demographics, health behaviours, medical history, clinical outcomes (including patient-reported outcomes), patient or user experiences, resource use, costs, omics, laboratory measurements, imaging, free text, test results and patient-generated data. We consider both national data collections and international data when making recommendations.

Table 2

Common sources of real-world data

Data source	Description	Examples
Electronic health records	Computerised individual patient records. These are typically used to inform the clinical management of patients. These sometimes integrate data from other information systems including laboratory, genomic, and imaging systems.	The Clinical Practice Research Datalink (CPRD) GOLD contains demographic and clinical information on patients enrolled in participating general practices across the UK.
Administrative data	Data collected for administrative purposes by health and social care services.	The Hospital Episode Statistics (HES) Admitted Patient Care dataset contains information on diagnoses and procedures done for all patients admitted to NHS hospitals or NHS-funded treatments in private hospitals. Its primary purpose is to inform the reimbursement of hospitals through payment by results and other operational activities.
Claims data	A type of administrative data on healthcare service use often collected from insurance-based systems.	Centers for Medicare & Medicaid Services data contains data on individuals in receipt of Medicare services derived from reimbursement information or payment of bills. The NHS Business Services Authority provides data on medicines dispensed in primary care in England.

Data source	Description	Examples
Patient registries	Registries are organised systems that collect uniform data (clinical and other) to identify specified outcomes for a population defined by a particular disease, condition or exposure. Registries can serve several purposes including research, clinical care or policy. Registries can include interventional studies.	The Systemic Anti-Cancer Therapy (SACT) dataset contains information on all patients treated with anticancer therapies from NHS England providers. This data is widely used within NICE to provide information on drugs approved for use within the Cancer Drugs Fund. The UK Cystic Fibrosis Registry collects data on consenting people with cystic fibrosis across specialist centres in the UK. The registry data is used to improve the health of people with cystic fibrosis by facilitating research, guiding quality improvement at care centres and monitoring the safety of new drugs.
Patient-generated health data	Data generated directly by patients or their carers including from wearable medical or personal devices, mobile apps, social media, and other internet-based tools. Data can be collected actively (for example, by people entering data on a form) or passively (for example, a smart watch that measures people's activity level).	Pulse oximeters used to monitor people with COVID-19 treated at home to alert need for hospital admission (Greenhalgh et al. 2021). Self-reported data on COVID-19 and long-COVID symptoms from the ZOE app.
Chart reviews	Data extracted retrospectively from a review of patient health records (including paper or electronic records). Chart reviews are widely used in natural history studies. They may allow the extraction of data not reported in routine data sources.	Retrospective chart reviews are especially common in studies of rare diseases to model natural history of disease and treatment pathways (Garbade et al. 2021).

Data source	Description	Examples
Audit and service evaluation	Clinical audits are done to understand how current standards of care measure against best practice or a set standard, and subsequently inform quality improvement. Data can be collected prospectively or retrospectively. Service evaluations are done to define and judge current care.	The Healthcare Quality Improvement Partnership manages national clinical audit programmes such as the Sentinel Stroke National Audit Programme (SSNAP). SSNAP is used to assess the quality of the organisation and delivery of multidisciplinary inpatient stroke health services in England, Wales and Northern Ireland.
Observational cohorts with primary data collection	Traditional prospective studies designed to answer one or more research questions.	The UK Biobank collects data on patient medical histories and genetics. It links to patient records for health outcomes. It was not designed for a specific research question but to enable epidemiological research. EMBRACE-I is a multicentre prospective cohort study to evaluate local tumour control and morbidity in patients undergoing MRI-based image guided adaptive brachytherapy for locally advanced cervical tumours.
Health surveys, interviews and focus groups	Health surveys involve systematic collection of data about health and disease in a human population through surveys. They have various purposes including understanding trends in health in a population or understanding patients' experiences of care. Interviews and focus groups are done to collect qualitative data such as patient perception and experiences.	The Health Survey for England is an annual representative household survey measuring trends in health in England. The 'Living with Lipoedema' 2021 survey by patient charity Lipoedema UK collects patient experience data from individuals with lipoedema. It evaluates experiences of patients having non-cosmetic liposuction or other treatments for lipoedema.

What is real-world evidence?

We define real-world evidence as evidence generated from the analysis of real-world data. It can cover a large array of evidence types including disease epidemiology, health service research or causal estimation (see the section on uses of real-world data in NICE guidance). It can be generated from a large range of study designs and analytical methods (including quantitative and qualitative methods) depending on the research question or use case. A real-world evidence study may use routinely collected data, bespoke data collection, or a combination of the two. We consider single-arm trials that use real-world data sources to create an external control to be real-world evidence studies.

Uses of real-world evidence in NICE guidance

NICE guidance

NICE has several guidance products that use the best available evidence to develop recommendations that guide decisions in health, public health and social care, including:

- guidelines for clinical, social care and public health topics, which offer advisory guidance to health and social care professionals
- evaluations of medical technologies including medicines, diagnostics, medical devices, digital health technologies and interventional procedures.

Guidelines are developed internally by NICE. Technology evaluations are usually informed by company submissions but may also use evidence submitted by manufacturers or other stakeholders or research commissioned from independent academic centres.

The processes and methods for technology evaluations differ across NICE's programmes. The Technology Appraisal Programme evaluates mostly medicines (including highly specialised technologies) but can also include medical devices and diagnostics. The Technology Appraisal and Diagnostic Guidance Programmes both consider the cost effectiveness of medical technologies. The Medical Technologies Evaluation Programme evaluates medical technologies including medical devices, digital health technologies and diagnostics that are expected to be cost-saving or cost-neutral and uses cost-consequence analysis considering patient and system outcomes. The Interventional Procedures Programme evaluates the efficacy and safety of interventional procedures without analysis of cost.

When NICE recommends a treatment 'as an option' through its Technology Appraisal Programme, the NHS must make sure it is available within 3 months (unless otherwise specified) of its date of publication. If a technology is potentially cost effective but there is substantial and resolvable uncertainty about its value, it can be recommended for use in a managed access agreement. After a specified period of collecting real-world data, the technology is reassessed through the Technology Appraisal Programme. Selected devices, diagnostic or digital technologies that are recommended in NICE guidance and are likely to be affordable and produce cost savings within 3 years of adoption can be funded through NHS England's MedTech funding mandate.

Methods and process manuals have been developed for different NICE programmes. Users of this framework should consult these manuals as appropriate:

- Developing NICE guidelines: the manual explains the processes and methods used to develop and update NICE guidelines, the guidance that NICE develops covering topics across clinical care, social care and public health.
- NICE's health technology evaluations manual describes the methods and processes for developing health technology evaluation, including for the Diagnostics Assessment Programme, the Medical Technologies Evaluation Programme, the Highly Specialised Technologies Programme, and the Technology Appraisal Programme.
- NICE's interventional procedures programme manual describes the processes and methods for developing guidance in the Interventional Procedures Programme.

NICE's evidence standards framework for digital health technologies sets out what good levels of evidence for digital health technologies look like. It is aimed at innovators and commissioners of digital health technologies.

Use cases for real-world data

The differences between NICE's guidance programmes lead to variation in the uses and acceptability of real-world evidence.

Real-world data is already used across NICE programmes to generate different types of evidence, especially for questions that are not about the effects of interventions. Examples from previous NICE guidance include:

- characterising health conditions, interventions, care pathways, and patient outcomes

and experiences including natural history: NICE highly specialised technologies guidance on onasemnogene abeparvovec for treating spinal muscular atrophy used multiple sources of real-world data to characterise spinal muscular atrophy

- estimating economic burden: NICE technology appraisal guidance on benralizumab for treating severe eosinophilic asthma reported data from CPRD GOLD linked to HES
- designing, populating and validating economic models. Common types of evidence include:
  - patient starting characteristics: NICE diagnostics guidance on QAngio XA 3D QFR and CAAS vFFR imaging software for assessing coronary stenosis during invasive coronary angiography reported data from the IRIS-FFR registry
  - baseline rates of events: NICE guideline on chronic obstructive pulmonary disease (COPD) in over 16s: diagnosis and management reported data from CPRD GOLD on baseline COPD exacerbation rates by disease severity
  - characterisation of treatment in routine practice: NICE technology appraisal guidance on fenfluramine for treating seizures associated with Dravet syndrome used multiple real-world studies to assess the average dose for comparator treatments in routine practice
  - transition probabilities between health states or disease progression: NICE technology appraisal guidance on patiromer for treating hyperkalaemia used CPRD data to model transition between disease states for people with chronic kidney disease
  - resource use and costs: NICE medical technologies guidance on HeartFlow FFRCT for estimating fractional flow reserve from coronary CT angiography used cost data on coronary revascularisation from NHS reference costs
  - patient-reported outcomes, including quality of life: NICE highly specialised technologies guidance on elosulfase alfa for treating mucopolysaccharidosis type 4a used quality of life data from a survey
  - extrapolation: NICE technology appraisal guidance on atezolizumab with carboplatin and etoposide for untreated extensive-stage small-cell lung cancer used data from the Flatiron Health database, which is derived from US electronic health records.
- measuring patient experience: NICE medical technologies guidance on myCOPD for

managing COPD used patient survey data on the ease of use of the technology

- developing and validating digital health technologies including prognostic models: see the NICE evidence standards framework for digital health technologies for further information
- identifying, characterising and addressing health inequalities: NICE technology appraisal guidance on crizanlizumab for preventing sickle cell crises in sickle cell disease reported evidence from the National Haemoglobinopathy Registry on the health and disproportionate burden of sickle cell disease in certain minority ethnic groups
- estimating test accuracy or reproducibility of test results such as biomarkers: NICE medical technologies guidance on Zio XT for detecting cardiac arrhythmias reported data from a retrospective observational cohort study
- estimating device or procedure failure rates: NICE guideline on joint replacement (primary): hip, knee and shoulder used data from the National Joint Registry on revision rates of knee replacements
- measuring the impact of interventions (including tests) on service delivery and decisions about care: NICE diagnostics guidance on tumour profiling tests to guide adjuvant chemotherapy decisions in early breast cancer reported results from several prospective observational studies.

Real-world data can also be used to assess the applicability of trial results to patients in the NHS or even to estimate intervention effects (for further information, see the section on estimating intervention effects using real-world data).

While real-world evidence is already widely used for many of these types of evidence (Leahy et al. 2020, Makady et al. 2018), its use could be more commonplace. When data is representative of the target population and of sufficient quality it may be the preferred source of data. Background event rates or natural history data from trials may sometimes overestimate or underestimate event rates in the target population because of selective recruitment (Bloudek et al. 2021). In some cases, there may be value in performing studies using routinely collected data rather than relying on published evidence that has lower applicability to the research question.

Estimating intervention effects using real-world

data

Uses and challenges of randomised controlled trials

Randomised controlled trials are the preferred study design for estimating the causal effects of interventions. This is because randomisation ensures that any differences in known and unknown baseline characteristics between groups are because of chance. Blinding (if applied) prevents knowledge of treatment allocation from influencing behaviours, and standardised protocols ensure consistent data collection.

However, randomised controlled trials are not always available or may not be sufficient to address the research question of interest.

Randomised trials may not be available for several reasons, including:

- randomisation is considered unethical, for instance because of high unmet need
- patients are unwilling to be allocated to one of the interventions in the trial
- healthcare professionals are unwilling to randomise patients to an intervention which they consider less effective
- a small number of eligible patients
- financial or technical constraints on studies
- not all treatment combinations (including treatment sequences) can be directly assessed.

Randomised controlled trials may be especially difficult to do for rare diseases, innovative and complex technologies, or in certain populations.

Similarly, high-quality randomised controlled trials can be challenging for medical devices and interventional procedures because of the difficulty of blinding, the importance of learning effects, changes to standard of care making the choice of comparator challenging, changes to the characteristics of the technology over time that may impact on performance, and limited research capacity or access to funding (Bernard et al. 2014).

Even if trials are available, they may not be directly applicable to the research question or to routine care in the NHS because of:

- use of comparators that do not represent the standard of care in the NHS (including placebo control)
- use of unvalidated surrogate outcomes
- limited follow up
- exclusion of eligible population groups (for example, individuals with comorbidities, pregnant women, and children)
- differences in populations, care pathways, or settings that impact on the transferability of results to the target population in the NHS
- differences in patient's use of a technology
- clinical support that differs from routine practice
- learning effects (that is, the effect of an intervention changes over time as users become more experienced)
- methods used to address post-randomisation events such as treatment switching, loss to follow up or missing data.

Some of these challenges, such as the use of comparators that do not represent the standard of care in the NHS, can potentially be addressed through other approaches such as network meta-analysis under certain assumptions about the comparability of the trials. See the NICE Decision Support Unit report on sources and synthesis of evidence for further information.

Real-world evidence can also be generated from randomised controlled trials that use real-world data in their design or for measuring outcomes, such as pragmatic clinical trials. Such trials may provide substantial value in combining the internal validity from randomisation with the greater generalisability of data from routine practice. The UK MHRA has published guidance on producing real-world evidence from randomised controlled trials.

Real-world evidence

Real-world data can be used to contextualise randomised trials, to estimate effects of interventions in the absence of trials, or to complement trials to answer a broader range of questions about the impacts of interventions in routine settings.

Contextualisation

Contextualisation involves assessing whether the results from trials will translate well to the target population in the NHS. While this is an important use of real-world data across NICE programmes, NICE may require the collection of further data through managed access arrangements for medicines that are potentially cost effective and if uncertainties can be addressed through further data collection. This data is often used to understand the relevance of trials to the NHS.

Real-world data has been used in NICE guidance to contextualise clinical trials including for:

- differences in eligible population in the NHS, treatment pathways, care settings and outcomes: NICE technology appraisal guidance on pegcetacoplan for treating paroxysmal nocturnal haemoglobinuria used UK registry data to show that urinary haemoglobin levels in UK practice were in line with the eligibility threshold for the randomised controlled trial
- modelling the relationship between surrogate outcomes and final outcomes (including patient-reported outcomes): NICE highly specialised technologies guidance on lumasiran for treating primary hyperoxaluria type 1 used a registry-based study to model the relationship between plasma oxalate, a surrogate outcome, and kidney function
- measuring the use of, and adherence to, interventions: NICE medical technologies guidance on Sleepio to treat insomnia and insomnia symptoms used data on usage collected from the app or website
- assessing the appropriateness of assumptions about long-term outcomes or treatment effects beyond trial periods: NICE technology appraisal guidance on nintedanib for treating progressive fibrosing interstitial lung diseases used registry data to validate extrapolations of long-term outcomes.

NICE technology appraisal guidance on osimertinib for treating EGFR T790M mutation-positive advanced non-small-cell lung cancer used data from the Systemic Anti-Cancer Therapy (SACT) dataset to assess the relevance of results from the AURA3 trial to NHS patients. In particular, SACT data was used to compare:

- overall survival

- differences in patient characteristics including age, ethnicity, performance status and treatment history.

Estimation

Effects can be estimated for a range of different outcomes, including:

- patient outcomes – clinical outcomes, biomarkers, patient-reported outcomes, behaviour change, user satisfaction and engagement
- system outcomes – resource use, costs and processes of care.

Real-world data can be used to better understand the effects of an intervention over its life cycle. The potential uses of real-world data for estimating effects of interventions depend on the stage in their life cycle.

For new interventions (for example, those with recent marketing authorisation in the UK), there will be limited real-world data on their use and outcomes in the NHS. The uses of real-world data include:

- creating a comparator arm (that is, external control) to estimate effects against a single-arm trial or to add to controls from a randomised controlled trial: NICE highly specialised technologies guidance on metreleptin for treating lipodystrophy used a natural history study to form an external control to a single-arm trial
- using data from early access to medicines schemes: NICE technology appraisal guidance on berotralstat for preventing recurrent attacks of hereditary angioedema included early access to medicines scheme data to reduce uncertainty around long-term outcomes
- estimating comparative effects in other countries in which the technology was available earlier than in the UK (Jonsson et al. 2021)
- predicting outcomes and treatment effects in routine settings, for example, by reweighting results from trials to reflect characteristics of all eligible patients: NICE technology appraisal guidance on pembrolizumab with carboplatin and paclitaxel for untreated metastatic squamous non-small-cell lung cancer used prescribing data from the Cancer Drugs Fund to estimate outcomes weighted by subgroup prevalence.

Once medical technologies are used routinely or in pilot projects, the opportunities for

real-world data are greater and include:

- estimating effects of interventions in routine settings (see NICE medical technologies guidance on DyeVert Systems for reducing the risk of acute kidney injury in coronary and peripheral angiography)
- providing head-to-head comparisons with preferred comparators: NICE technology appraisal guidance on mogamulizumab for previously treated mycosis fungoides and Sezary syndrome used HES data to provide a UK-specific standard-of-care comparator arm to the intervention arm of a randomised controlled trial
- estimating effects in populations excluded from, or under-represented in, the available randomised controlled trials, or extrapolating results from trials: NICE technology appraisal guidance on casirivimab plus imdevimab, nirmatrelvir plus ritonavir, sotrovimab and tocilizumab for treating COVID-19 used OpenSAFELY electronic health records data to support outcomes observed in trial data, and included high-risk populations excluded from trial data
- exploring heterogeneity in intervention effects: NICE technology appraisal guidance on pembrolizumab for treating relapsed or refractory classical Hodgkin lymphoma after stem cell transplant or at least 2 previous therapies used SACT data to model overall survival among those without previous stem-cell transplant
- estimating effects on final outcomes of interest (rather than surrogate outcomes) and over longer time periods
- estimating effects for combination therapies (including sequences) or decision strategies not examined in randomised controlled trials (Fu et al. 2021)
- incorporating into evidence synthesis, for example, informing priors, increasing power or filling evidence gaps in a network meta-analysis (NICE Decision Support Unit report on sources and synthesis of evidence, Sarri et al. 2020).

The validity of real-world evidence for estimating intervention effects

A growing body of literature aims to understand the internal validity of real-world evidence (or, more generally, non-randomised studies) in comparison with randomised controlled trials. This includes meta-epidemiological studies, which compare results from studies of different designs addressing the same question (Woolacott et al. 2017), individual case

studies (Dickerman et al. 2020) and systematic replication studies such as RCT Duplicate (Franklin et al. 2020).

These studies have demonstrated that high-quality non-randomised studies can produce valid estimates of relative treatment effects in many, but certainly not all, situations. There are some common design principles that improve the likelihood of valid estimates including:

- the use of active comparators (alternative interventions for the same or similar indication, usually of the same modality) and
- comparing new users (or initiators) of interventions rather than those who have been using an intervention for some time (prevalent users).

Validity may also depend on other factors including the characteristics of the disease, type of outcome (objective clinical outcomes are preferred), the treatment landscape, and data content and quality.

Challenges in generating real-world evidence

Real-world data has great potential for improving our understanding of the value of interventions in routine settings. However, there are important challenges that must be addressed to generate robust results and improve trust in the evidence. We describe key challenges below.

Trust in real-world evidence studies

Real-world data is often complex and requires substantial preparation before it can be analysed. Also, for some applications, such as the estimation of comparative effects, the methods of analysis can be advanced. When making use of already collected data, researchers may have access to data before finalising their statistical analysis plans. Data preparation and analytical decisions can have important effects on the resulting estimates.

Therefore, concerns about the integrity and trustworthiness of the resulting evidence (for example, resulting from data dredging or cherry-picking) need to be addressed. Concerns about the legitimate use of data have been highlighted by the retraction of high-profile studies about the effectiveness of repurposed medicines for treating COVID-19 from prominent medical journals.

Trust in real-world evidence studies can be improved by:

- registering the study protocol before implementing the study (see the Real-World Evidence Transparency Initiative)
- reporting checklists or tools (see Enhancing the Quality and Transparency of Health Research [EQUATOR] network)
- requiring author statements to confirm the integrity of data access and study conduct (see learning from a retraction by the editors of the Lancet Group, 2020)
- open publishing of data, code lists and analytical code
- providing access to data through secure data environments and maintaining audit trails (see the Department of Health and Social Care's report on better, broader, safer: using health data for research and analysis).

See guidance on planning, conducting and reporting real-world evidence studies in the section on conduct of quantitative real-world evidence studies to generate real-world evidence.

Data quality and relevance

There are several common challenges with using real-world data. Some types of data are often, though not always, absent from real-world data sources (such as measures of tumour size or functional status). However, methods to extract data elements from unstructured data, such as doctor's notes, are increasingly used.

Other variables may be collected at an insufficiently granular level. For instance, a study may need knowledge of a specific drug or medical device, but the data may include only drug or device class. Similarly, a study may need to distinguish between haemorrhagic and ischaemic strokes while a data source may contain data on all strokes without further detail. Even if relevant items are collected with the needed granularity, the data may be missing or inaccurate, which can cause information bias. In addition, there may be variation in data-recording practices and quality across centres or individuals, and in the quality management processes for different sources of data.

In addition to the availability of data on relevant study elements, the relevance of a given data source to a research question may be affected by several factors. This includes the representativeness of the study sample and similarities in treatment patterns and

healthcare delivery to routine care in the NHS, the timeliness of data, sample size and length of follow up. The key questions are whether the data is sufficient to produce robust estimates relevant to the decision problem and whether results are expected to translate or generalise to the target population in the NHS.

See the section on assessing data suitability for further information.

Risk of bias

Studies using real-world data are at risk of bias from a number of sources, depending on the use case. We describe key risks of bias that threaten validity in individual real-world evidence studies below. Detailed descriptions of risks of bias in non-randomised studies are available, such as the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCEPP) guide on methodological standards in pharmacoepidemiology and chapter 25 in the Cochrane handbook for systematic reviews of interventions.

Selection bias

In non-comparative studies, selection bias can occur if the people studied are not representative of the target population. This might result from non-random sampling of the source population, non-response to a questionnaire, or differences in behaviours and outcomes of those who volunteer to be part of research studies.

In comparative effect studies, selection bias occurs if the selection of participants or follow-up time is related to both the interventions and the outcomes. A lack of representativeness of the target population is not itself necessarily a cause of selection bias in comparative studies. Selection bias in comparative studies is distinct from confounding.

Common causes of selection bias at study entry include:

- including prevalent users of a technology compared with non-users (users who had already experienced the event or not tolerated the intervention would be excluded from analysis)
- excluding a period of follow up in which the outcome cannot occur (known as immortal time bias for survival outcomes)

- selection into the study based on a characteristic (for example, admission to hospital) that is related to the intervention and outcome.

A common cause of selection bias at study exit is loss to follow up. Selection bias can also be caused by excluding participants from analysis, such as those with missing data.

Information bias

Information bias may result from missing or inaccurate data on population eligibility criteria, interventions or exposures, outcomes and covariates (as relevant). These limitations may occur because of low data quality, care patterns or data collection processes. They may also result from misspecification of the follow-up period.

The consequences of these issues depend on factors including the study type, whether limitations vary across intervention groups, whether they are random or systematic (that is, the missing data mechanism), the magnitude of the limitation and in which variables they occur. One common cause of differential misclassification across groups is detection bias. This occurs when the processes of care differ according to intervention status such that outcomes are more likely to identified in 1 group than in another. See the section on measurement error and misclassification for further information.

Confounding

Confounding occurs when there are common causes of the choice of intervention and the outcome. This is expected to be common in healthcare because healthcare professionals and patients make decisions about treatment initiation and continuation based on their expectations of benefits and risks (known as confounding by indication or channelling bias). Confounding bias may be intractable when comparing treatments with different indications and across types of intervention (for example, interventional procedure compared with drug treatment) and for studies of environmental exposures.

Bias may also arise because of inappropriate adjustment for covariates, for example, if a study controls for covariates on the causal pathway (such as blood pressure in the effect of anti-hypertensive medication on stroke), colliders (a variable influenced independently by both the exposure and the outcome), or instruments (defined as a variable that is associated with the exposure but unrelated with the outcome except through the exposure).

External validity bias

External validity refers to how well the findings from the analytical sample apply to the target population of interest. Study findings may be intended to be applied to a target population from which the study sample was drawn ('generalisability'), or to another target population, from which the study sample was not derived ('transportability').

Differences can occur between the study sample and target population for factors that affect outcomes on the scale of estimation (for example, relative versus absolute effects). These may include differences in patient or disease characteristics, healthcare settings, staff experience, treatment types and clinical pathways. Further differences may result from patient exclusions, drop out and data missingness in the analytical sample.

Methods to assess and adjust for some elements of external validity bias (those relating to differences in patient characteristics in studies of comparative treatment effects) are discussed in the section on addressing external validity bias.

Other forms of bias

Reverse causation (or protopathic bias) occurs when the intervention is a result of the outcome or a symptom of the outcome. This is most problematic in conditions with long latency periods such as several cancers. If present, this is a severe form of bias with major implications for internal validity.

Biases may also result from the statistical analysis of data (for example, model misspecification).

When assessing the body of literature on a research question there are further concerns about publication bias because of non-reporting of real-world evidence studies, especially if they show null results (Chan et al. 2014).

Conduct of quantitative real-world evidence studies

Key messages

- Transparent and reproducible generation of real-world evidence is essential to improve trust in the evidence and enable reviewers to critically appraise studies.
- The following principles underpin the conduct of real-world evidence studies:
  - Ensure data is of good provenance, relevant and of sufficient quality to answer the research question.
  - Generate evidence in a transparent way and with integrity from study planning through to study conduct and reporting.
  - Use analytical methods that minimise the risk of bias and characterise uncertainty.
- The required level of evidence may depend on the application and various contextual factors (see the section on considerations for the quality and acceptability of real-world evidence). Users should refer to relevant NICE manuals for further information on how recommendations are made.

Introduction

Principles for evidence generation

This section describes NICE's preferred approaches for planning, conducting and reporting real-world evidence studies.

The following principles underpin the conduct of all real-world evidence studies:

- Ensure data is of good and known provenance, relevant and of sufficient quality to

answer the research question.

- Generate evidence in a transparent way and with integrity from study planning through to study conduct and reporting.
- Use analytical methods that minimise the risk of bias and characterise uncertainty.

The focus here is currently on real-world evidence studies of quantitative data. However, several aspects of planning, conducting and reporting that we describe are also applicable to qualitative studies. For aspects that differ, recognised methods of analysing, synthesising, and presenting qualitative evidence should be applied.

Patients should be consulted throughout all aspects of study planning and conduct.

Considerations for the quality and acceptability of real-world evidence

All studies should aim for the highest level of transparency and rigour. However, the large number and variety of real-world evidence studies that can inform a single piece of guidance means there may be reasonable trade-offs between the extent of analysis and reporting and the context of use, including:

- the contribution of the study to the final recommendation
- the impact of the recommendation on health and system outcomes
- other contextual factors.

The contribution of a particular type of evidence will vary across applications depending on the key drivers of uncertainty (that is, the evidence gap). For instance, in oncology, assumptions around long-term outcomes such as overall survival and the applicability of global trials to the NHS are often key (Morrell et al. 2018). In cost-effectiveness or cost-comparison models, a number of different parameters could be important determinants of cost effectiveness including event incidence, prevalence, natural history of disease, test performance, costs or quality of life.

In general, non-randomised studies of clinical effects will need higher levels of rigour and transparency than simple characterisation studies. Estimates of clinical effectiveness are usually a key driver of recommendations and non-randomised studies can be at risk of bias.

The contextual factors that influence the acceptability of evidence include the level of decision uncertainty, disease prevalence, impact on health inequalities and the possibility of generating high-quality evidence. Users should refer to the relevant NICE manual for further information on how recommendations are made (see the section on NICE guidance).

High-quality real-world evidence may be more difficult to generate in certain contexts. These include for rare diseases, and some medical devices (including digital health technologies), interventional procedures or other complex interventions. Conducting randomised controlled trials may also be challenging in these contexts (see the section on uses and challenges of randomised controlled trials).

Common challenges in the evaluation of medical devices and interventional studies using real-world data include:

- limited integrated national data collections of medical device use and outcomes
- lack of granularity in many routinely collected data sources to identify specific devices (and unique device identifiers) or procedures
- identifying appropriate comparators, changes to technologies over time and learning effects.

These challenges are not universal and there are ongoing improvements to the availability of high-quality data collections for medical devices and procedures including registries and electronic health record systems. When possible the highest quality data should be used.

Common challenges in rare diseases include:

- a lack of systematic identification of the target population
- small sample sizes or the need to combine multiple sources of data with different data models and data collection processes
- a lack of agreed common data elements
- substantial variation in natural history of disease
- complex treatment pathways.

Study planning

Defining the research question

Evidence developers should clearly specify their research question irrespective of the study design. While the specific elements of the research question will vary the following are common to many study designs:

- conceptual definitions of key study variables including, as relevant, population eligibility criteria, interventions or exposures, outcomes (patient or system outcome) and covariates (including confounders and effect modifiers)
- subgroups, including specifying whether the subgroup categories are validated or commonly used in the relevant area of research
- the target quantity that is to be estimated, for example, disease prevalence or average effect of adhering to an intervention on overall survival.

Patient outcomes should reflect how a patient feels, functions or how long a patient lives. This includes clinical outcomes such as survival as well as patient-reported outcomes. Outcomes should be reliable and valid for the context of use. Choice of outcomes may be supported by high-quality core outcome sets such as those listed in the Core Outcome Measures in Effectiveness Trials (COMET) database.

The target quantity to be estimated should address the overall research question of interest. For example, prevalence can reflect the quantity of a population who might need access to services at a point in time. It represents a function of incidence and duration of the condition; this may be useful for public health planning. Incidence captures rates of events across different subgroups or those with different exposures but assumes a constant rate across defined time intervals. Plausibility of the average rate should therefore be considered.

For non-randomised studies of comparative effects, developers should provide clear justification for the study, considering reasons for the absence of randomised evidence, the limitations of existing trials and the ability to produce robust real-world evidence for the research question.

Planning study conduct

Developers should aim to pre-specify as much of the study plan as possible. Protocols should describe the objectives of the study, data identification or collection, data curation, study design and analytical methods for all pre-planned analyses including subgroup and sensitivity analyses. We recognise that the complexity of data curation in many real-world evidence studies means not all analytical decisions can be pre-specified. When decisions will be driven by the data these should be clearly described and planned approaches justified. The HARmonized Protocol Template to Enhance Reproducibility (HARPER) tool provides a protocol structure for supporting transparent and reproducible real-world studies of comparative effects.

Planning studies before conduct improves the quality of studies and can reduce the risk of developers performing multiple analyses and selecting those producing the most favourable results.

Pre-specifying analysis plans is especially important for studies of comparative effects. For such studies, we encourage publishing the study protocol on a publicly accessible platform, with any changes to the protocol registered and justified. We do not recommend a specific platform but options include the ClinicalTrials.gov database, the International Standard Randomised Controlled Trial Number (ISRCTN) registry, the European Union electronic Register of Post-Authorisation Studies (EU-PAS), and the Open Science Framework (OSF). Some of these databases are currently more suited to real-world evidence studies than others.

Further guidance on registration of study protocols is provided by the Real-World Evidence Transparency Initiative. NICE's Advice service provides advice on how technology developers can make best use of real-world data as part of their evidence-generation plans.

When planning the study, developers should consider any equality or diversity issues that should be addressed in design, analysis, or interpretation of the study.

Choosing fit-for-purpose data

Developers should justify the selection of the final data sources, ensuring the data are of good provenance and fit for purpose for the research question (see the section on assessing data suitability).

We encourage developers to identify candidate data sources through a systematic, transparent and reproducible search, including:

- pre-specification of search strategy and defined criteria for selection and prioritisation of datasets
- expert consultation to inform the search strategy and selection criteria and to highlight known suitable datasets
- an online search and systematic literature search, and correspondence with lead authors of relevant publications, when necessary, to gain information on access to and suitability of potential data sources
- a direct search of data sources. In the UK this may be supported by registries of data sources such as the Health Data Research UK Innovation Gateway
- a flow diagram outlining the total number of potential data sources identified and the number excluded and reasons why (including for reasons of poor data suitability and feasibility of access).

This approach can be informed by the considerations outlined in the section on assessing data suitability or by following external guidance (Hall et al. 2012, Gatto et al. 2021).

The efforts made to identify data sources should be proportional to the overall importance of the study. We also recognise that currently, registries of data sources are not always available or may have limited metadata.

Data should be accessed and used in accordance with local law, governance arrangements, codes of practice and requirements of the data controller. In the UK, the Health Research Authority (HRA) provides guidance around research and use of data in accordance with the UK Policy Framework for Health and Social Care Research.

Making early contact with data controllers and data processors is prudent to ensure data are available when needed. Developers should ensure they have appropriate ethical (or other) approval for the research study if needed. Developers should also create a plan for sharing data with independent researchers and NICE collaborating centres, when appropriate.

Data collection

For some use cases, primary data collection may be needed. Examples include:

- a new observational cohort study
- additional data collection to complement an existing data source, for example, adding a quality-of-life questionnaire to a patient registry or performing a subsample validation study
- a health survey.

When planning primary data collection, consider how to implement this collection in a patient-centred manner while minimising the burden on patients and healthcare professionals. Assess the feasibility of additional data collection before proceeding.

Sampling methods reduce the burden of data collection but can introduce selection bias. Methods such as simple random sampling support external validity but tend to be feasible only when the target population is small and homogenous. Alternative sampling techniques are available, for example:

- stratified selection divides the target population into subgroups based on important characteristics, such as prognostic factors or treatment effect modifiers, sampling from each strata to ensure representation of all important subgroups
- balanced sampling for site selection considers important variation across sites in the target population. Recruitment focuses on sufficient representation of sites within each subgroup. Potential sites are ranked, allowing for quick identification of replacements due to non-participation
- purposive sampling selects individuals based on their likelihood of being informative, rather than to generalise findings to a larger population. For example, to investigate heterogeneity across characteristics or settings. This approach is common in qualitative research.

Data collection should follow a predefined protocol and quality assurance processes should be put in place to ensure the integrity and consistency of data collection. This also applies to the extraction of structured information in retrospective chart reviews or when using data science methods to derive structured data elements from already collected data sources.

Data collection should follow best-practice standards for 'Findable, Accessible, Interoperable, and Reusable (FAIR)' data using open data standards (see the UK Health Data Research Alliance's Data Standards White Paper 2021).

Data should be collected, stored, processed and deleted in accordance with the current data protection laws with appropriate transparency information provided and safeguards implemented. Approvals from the HRA or local organisation review and agreement as appropriate should be in place. When appropriate, consent from participants should be provided.

Please refer to Health Research Authority guidance on governance requirements and data protection regulation for research and non-research use of healthcare data.

Study conduct

Choosing study design and analytical methods

Real-world data can be used to generate several types of evidence including disease prevalence or incidence, healthcare utilisation or costs, treatment pathways, and patient characteristics, outcomes, and experiences (see the section on use cases for real-world data). The appropriate study designs and analytical methods used should be relevant to the research question and reflect the characteristics of the data, including:

- the nature and distribution of the outcome variable
- sample size
- the structure of the data including data hierarchies or clustering (for example, patients may be clustered within hospitals or data may be collected on a patient at multiple timepoints)
- heterogeneity in outcomes across population groups
- whether data is cross-sectional or longitudinal.

Diagnostic checks should be used to assess the appropriateness of the selected statistical model, if relevant. The appropriate checks will depend on the purpose of the study and methods used.

Further information on the design and analysis of comparative effect studies is provided in the methods section.

Minimising risk of bias

Threats to internal validity from sources of bias should be identified and addressed through data collection and analysis as appropriate. Key threats to internal validity come from selection, information, confounding and other biases depending on the use case (see the section on risk of bias).

The risk of bias from using a particular data source will be informed by the information considered during data suitability assessment.

More detailed guidance on minimising bias in studies of comparative effects is provided in the methods section.

Assessing robustness of study results

Developers should seek to minimise bias at the study design and analysis stages. However, because of the range of possible biases and the complexity of some real-world data sources and analytical methods, some concerns about residual bias will often remain.

Sensitivity analyses should reflect areas with the greatest concerns about risk of bias, or when data curation or analytical decisions were made despite notable uncertainty. Common considerations include:

- varying operational definitions of key study variables
- differing time windows to define study variables and follow up
- using alternative patient eligibility criteria
- addressing missing data and measurement error
- alternative model specifications
- addressing treatment switching or loss to follow up
- adjusting for non-adherence.

If concerns about residual bias remain high and impact on the ability to make recommendations, developers could consider using quantitative bias analysis. These methods provide quantitative estimates of the impact of bias on study results (Lash et al. 2014). If external data on bias is incorporated, this should be identified in a transparent and systematic way. For parameters of economic models including relative effects, sensitivity analysis may consider the impact of bias on cost effectiveness as well as the parameter value.

Using proportionate quality assurance processes

Quality assurance of data management, analytical code and analysis is essential to ensure the integrity of the study and reduce the risk of coding errors. Quality assurance processes should be proportional to the risks of the study.

For further information on quality assurance please see the Office for National Statistic's Quality Assurance of Code for Analysis and Research and the UK Government's Aqua Book. This may be supported by using validated analytical platforms.

Study reporting

Reporting of studies should be sufficient to enable an independent researcher with access to the data to reproduce the study, interpret the results, and fully understand its strengths and limitations. Several reporting checklists identify key reporting items for:

- observational studies (see the EQUATOR network for reporting checklists by study design, and the Strengthening the Reporting of Observational Studies in Epidemiology [STROBE] guidelines)
- observational studies of routinely collected data (REporting of studies Conducted using Observational Routinely collected Data [RECORD]), and
- studies of comparative effects (the RECORD statement for pharmacoepidemiology [RECORD-PE]; although this tool was initially designed for phamacoepidemiological studies the items are relevant to other comparative studies).

Also, the STaRT-RWE tool has been developed to help the presentation of study data, methods and results across use cases.

Below we describe key issues across data sources, data curation, methods and results

that are especially important to cover in reporting the study.

Reporting on data sources

Sufficient information should be provided to understand the data source, its provenance, and quality and relevance in relation to the research questions. This should be informed by the considerations described in the data suitability assessment.

Developers should provide additional information:

- Ethical (or other) approval for the research study or explain why such approval was not necessary.
- A statement that the data was accessed and used in accordance with approvals and information governance requirements.
- A description of how others can access the data (that is, a data sharing statement; for an example, see the BMJ policy on data sharing).

Reporting on data curation and analysis

Many real-world evidence studies, especially those using routinely collected data, need considerable processing (or curation) before analysis is done. The decisions made in data curation (including linkages, transformations and exclusions) may have substantial effects on study results. Data curation should be well described, such that reviewers can understand what was done and how it may impact on results. This should include any curation performed before the evidence developer accessing the data wherever possible.

For each individual study, developers should provide information on the software used to perform analyses including the version system and any external packages used. Ideally, analytical code should follow best practice in code structure, formatting and comments and be publicly available (for example, through a code repository such as GitHub) or made available on request to enable reproduction. When human abstraction or artificial intelligence tools are used to construct variables from unstructured data, the methods and processes used should be clearly described and their validity documented.

It may not be feasible to provide fully open code in all situations, for instance, when using proprietary software or identifiable personal information. Developers should provide clear information on the methods used and their validity. They should also seek to provide

access to the algorithms necessary to replicate and validate the analyses on request, with necessary intellectual property protections in place.

Trust in the integrity of study conduct can be further improved by providing evidence that the study was done appropriately, for example, by showing an audit trail of the analysis, if this is feasible. This could demonstrate, for instance, that developers prepared analysis and finalised protocols before the relevant results were revealed (MacCoun and Perlmutter 2015).

Reporting on methods

Below we describe key items that should be reported. This information should be presented for all analyses including subgroup and sensitivity analyses. Methods should be consistent with the study protocol, and deviations should be identified and justified.

Study design

Clear operational definitions should be given for all study variables and details of follow up, if relevant. Study variables typically include patient eligibility criteria, interventions or exposures, outcomes and covariates.

For each variable, information should be provided on:

- the operational definition of the variable including code lists and algorithms when possible
    - how code lists or algorithms have been developed and, when possible, validated.
- the time period over which information for each variable is sought, defined in relation to an index date (for example, 12 months before starting treatment)
- the grace period between observations that are assumed to represent continued use of an intervention, if relevant.

For studies of comparative effects, the process by which potential confounders were identified should be described alongside assumptions about the causal relationships between study variables.

The following information on follow up should be described when applicable:

- the start and end of follow up in relation to the index date
- for interventions, assumptions about the minimum time between intervention and outcome occurrence (latency period) and the likely duration of effects (exposure-effect window).

In longitudinal studies, this information can be usefully summarised using a study design diagram (Schneeweiss et al. 2019). The Reproducible Evidence: Practices to Enhance and Achieve Transparency (REPEAT) initiative's project page hosts the paper and design diagram templates.

Statistical methods

The statistical methods used should be clearly described. Information should be sufficient to:

- understand what methods were used and why they were chosen
- demonstrate the validity of modelling assumptions
- understand how the analysis addresses different risks of bias including selection bias, information bias and, if relevant, confounding (also see the section on quality appraisal).

Reporting results

The following information should be presented in all studies:

- flow (or patient attrition) diagrams to report number of patients at each stage of the study from raw data to the final analytical sample with reasons for exclusion
- patient characteristics (including missing data) and details of follow up including event rates (or other distributional information on outcomes). For comparative studies these should be presented across groups or levels of exposure and, if relevant, before and after adjustment
- differences in patient characteristics in the analytical sample and target population.

Results should include central-point estimates, measures of precision and other relevant distributional information if needed. Results should be presented for the main analysis and

all subgroup and sensitivity analyses. It should be clear which of these analyses were pre-specified and which were not. For analyses that use adjustment to deal with confounding, unadjusted results should also be presented.

Ensure that information in figures and tables cannot inadvertently identify patients. The Office for National Statistics has guidance on maintaining confidentiality when disseminating health statistics.

Interpreting the results

Provide information to help interpret what the results mean. Discuss limitations in data sources, study design and analysis.

Communicating real-world evidence studies clearly

Real-world evidence studies can be technically complex. To help readers understand them, studies should be documented clearly by:

- following advice on writing understandable scientific material (see Gopen and Swann 1990, Greene 2013)
- avoiding jargon; if this is not possible, explain terms in plain English
- avoiding abbreviations (see Narod et al. 2016)
- labelling tables, graphs, and other non-text content clearly and explaining how to interpret them.

Assessing data suitability

Key messages

- Transparent reporting of data sources is essential to ensure trust in the data source and understand its fitness for purpose to address the research question
- Data should be of good and known provenance
  - Reporting on data sources should cover the characteristics of the data, data collection, coverage and governance
- Data fitness for purpose can be summarised by the data quality and relevance
  - data quality relates to the completeness and accuracy of key study variables
  - data relevance is determined by the data content, differences in patients, interventions and care settings between the data and the target population in the NHS, and characteristics of the data such as sample size and length of follow up.
- The Data Suitability Assessment Tool (DataSAT) in Appendix 1 may be used to provide consistent and structured information on data suitability.
- There are reasonable trade-offs between different data sources in terms of quality, size, clinical detail and locality.
- The acceptability of a given data source may depend on the application and various contextual factors.

Introduction

Data used to inform NICE guidance should be reported transparently and be of good provenance and fit for purpose in relation to the research question. The primary aims of this section of the framework are to:

- provide clear guidance to evidence developers about expectations for clear and transparent reporting on data and its fitness for purpose
- enable evidence reviewers and committees to understand data trustworthiness and suitability when critically appraising the study or developing recommendations.

This section should be read alongside the section on conduct of quantitative real-world evidence studies.

We do not define minimum standards for data suitability beyond that the data should be used in accordance with national laws and regulations concerning data protection and information governance (see the section on reporting on data sources). The considerations for data suitability are broadly applicable across different types of real-world data and use cases but are largely focused on quantitative studies.

The acceptability of a data source will depend on the use case, and contextual factors (see the section on considerations for the quality and acceptability of real-world evidence studies). We recognise the need for trade-offs between different characteristics of data sources including quality, size, clinical detail and locality. International data may be appropriate for some questions in the absence of sufficient national data or when results are expected to translate well between settings. We also recognise that there may be challenges in identifying or collecting the highest quality evidence in some applications including in rare diseases and for some medical devices and interventional procedures (see the section on challenges in generating real-world evidence).

We do not request a particular format for the overall presentation of this information. However, we have developed the Data Suitability Assessment Tool (DataSAT) to help the consistent and structured presentation of data suitability at the point of assessment. The concepts presented in the tool may also help developers choose between potential data sources and in performing feasibility studies, but this is not its primary purpose. The tool template and example applications are presented in appendix 1.

Data provenance

A full understanding of data provenance is essential to create trust in the use of data and understand its fitness for purpose for a given application. In this section we present data provenance considerations across 4 themes: basic characteristics of the data source, data collection, coverage and governance.

Many real-world evidence studies will combine more than 1 data source, either by data linkage or data pooling. Data linkage is often done to extend the information available on individual patients, for example, by combining data from a prospective observational cohort study with hospital discharge or mortality records, or patient-generated health data. Data pooling is used to extend sample size or coverage of data and is common in studies of rare diseases.

The reporting of data sources should primarily refer to the combined data used for the research study. However, important differences between contributing datasets should be clearly described.

Basic characteristics of data

Information that allows identification of the data sources should be clearly reported. This includes the names of the overall and contributing data sources, versions (if available) and the dates of data extraction.

Common data models are used to standardise the structure and sometimes coding systems of different data sources. If data has been converted to a common data model, the model and its version should be reported and full details of the mapping made available, including any information loss. This information is essential to allow the study to be reproduced.

Common data models can also support the use of federated data networks. These allow individual patient health data to stay under the protection of partnering data holders who will run standardised analyses before results are aggregated across datasets. Reporting of federated data networks should be sufficient to understand the process of recruiting data partners, feasibility assessments, and the common analytical framework used.

While complete and accurate data linkage will improve the quality and value of data, imperfect linkage could exclude patient records or lead to data misclassification. Therefore, when multiple sources of data are linked the following information should be reported:

- who did the linkage (for example, NHS Digital)
- methods of linkage including whether deterministic or probabilistic, and the variables used for linkage

- the performance characteristics of data linkage (see the Government Analysis Function guidance on quality assessment in data linkage).

Data collection

An understanding of a data source requires knowledge of the purpose and methods of data collection.

Information on the original purpose of data collection should include:

- whether the data was routinely collected or collected for a specific research purpose (or a combination)
- the type of data source and primary use, for example:
  - electronic health records for patient care
  - administrative data for reimbursement of providers
  - registry for assessing medical device safety
  - prospective observational cohort study to estimate quality of life after an intervention
  - retrospective chart review to model the natural history of a condition.

Additional information on important data types should cover:

- which types of data were collected, for example, clinical diagnoses, tests, procedures and prescriptions
- how these were coded or recorded, for example, using ICD-10 codes for clinical diagnoses, or free text data on cancer stage or biomarkers
- how data was collected, for example, directly by healthcare professionals in clinical examinations, by remote monitoring or by administrative staff. If data is captured by a digital health technology, the validity of the technology should be reported
- changes to data collection over time, for example:
  - addition of new data elements (for example, a quality-of-life questionnaire)

  - removal of data elements
  - changes to the method of data collection (for instance, a switch to routine monitoring of patient outcomes)
  - changes to coding systems (for example, the switch from Read v2 to SNOMED-CT codes in UK primary care). Information on any mapping between coding systems should be made available
  - software updates to data capture systems including digital health technologies that had substantial impacts on data capture.
- quality assurance processes for data collection that were in place (including training or blinded review)
- transformations performed on the data such as conversion to a common data model or other data standards.

Any differences between data providers in how and what data were collected, and its quality, should be described. This is especially important when data sources are pooled from different systems and across countries.

Data coverage

Providing clear information on data coverage is essential, including the population, care settings, geography and time. Such information has important implications for data relevance that can inform later assessments of data suitability.

Information should be provided on:

- the extent to which the data source captures the target population:
  - if a data source does not include the full target population, the representativeness of the data captured should be noted
  - for studies involving prospective data collection including patient registries, information on patient accrual should be reported.
- the settings in which data collection was based:
  - this should distinguish between care settings (for example, primary care

compared with secondary care), type of providers (for example, specialist medical centres compared with general hospitals) and other factors when relevant

  - if information was collected outside of the health or social care system, this should be described (for instance, remote monitoring of activities of daily living).

- the geographical coverage of the data including countries and regions, if relevant
- the time period of data collection.

Data governance

Information about data governance is important for understanding the maturity of data and its reliability. This should include the following information:

- the name of the data controller
- the funding source for data collection and maintenance
- data documentation including items such as a data dictionary and data model
- details of the quality assurance and data management process including audit.

Data fitness for purpose

The section on data provenance described important characteristics of data sources distinct from the planned study. In this section we focus on the fitness for purpose of data to answer specific research questions considering its quality and relevance. A dataset may be of value for 1 application but not another.

Substantial data curation including data cleaning, exclusions and transformations is needed to prepare original data sources for analysis. Data curation and quality assurance should be reported transparently as described in the section on study reporting.

Data quality

Limitations to data quality include missing data, measurement error, misclassification and incorrect reporting of dates. These issues can apply to all study variables including patient eligibility criteria, outcomes, interventions or exposures, and covariates. They can create

information biases that cause real-world evidence studies to produce biased estimates. Transparent reporting of data quality is essential for reviewers to understand the risk of bias and whether it has been adequately addressed through data analysis or explored through sensitivity analysis. We focus on 2 main aspects of data quality: completeness and accuracy.

Information on completeness and accuracy should be provided for all key study variables. Study variables can be constructed by combining multiple data elements, including both structured data and unstructured data, and may come from different linked data sources. The complexity of these study variables will vary according to the data sources and applications. For instance, in some applications an asthma exacerbation may be identified from a single data field (such as the response to a questionnaire), while in others it may need to be constructed from combinations of diagnostic codes, prescriptions, tests, free text or other data.

As described in the section on study reporting, it is essential that clear and unambiguous definitions are given for each study variable including types of data, code lists, extraction from unstructured data, and time periods, when possible. These operational definitions including code lists should be made available to others and reused, if appropriate. The validity of an existing code list should be reviewed before use. When unstructured data is used, information should be provided on data extraction and the reliability of these methods.

These considerations also apply to data from digital health technologies producing patient-generated data, including patient-reported outcomes and digital biomarkers. Further information on the validity of data generated from the technology and user accessibility should be provided.

To interpret study results, further information is needed on reasons for data missingness and inaccuracy and whether these are random or systematic. For comparative studies, it is important to understand the extent to which missingness or inaccuracy differ across intervention groups. The section on addressing information bias has further information on methods for dealing with missing data, measurement error and misclassification. We have not set minimum thresholds for data completeness or accuracy because the acceptable levels will depend on the application (see the section on considerations for the quality and acceptability of real-world evidence studies).

Completeness

Data completeness refers to the percentage of records without missing data at a given time point. It does not provide information on the accuracy of that data. The percentage is often easily calculated from the data source and should be calculated before excluding relevant data or imputation. For outcomes such as experiencing a myocardial infarction, issues of data missingness should be clearly distinguished from misclassification. For binary variables, the absence of an event (when it has occurred) may be best summarised as a data accuracy issue (misclassification due to false negatives).

Accuracy

Measuring accuracy, or how closely the data resemble reality, depends on the type of variable. Below we describe common metrics of accuracy for different types of variables:

- continuous or count variables (mean error, mean absolute error, mean squared error)
- categorical variables (diagnostic accuracy measures such as sensitivity, specificity, positive predictive value, and negative predictive value; Fox et al. 2022)
- time-to-event variables (difference between actual time of event and recorded time of event).

Gold standard approaches for measuring accuracy of the data include:

- comparison with an established gold standard source (for example, UK Office for National Statistics mortality records)
- medical record review.

These approaches may be taken for a subset of the analytical population or be based on a previous study in the same or similar population and data source.

These gold standard approaches are not always possible or feasible. Other approaches that can show approximate accuracy include:

- comparing different variable definitions, for example, by using additional codes, requiring multiple codes, or combining different data types
- comparing sample distributions with population distributions or previous studies

- exploring plausibility of the data, informed by expert opinion
- checking consistency (agreement in patient status in records across the data sources)
- assessing conformance (whether the recording of data elements is consistent with the data source specifications)
- checking persistence (whether the data are consistent over time).

Transparent reporting of data accuracy for key study variables includes:

- Quantitative information on accuracy, if available, including means and confidence intervals. Additional distributional information may also be valuable.
- Describing the methods and processes used to quantify accuracy including any assumptions made. When this is based on previous studies, the applicability to the present analysis should be discussed, and may consider differences in study variable definitions, populations, data sources, time periods or other relevant considerations.

Data relevance

The second component of data fitness for purpose is data relevancy. Key questions of data relevancy are whether:

- the data provides sufficient information to produce robust and relevant results
- the results are likely to generalise to patients in the NHS.

The assessment of data relevancy should be informed by the information provided in the section on data provenance.

NICE prefers data relating directly to the UK population that reflects current care in the NHS. However, we recognise the potential value of international data if limited information is available for the NHS or if results can be expected to translate well between settings. In some applications there will be a trade-off between using local data and other important characteristics of data including quality, recency, clinical detail, sample size and follow up. International data is likely to be of particular value when an intervention has been available in another country before becoming available in the UK, or in the context of rare diseases. Similar considerations apply to using data from regional or specialist healthcare providers within the NHS.

We describe key aspects of data relevancy below distinguishing between data content, coverage and characteristics.

Data content

There are 3 key considerations for understanding whether the data content is sufficient for a research question:

- Does the data source contain sufficient data elements to enable appropriate definitions of population eligibility criteria, outcomes, interventions and covariates, as relevant?
- Are the data elements collected with sufficient granularity (or detail)?
- Are measurements taken at relevant time points?

To help understand whether data elements are sufficient, it is useful to first define the target concept and judge the extent to which this can be proxied using real-world data. The implications of insufficient data will vary depending on the study variable and use case. Key endpoints necessary to answer the research questions should be available and should be sufficiently objective and detailed to support an evaluation. Insufficient information to define the population, interventions or outcomes appropriately will limit the relevance of the research findings. Insufficient information on confounders will limit the ability to produce valid findings.

The needed granularity of data will vary across research questions. For example, when considering the effect of knee replacement on quality of life we may be interested in the effect compared with physiotherapy alone, total versus partial knee replacement, or of different implanted devices. Similarly, any stroke may be appropriate as an outcome for some research questions, while others will need haemorrhagic and ischaemic strokes to be separated.

Finally, we may be interested in the effect of knee replacement on quality of life at 1 year after the procedure. In routinely collected data, the recording of such information does not follow a strict protocol with measurements missing or taken at irregular time points.

Data coverage

The generalisability of research findings to patients in the NHS will depend on several

factors, including:

- the similarity in patient characteristics between the analytical sample and target population
- the similarity in care pathways and treatment settings
- changes in care pathways (including diagnostic tests) and outcomes over time.

The similarity of the analytical sample to the target population is especially important in descriptive studies, such as those estimating disease prevalence. In comparative studies this may be less important if the intervention effects are expected to transfer across patients with different characteristics, and the emphasis should be on ensuring internal validity. If there is substantial heterogeneity in treatment effects across subgroups, similarity in patient characteristics becomes more important. Effect estimates on the relative scale usually transfer better across subgroups than estimates on absolute scales (Roberts and Prieto-Merino 2014). In other applications, such as prognostic modelling, non-representative sampling may be preferred to ensure adequate representation of important patient subgroups.

Consideration needs to be given to how any differences in the treatment pathways or care settings seen in the analytical sample and the NHS may impact on the relevance of results. This is especially important when using international data. Even within the NHS, the data may relate to specific regions that are not representative of the country or focus on specialist providers rather than all providers. Finally, changes to care pathways including diagnostic tests as well as background trends in outcomes (such as mortality) may limit the value of historical data even from the NHS. These issues need to be carefully considered and reported when discussing the relevance of data for use in NICE guidance.

Data characteristics

The final category of data relevancy concerns the size of the analytical sample and the length (and distribution) of follow up. The sample size should be large enough to produce robust estimates. However, we recognise that sample size will always be limited in some contexts. The follow up should be long enough for the outcomes of interest to have occurred or accrued (for outcomes such as healthcare costs). The amount of data available before the start of follow up may also be important to provide information on confounders and identify new users of an intervention. Using data sources with a lower time lag between data collection and availability for research may allow for longer follow

up to be available for analyses.

Methods for real-world studies of comparative effects

Key messages

- Non-randomised studies can be used to provide evidence on comparative effects in the absence of randomised controlled trials or to complement trial evidence to answer a broader range of questions about the effects of interventions in routine settings.
- The recommendations presented here focus predominantly on cohort studies including those using real-world data to form external control arms.
- Study design
  - Design studies to emulate the preferred randomised controlled trial (target trial approach).
  - Avoid time-related biases due to differences between patient eligibility criteria being met, treatment assignment, and start of follow up.
  - For studies using external control, select and curate data to minimise differences between data sources including availability and operational definitions of key study variables, data collection processes, patient characteristics, treatment settings, care pathways, and time periods, and consider the implications for study quality and relevance.
- Analysis
  - Identify potential confounders (including time-varying confounders) using a systematic approach and clearly articulate causal assumptions.
  - Use a statistical method that addresses confounding considering observed and unobserved confounders.
  - Consider the impact of bias from informative censoring, missing data, and measurement error and address appropriately, if needed.
  - Assess the external validity of findings to the target population and consider if adjustment methods are suitable or needed.
  - Use sensitivity and bias analysis to assess the robustness of results to main

risks of bias and uncertain data curation and analysis decisions.

- Reporting
  - Justify the need for non-randomised evidence.
  - Provide a study protocol and statistical analysis plan before performing final analyses.
  - Report studies in sufficient detail to enable independent researchers to reproduce the study and understand what was done and why.
  - Assess the risk of bias and relevance of the study to the research question.
- The acceptable quality of evidence may depend on the application and various contextual factors (see the section on considerations for the quality and acceptability of real-world evidence).

Introduction

We previously outlined principles for the robust and transparent conduct of quantitative real-world evidence studies across different use cases. In this section we provide more detailed recommendations for the conduct of studies of comparative effects using real-world data. This includes traditional observational studies based on primary or secondary data collection and trials in which real-world data is used to form an external control. We do not provide specific considerations for purely interventional studies (whether randomised or not) or external control studies using only interventional data. We focus here on quantitative studies but recognise that qualitative evidence can play an important role in improving our understanding of the value of interventions.

Randomised controlled trials are the preferred study design for estimating comparative effects. Non-randomised evidence may add value if randomised controlled trials are absent, not directly relevant to the research question or of poor quality (see the section on uses and challenges of randomised controlled trials). They can also complement trial evidence to answer a broader range of questions (see the section on estimating intervention effects using real-world data).

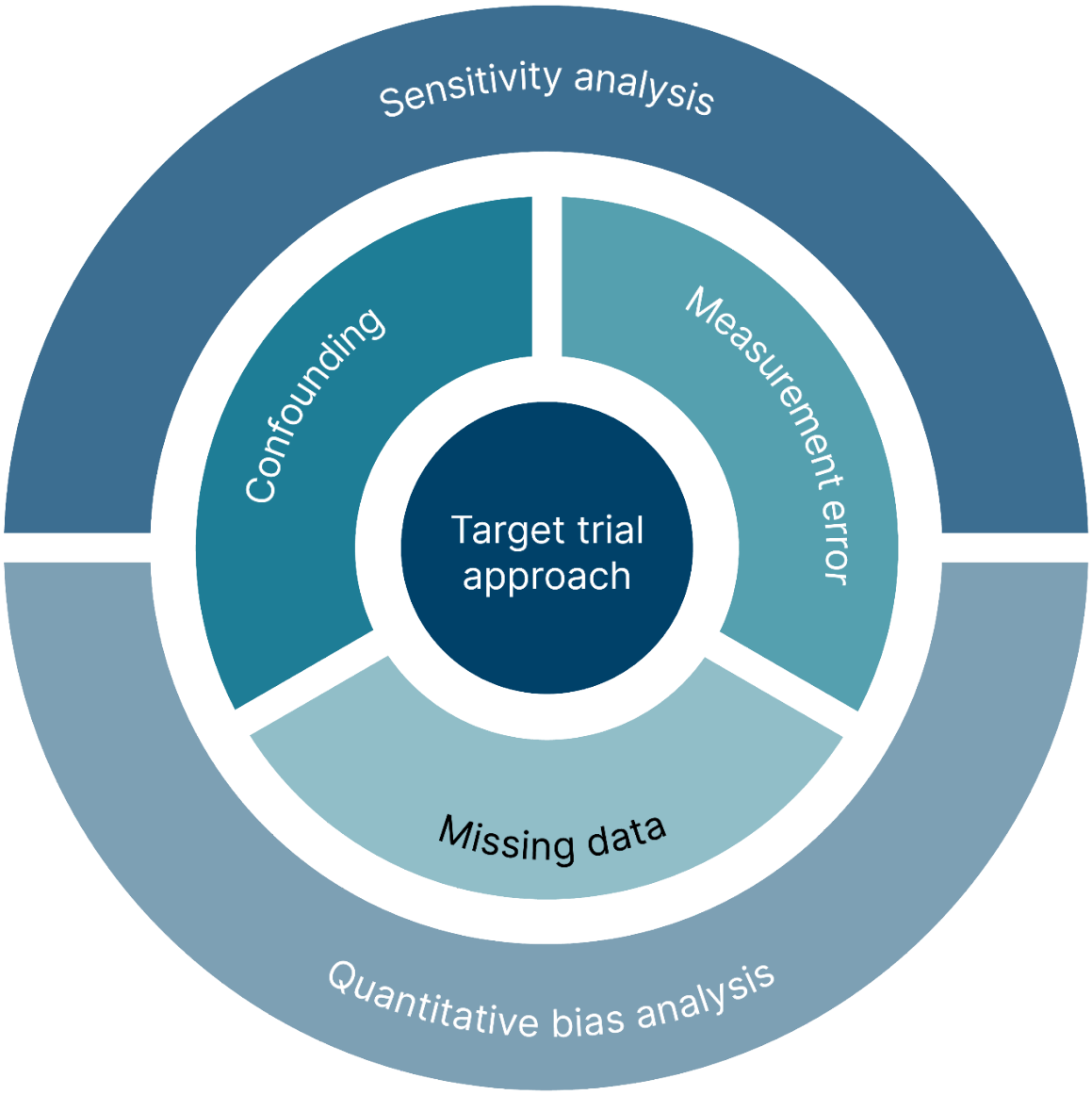
If real-world evidence on comparative effects may improve the evidence base, it is essential that studies are done using robust and transparent methods. We recommend designing real-world evidence studies to emulate the randomised trial that would ideally have been done (see the section on study design), using appropriate statistical methods to address confounding and informational biases (see the section on analysis), and assessing the robustness of results using sensitivity and bias analysis (see the section on assessing robustness). This approach is summarised in figure 1 a visual summary of key considerations for planning and reporting cohort studies using real-world data.

The recommendations provided here are intended to improve the quality of real-world studies of comparative effects, both in terms of methodological quality and validity, and the transparency of study conduct. They were derived from best-practice guidance from the published literature, international research consortia, and international regulatory and payer bodies, and will be updated regularly in line with developing methodologies. They build on NICE Decision Support Unit's technical support document 17, which presents statistical methods for analysing observational data.

We recognise that not all studies will be able to meet all recommendations in full. The ability to perform studies of the highest quality will depend on the availability of suitable data (see the section on assessing data suitability) and characteristics of the condition and intervention. Simpler methods may be appropriate for other applications including assessing non-health outcomes like user experience or some system outcomes. In addition, the acceptability and contribution of specific studies to decisions will depend on the application as well as several contextual factors (see the section on considerations for the quality and acceptability of real-world evidence studies).

Figure 1

Visual summary of key considerations for planning and reporting cohort studies using real-world data



Types of non-randomised study design

Overview

A large variety of study designs can be used to estimate the effects of interventions, exposures or policies. The preferred study design will be context dependent. It may depend on whether variation in the exposure is within individuals over time, between individuals, or between other groups such as healthcare providers. In general, confidence in non-randomised study results is strengthened if results are replicated using different study designs or analytical methods, known as triangulation (Lawlor et al. 2016).

One important distinction is between interventional and observational studies. In interventional studies, individuals (or groups of individuals) are allocated to 1 or more interventions according to a protocol. Allocation to interventions can be random, quasi-random or non-random. In observational studies, interventions are not determined by a protocol but instead according to the preferences of health and social care professionals and patients. Hybrid studies may make use of both interventional and observational data. In this section we focus on observational and hybrid studies only.

Both interventional and observational studies can be uncontrolled. Uncontrolled studies are appropriate only in rare cases, in which the natural course of the disease is well understood and highly predictable and the treatment effect is very large (see ICH E10 choice of control group in clinical trials and Deeks et al. 2003). In most cases a comparison group is needed to generate reliable and informative estimates of treatment effects. Controlled studies can make use of variation in exposures and outcomes across individuals (or groups), within individuals (or groups) over time, or both. In this section we focus on controlled studies.

Below we discuss types of comparative studies. Some taxonomies distinguish between prospective studies (involving primary data collection) and retrospective studies (based on already collected data). This distinction does not necessarily convey information about study quality and so we advise against its use (Dekkers and Groenwold 2020).

Cohort studies

In cohort studies, individuals are identified based on their exposures and outcomes compared during follow up. Usually, cohort studies will compare individuals subject to different exposures from the same data source. However, they can also combine data from different sources including from interventional and observational data sources. In this case, the observational data is used to form an external control to the intervention used in the trial. The trial will often be an uncontrolled single-arm trial but could also be an arm from a controlled trial. External data can also be used to augment concurrent controls within a randomised controlled trial.

External controls can also be formed from data from previous clinical trials. A potential advantage of such studies is greater similarity in patient inclusion criteria, follow up and outcome determination. Often only aggregate rather than individual patient-level data will be available from previous trials. NICE Decision Support Unit's technical support document 18 describes methods for unanchored indirect comparisons with aggregated data.

In the following study design and analysis sections, we focus on cohort studies including those using external control from real-world data sources which are the most common non-randomised study designs informing NICE guidance. Other study designs including quasi-experimental designs or self-controlled studies may be relevant in some contexts as outlined below.

Self-controlled studies

Self-controlled, or 'within-subject', designs make use of variation in exposure status within individuals over time. These include case-crossover, self-controlled case series, and variants of these designs. They are most appropriate for transient exposures with acute-onset events (Hallas and Pottegard 2014). While primarily used in studies of adverse effects of medicines (including vaccines), they have been used to assess the effects of oncology medicines using the experiences of individuals on prior lines of therapy (Hatswell and Sullivan 2020). This is most relevant if appropriate standard-of-care comparators are not available.

A key advantage of self-controlled methods is the ability to control for confounders (including unmeasured or unknown confounders) that do not vary over time, such as genetic inheritance, or vary slowly like many health behaviours. However, it is still necessary to adjust for covariates that may change over time (for example, disease severity). Such methods generally either assume no time-based trends in outcomes or try to model the trend statistically. These approaches can often be strengthened by the addition of control groups of people not exposed to the interventions.

Cross-sectional studies

In cross-sectional studies information on current exposures and outcomes is collected at a single time point. While they can be used to estimate intervention effects, they are less reliable than longitudinal studies (such as cohort studies) if there is need for a clear temporal separation of exposures and outcomes.

Case-control studies

In case-control studies individuals are selected based on outcomes, and odds of exposures are compared. Case-control studies embedded within an underlying cohort are known as nested case-cohort studies. Case-control studies conducted within existing database studies are generally not recommended because they use less information than

cohort studies (Schuemie et al. 2019). Case-control studies are most useful for rare outcomes or if there is a need to collect further information on exposures, for example, from manual medical record review or primary data collection.

Quasi-experimental studies

Quasi-experimental studies and natural experiments exploit external variation in exposure across people or over time (an 'instrument') that is otherwise unrelated to the outcome to estimate causal effects (Reeves et al. 2017, Matthay et al. 2019). Common quasi-experimental methods include instrumental variable analysis, regression discontinuity, interrupted time series and difference-in-difference estimation. They are frequently used in public health settings when randomisation is not always feasible but have also been used in medical technologies evaluations (see NICE medical technologies guidance on Sleepio to treat insomnia and insomnia symptoms).

Instrument-based approaches may be useful if:

- confounding because of unknown or poorly measured confounders is expected
- an appropriate instrument is available that is associated with the exposure of interest and does not affect the outcome except through the exposure.

Examples of instruments that have been used in healthcare applications include variation in physician treatment preferences or hospital formularies, genes, distance to healthcare providers or geographic treatment rates, arbitrary thresholds for treatment access, or time (for example, time of change to clinical guidelines that have immediate and substantial impacts on care patterns).

A key advantage of these approaches is in addressing confounding due to unobserved or poorly measured covariates. However, consideration needs to be given to the validity of the instrument in addition to other methodological challenges depending on the particular design used (see NICE Decision Support Unit's technical support document 17). Applications of these methods are usually strongly dependent on assumptions that are difficult to test, and a clear case for validity based on substantive knowledge and empirical justification is required.

Study design

In this section we present study design considerations for cohort and external control studies using real-world data. These approaches may also be useful for other non-randomised study designs.

The target trial approach

Non-randomised studies should be designed to mimic the randomised trial that would ideally have been performed unconstrained by ethical or feasibility challenges (Hernán and Robins 2016, Gomes et al. 2022). This process, known as the target trial approach (or trial emulation), requires developers to clearly articulate the study design and helps avoid selection bias because of poor design (Bykov et al. 2022). Usually, the target trial would be a pragmatic randomised trial representing the target population of interest and reflecting routine care. This approach forms the basis of the Cochrane ROBINS-I risk of bias tool for non-randomised studies (Sterne et al. 2016).

Studies should aim to emulate the target trial as closely as possible and, if this is not possible, trade-offs should be clearly described. In some cases, a data source may not be of sufficient relevance or quality to allow trial emulation. This can be particularly problematic for studies using real-world data to form an external control because differences in terms of patients, settings, care, data collection and time periods can limit the comparability between the trial and the real-world data (Gray et al. 2020, Pocock 1976). Sometimes it will not be possible to adequately emulate a target trial with real-world data and bespoke data collection may be needed.

The target trial can be defined across 7 dimensions: eligibility criteria, treatment strategies, assignment procedure, follow-up period, outcomes, causal effect of interest and analysis plan. We describe each dimension below and provide considerations for those developing evidence to inform NICE guidance.

Eligibility criteria

For most studies, the eligibility criteria should mimic a hypothetical pragmatic trial by reflecting the clinical pathways (including diagnostic tests) and patients seen in routine care in the NHS. For external control studies, the focus should be on matching the eligibility criteria from the interventional study rather than the broader target population. As in a trial, eligibility criteria should be based on variables recorded before treatment

assignment.

If heterogeneity is anticipated in the intervention effects, subgroup analysis can be done. The subgroups should be defined upfront when planning the study.

Treatment strategies

Treatment strategies include the intervention of interest and any comparators. Comparators could be different levels of an exposure (for example, different doses of a medicine), a different intervention, or the absence of intervention. In observational data it is very difficult to emulate a placebo-controlled trial because of higher risk of selection bias and intractable confounding.

Comparators that are for the same (or similar) treatment indication (that is, active comparators) are preferred to comparison with those not receiving an intervention. Active comparators reduce the risk of confounding by indication by ensuring greater similarity of patients having different interventions. If routine follow-up procedures are similar across interventions this also reduces the risk of detection bias. The active comparator should ideally reflect established practice in the NHS.

For studies of interventions, new (or incident) user designs are generally preferred to studies of prevalent users (those who have already been using the intervention for some time) because of the lower risk of selection bias and better emulation of trial designs. Prevalent users have, by definition, remained on-treatment and survived for some period of follow up. When making use of already collected data, new users are typically defined using an initial period in which the individual was not observed to use the intervention of interest (known as the 'washout' period in pharmacoepidemiology). A further advantage of new-user designs is the ability to estimate time-varying hazards from treatment initiation. The inclusion of prevalent users may be needed if the effects of interventions are cumulative, there are too few incident users in the data, or follow up is limited (Vandenbroucke and Pearce 2015, Suissa et al. 2016).

Data on comparators would ideally come from the same period as the intervention as well as from the same healthcare system and settings. This is to minimise any differences between treatment groups resulting from differences in care access, pathways (including diagnostic tests) or time-based trends in outcomes.

Assignment procedure

In randomised controlled trials, individuals (or groups) are randomly assigned to interventions. If possible, providers, patients and analysts are blinded to this assignment. Neither random assignment nor blinding are possible in observational studies. With sufficient information on confounders, random assignment can, however, be approximated through various analytical approaches (see the section on analysis).

In some applications, individuals will meet eligibility criteria at multiple time points. For example, they may start treatment more than once after a sufficient period without exposure (or 'washout' period). There are several approaches to deal with this including using only the first eligible time point, a random eligible time or all eligible time points (Hernán and Robins 2016).

Follow-up period

The start and end of follow up must be defined. The start of follow up should ideally begin at the same time at which all eligibility criteria are met and the intervention is assigned (or just after). If a substantial latency period is expected between treatment initiation and outcomes, it may be necessary to define an induction period before which outcomes are not counted. This can reduce the risk of reverse causation, in which the outcome influences the exposure.

The follow-up period should be long enough to capture the outcomes of interest but should not exceed the period beyond which outcomes could be reasonably impacted by the intervention (known as the exposure-effect window). Censoring events should be clearly defined and will depend on the causal effect of interest.

Outcomes

Primary and secondary outcomes should be defined and can include both patient and health system outcomes (such as resource use or costs). Patient outcomes should reflect how a patient feels, functions, or how long a patient lives. This includes quality of life and other patient-reported outcome measures. Objective clinical outcomes (such as survival) are typically subject to a lower risk of bias than subjective outcomes if outcome detection or reporting could be influenced by known treatment history.

For a surrogate outcome there should be good evidence that changes in the surrogate

outcome are causally associated with changes in the final patient outcomes of interest (Ciani et al. 2017).

While outcome ascertainment is not blinded in observational data, analysts can be blinded to outcomes before finalising the analysis plan (see the section on analysis).

Causal effect of interest

Researchers should describe the causal effect of interest. Trials are usually designed to estimate 1 of 2 causal effects: the effect of assignment to an intervention (intention-to-treat) or the effect of adhering to treatment protocols (per-protocol). It is not usually possible to estimate the effect of treatment assignment using observational data because this is not typically recorded. However, it can be proxied using treatment initiation (the as-started effect). The equivalent of the per-protocol effect is sometimes called the on-treatment effect.

The as-started effect is usually of primary interest to NICE. However, if treatment discontinuation (or switching) is substantial or is not expected to reflect routine practice or outcomes in the NHS, it is important to present results from the on-treatment analysis. On-treatment analyses may also be most appropriate for the analysis of safety and adverse events. The on-treatment effect can also be extended to cover dynamic treatment strategies such as treatment sequences or other complex interventions which are of interest to NICE.

Analysis plan

The analysis plan should describe how the causal effect of interest is to be estimated, taking into account intercurrent events. Intercurrent events are events occurring after treatment initiation (such as treatment switching or non-adherence) that affect the interpretation of the outcome of interest. This is supported by the estimand framework (for further information, see ICH E9 [R1] addendum on estimands and sensitivity analysis in clinical trials).

The relevance of intercurrent events will depend on the causal effect of interest. In an as-started analysis, treatment discontinuation, switching or augmentation can usually be ignored. However, if these changes are substantial there is a risk of increasing exposure misclassification over time. In most cases this would bias estimates of effect towards the null.

In an on-treatment analysis or when modelling dynamic treatment strategies, the follow up is often censored once the patient stops adhering to the treatment plan plus some biologically informed effect window. For medicines (and some devices) continued exposure is proxied by dates of prescriptions and expected period of use (for example, derived from number of days' supply), with some grace period between observations permitted. Particular attention needs to be given to the possibility of informative censoring, which causes bias if censoring depends on outcomes and differs across interventions, and time-varying confounding.

Further content on statistical analysis including addressing confounding, informative censoring, missing data and measurement error is presented in the analysis section.

Panel 1 shows examples of using the target trial approach:

Panel 1: examples of the target trial approach

Example 1: What is the effect of initiating HRT on coronary heart disease in postmenopausal women?

The Women's Health Initiative randomised controlled trial showed that initiating treatment with hormone replacement therapy increased the risk of coronary heart disease in postmenopausal women. This contradicted earlier observational studies that found a reduction in the risk of coronary heart disease. Hernán et al. 2008 followed a target trial approach, replicating as far as possible the Women's Health Initiative trial using data from the Nurses' Health Study. They were able to show that the difference in results between the trial and observational studies resulted from the inclusion of prevalent users of hormone replacement therapy in the observational cohort. These women had already survived a period of time on-treatment without experiencing the outcome. Following a new-user design (as well as other principles of the target trial approach) they were able to produce effect estimates consistent with the trial.

Example 2: What is the optimal estimated glomerular filtration rate (eGFR) at which to initiate dialysis treatment in people with advanced chronic kidney disease?

The IDEAL randomised controlled trial showed a modest reduction in mortality and cardiovascular events for early versus late initiation of dialysis. The average eGFR scores in the early and late treatment arms were 9.0 and 7.2 mL/min/1.73 m^2, respectively. There therefore remains considerable uncertainty about the optimal time to initiate dialysis. Fu et al. 2021 emulated the IDEAL trial using data from the National Swedish Renal Registry and were ability to produce similar results over the narrow eGFR separation achieved in the trial. They were then able to extend the analysis to a wider range of eGFR values to identify the optimal point at which to initiate dialysis therapy.

Example 3: What is the effect of initiating treatment with fluticasone propionate plus salmeterol (FP-SAL) versus 1) no FP-SAL or 2) salmeterol only on COPD exacerbations in people with COPD?

The TORCH trial found that treatment with FP-SAL was associated with a reduction in the risk of COPD exacerbations compared with no FP-SAL or salmeterol only. However, the trial excluded adults aged above 80 years and those with asthma or mild COPD. There is uncertainty about the extent to which results from the TORCH trial

apply to these patients. Wing et al. 2021 were able to replicate the findings of the TORCH trial for COPD exacerbations using primary care data from Clinical Practice Research Datalink in England for the comparison with salmeterol only but not with no FP-SAL. This reflects the challenge in emulating a trial with placebo control. By extending their analysis to a wider target population they were able to demonstrate evidence of treatment effect heterogeneity by COPD severity but not by age or asthma diagnosis.

Analysis

Addressing risk of confounding bias

Identification and selection of confounders

Potential confounders should be identified before analysis, based on a transparent, systematic and reproducible process. Key sources of evidence are published literature and expert opinion. Consideration should be given to the presence of time-varying confounders. These affect the outcome and future levels of the exposure and can be affected by previous levels of the exposure. They are especially relevant when modelling time-varying interventions or dynamic treatment strategies or addressing informative censoring.

Developers should outline their assumptions about the causal relationships between interventions, covariates and outcomes of interest. Ideally, this would be done using causal diagrams known as directed acyclic graphs (Shrier and Platt 2008).

Inappropriate adjustment for covariates should be avoided. This may result from controlling for variables on the causal pathway between exposure and outcomes (overadjustment), colliders or instruments. Confounders that may change value over time should be recorded before the index date, except when using statistical methods that appropriately address time-varying confounding.

The selection of covariates may use advanced computational approaches such as machine learning to identify a sufficient set of covariates, for example, when the number of potential covariates is very large (Ali et al. 2019, Tazare et al. 2022). The use of these

methods should be clearly justified and their consistency with causal assumptions examined. Choosing covariates based on statistical significance should be avoided.

Selecting methods for addressing confounding

Adjusted comparisons based on clear causal assumptions are preferred to naive (or unadjusted) comparisons. Statistical approaches should be used to address confounding and approximate randomisation (see the section on assignment procedure).

Various approaches can be used to adjust for observed confounders including stratification, matching, multivariable regression and propensity score methods, or combinations of these. These methods assume no unmeasured confounding. Simple adjustment methods, such as stratification, restriction and exact matching, may be appropriate for research questions in which confounding is well understood and there are only a small number of confounders that are well recorded.

If there are many potential confounders, more complex methods such as multivariable regression and propensity score (or disease risk score) methods are preferred. Propensity scores give the probability of receiving an intervention based on observed covariates. Several methods use propensity scores including matching, stratification, weighting and regression (or combinations of these). General discussions of the strengths and weaknesses of these different approaches can be found in Ali et al. 2019. The choice of method should be justified and should be aligned with the causal effect of interest.

There is mixed evidence on the relative performance of regression and propensity score methods for addressing confounding bias (Stürmer et al. 2006). However, using propensity score methods may have advantages in terms of the transparency of study conduct:

- Propensity scores are developed without reference to outcome data, which can reduce the risk of selective reporting of results when combined with strong research governance processes.
- With certain propensity score methods it is possible to examine the similarity of intervention groups in terms of observed covariates, providing evidence on the extent to which comparability was achieved. Absolute standardised differences of less than 0.1 are generally considered to indicate good balance although small absolute differences may still be important if the variable has a strong effect on the outcome.

Regression and propensity score methods may also exclude some participants to enhance

the similarity of people across intervention arms or levels. When using such methods, trade-offs between internal validity, power and generalisability should be considered. For studies of comparative effects, internal validity should generally be prioritised.

Time-varying confounders should typically not be adjusted for using the above methods. It may be acceptable for on-treatment analyses if confounders that vary over time are not affected by previous levels of the intervention but this is uncommon. G-methods including marginal structural models with weighting are preferred (Pazzagli et al. 2017, Mansournia et al. 2017). Adjustment for time-varying confounders requires high-quality data over the whole follow-up period.

Various sensitivity and bias analyses can be used to adjust for bias because of residual confounding or to explore its likely impact (see the section on assessing robustness of studies). This may be informed by external data on confounder-outcome relationships or data from a data-rich subsample of the analytical database, if available (Ali et al. 2019). Negative controls (that is, outcomes that are not expected to be related to the intervention) may also be useful (Lipsitch et al. 2010).

If there are multiple potential sources of suitable real-world data to provide external control to trial data, developers should consider whether to estimate effects separately for each data source or to increase power by pooling data sources. Data sources should only be pooled when there is limited heterogeneity between sources in terms of coverage and data quality. Individual estimates of effects for each data source should always be provided.

External controls can also be used to supplement internal (or concurrent) controls in randomised controlled trials. There are several methods available to combine internal and external controls, which place different weight on the external data (NICE Decision Support Unit report on sources and synthesis of evidence).

Instrument-based methods (or quasi-experimental designs) can be used to address unobserved confounding (Matthay et al. 2019). Further technical guidance on methods for addressing baseline confounding due to observed and unobserved characteristics using individual patient-level data is given in NICE's Decision Support Unit technical support document 17.

Addressing information bias

Limitations in data quality including missing data, measurement error or misclassification can cause bias and loss of precision. Here we describe analytical approaches to address information bias. The information needed to understand data suitability will provide an insight into the likely importance of information bias (see the section on assessing data suitability).

Informative censoring

Censoring occurs in longitudinal studies if follow up ends before the outcome is fully observed. It can happen because the data collection period ends (administrative censoring), loss to follow up, occurrence of events such as treatment switching, non-adherence, or death depending on the analysis. It may be induced by analytical strategies such as cloning to avoid time-related biases in studies without active comparators (Hernán and Robins 2016).

Censoring can create bias if it is informative (that is, it is related to the outcomes and treatment assignment). For example, in on-treatment analyses, if people on an experimental drug were less likely to adhere to the treatment protocol because of a perceived lack of benefit this could lead to informative censoring. When modelling effects on-treatment or dynamic treatment strategies, censoring because of treatment switching is likely to be informative. Methods to address informative censoring are similar to those for time-varying confounding such as marginal structural models with weighting or other G-methods (Pazzagli et al. 2017). Methods for dealing with missing data may also be used (see the section on missing data).

Missing data

The impact of missing data depends on the amount of missing data, the variables that have missing data, and the missing data mechanism. Developers should compare patterns of missingness across exposure groups and over time, if relevant, considering causes of missingness and whether these are related to outcomes of interest. Missing data on outcomes may arise for a number of reasons including non-response to questionnaires or censoring.

If the amount of missing data is low and likely to be missing completely at random, complete records analysis will be sufficient. Advanced methods for handling missing data

include imputation, inverse probability weighting and maximum likelihood estimation. Most of these methods assume the missing data mechanism can be adequately modelled using available data (that is, missing at random). If this is not the case, sensitivity or bias analysis may be preferred (see the section on assessing robustness). A framework for handling missing data is provided in Carpenter and Smuk 2021.

Measurement error and misclassification

Measurement error describes the extent to which measurements of study variables deviate from the truth. For categorical variables, this is known as misclassification. The impact of measurement error depends on the size and direction of the error, the variables measured with error, and whether error varies across intervention groups. Measurement error can induce bias or reduce the precision of estimates.

Random measurement error in exposures tends to (but does not always) bias estimates of treatment effects towards the null (van Smeden et al. 2020). Random measurement error in continuous outcomes reduces the precision of estimates but provides unbiased estimates of comparative effects. For risk ratios and rate ratios, non-differential misclassification of a categorical outcome provides unbiased estimates of comparative effects when specificity is 100%, even if sensitivity is low. So, it is often recommended to define outcome variables to achieve high specificity.

Differential measurement error in exposures, covariates or outcomes generally produces biased estimates of comparative effects but the direction of bias can be hard to predict. If data is available on the likely structure and magnitude of measurement error (for example, through an internal or external validation study), this information can be incorporated into analyses using calibration or other advanced methods (van Smeden et al. 2020).

Addressing external validity bias

Assessing external validity

This section focuses on methods to assess and address external validity bias resulting from differences in patient characteristics (for example, age, disease risk scores) between the analytical sample and the target population. Importantly, differences in patient characteristics may not be the only, or most important, sources of external validity bias. Developers should also consider differences in: setting (for example, hospital type and access to care), treatment (for example, dosage or mode of delivery, timing, comparator

therapies, concomitant and subsequent treatments) and outcomes (for example, follow-up, measurements, or timing of measurements). Identifying a suitable data source, using a target trial approach and using internally valid analysis methods remain the primary approaches by which external validity can be achieved.

To assess external validity, an explicit definition of the target population is needed and suitable reference information. Information can be drawn from published literature, context-relevant guidelines, or bespoke analysis of data from the target population alongside information gathered during the data suitability assessment.

To assess differences between the analytical sample and target population for patient characteristics, several tests are available:

- averages and distributions of individual variables can be compared (for example, using absolute standardised mean differences);
- multiple variables can be compared simultaneously using propensity scores (here, reflecting a patient's propensity for being selected into the study) which also support measures of differences arising from joint distributions of patient characteristics (for example, Tipton 2014).

In studies of relative treatment effects, differences observed between the analytical sample and target population do not necessarily lead to concerns about external validity bias unless those differences are considered to be important treatment effect modifiers. This depends on the causal effect of interest, the extent of heterogeneity in the treatment effect, and whether this has been adequately modelled. Assumptions about the causal relationships between interventions, outcomes, and other covariates can be outlined to help identify potential treatment effect modifiers, for example, using directed acyclic graphs (Shrier and Platt 2008). Under certain conditions, treatment effect modification can also be investigated statistically (Degtiar and Rose 2022).

In studies of absolute treatment effects, assessment of external validity requires consideration of all differences that are prognostic of the outcome of interest, not only treatment effect modifiers.

Methods to minimise external validity bias

Methods to adjust for external validity bias are similar to those which adjust for confounding bias, including matching, weighting, and outcome regression methods. These

approaches can also be combined for additional robustness.

- Matching and weighting methods balance individual characteristics associated with selection into the sample (for example, using propensity scores).
- Regression methods model outcomes in the analytical sample and then standardise model predictions to the distribution of covariates in the target population.

Degtiar and Rose 2022 provides further guidance on these methods, including approaches for when only summary-level data is available for the target population. Adjustment approaches are unlikely to perform well when the target population is poorly represented in the analytical sample, that is, where there is insufficient overlap for important covariates, or across strata of these variables. Successful application of these methods also depends on good internal validity of analyses and consistency in measurements of outcomes, treatments, and covariates across settings.

Where the sample is drawn from an entirely different population to the target population, judgements of similarity will require stronger assumptions. Pre-specified, empirical assessments of 'transportability' for the decision context could provide supportive evidence (for example, see Ling et al. 2023). In all cases, sensitivity analyses are recommended to explore potential violation of study assumptions (for example, see Dahabreh et al. 2023, Nguyen et al. 2018).

Assessing robustness of studies

The complexity of studies of comparative effects using real-world data means developers must make many uncertain decisions and assumptions during data curation and analysis. These decisions can have a large impact, individually or collectively, on estimates of comparative effects. It is therefore essential that the robustness of results to deviations in these assumptions is demonstrated. We describe key sensitivity analyses across several domains in table 3. Which sensitivity analyses to focus on will vary across use cases depending on the strengths and weaknesses of the data as well as the areas in which the impact of bias, study assumptions and uncertainty are greatest. These approaches can be applied directly to measures of clinical effectiveness or propagated through to cost-effectiveness analyses.

For key risks of bias (for example, those arising because of unmeasured confounding, missing data or measurement error in key variables), quantitative bias analysis may be

valuable. Quantitative bias analysis describes a set of techniques that can be used to:

- examine the extent to which bias would have to be present to change results or affect a threshold for decision making, or
- estimate the direction, magnitude and uncertainty of bias associated with measures of effect.

Methods that examine the extent to which bias would have to be present to change study conclusions tend to be simpler and include the e-value approach. These approaches are most useful when exploring a single unmeasured source of bias, however sources of bias are often multiple and may interact. Developers should consider and pre-specify a plausible level of bias in the parameter before application of these methods. More sophisticated approaches look to model bias and incorporate it into the estimation of effects (Lash et al. 2014). Bias parameters can be informed by external information or data-rich subsamples of the analytical data source. The identification and validity of external bias should be clearly described and justified.

Bias analysis may be particularly valuable in studies using real-world data external controls if differences between data collection, settings and time may reduce comparability of data. Panel 2 shows an example of bias analysis in practice, and table 3 shows examples of sensitivity analysis.

Panel 2: example of bias analysis

What is the effectiveness of the ALK-inhibitor alectinib compared with ceritinib in crizotinib-refractory, ALK-positive non-small-cell lung cancer?

The comparative effectiveness of alectinib versus ceritinib on overall survival in patients with ALK-positive non-small-cell lung cancer is uncertain because of a lack of head-to-head trials. Wilkinson et al. 2021 used real-world data on ceritinib from the Flatiron Health database (derived from US electronic health records) to form an external control to the alectinib arm of a phase 2 trial. The authors found a significant improvement in survival for those initiating alectinib. However, the study was at risk of residual bias from unmeasured confounding and missing baseline data on Eastern Cooperative Oncology Group Performance Status (ECOG PS) in patients having ceritinib (47% of patients had missing data).

Bias analysis methods were used to explore these risks. The e-value approach was used to estimate the relative risk of an unobserved confounder between intervention and mortality that would be needed to remove the treatment effect. The estimated relative risk of 2.2 was substantially higher than for any observed confounders and considered unlikely given the estimated imbalance for important but poorly captured confounders.

For missing ECOG data they assumed the causes of missing data were non-random and missing data values in the ceritinib arm were likely to be worse than expected based on multiple imputation. They argued that no plausible assumptions about missing data could explain the observed association between intervention and mortality.

Table 3

Examples of sensitivity analyses to examine robustness of results to data curation, study design, and analysis decisions

Domain	Example sensitivity or bias analysis
Exposure misclassification	• On-treatment analyses • Vary exposure definitions including, if relevant, days' supply, grace period, washout period, exposure-effect window and latency
Outcome misclassification	• Adjust for known performance metrics • Quantitative bias analysis
Population	• Alternative patient eligibility criteria
Detection bias	• Include measures of healthcare use as covariates • Restrict to those with regular contact with the health system before baseline
Follow-up time	• As-started and on-treatment analyses • Restrict outcome period so it is similar between groups for informative censoring • Prevalent-user and new-user analyses
Reverse causation	• Introduce or change lag time between exposure end and start of follow up for outcomes

Domain	Example sensitivity or bias analysis
Confounding	• Add or remove selected confounders • Extend look-back period over which covariates are identified • Use negative controls (also known as falsification endpoints or probe variables) to estimate comparative effects using the same model on outcomes, which should not be related to treatment (results from these can also be used to calibrate effect estimates) • Propensity score calibration to adjust observed effect estimates for unmeasured bias using variables observed in a validation study • Quantitative bias analysis
Missing data	• Use different methods • Include missing variable indicators for covariates in statistical models • Quantitative bias analysis (for instance, assuming missing not at random mechanisms)
Model specification	• Vary model specifications • Use analytical approaches with different assumptions (triangulation)
Data curation	• Alternative categorisations of continuous variable or adjust data exclusions

Reporting

We provide general principles for the transparent reporting and good conduct of real-world evidence studies in the section on conduct of quantitative real-world evidence studies. The following reporting considerations are especially important for comparative effects studies:

- Justification of the use of real-world evidence. This should cover, as relevant, the reasons for the absence of randomised evidence, the limitations of existing trials and the ability to produce meaningful real-world evidence for the specific research question.
- Publish a study protocol (including statistical analysis plan) on a publicly accessible platform before the analysis is done.
- Report studies in sufficient detail to enable the study to be reproduced by an independent researcher.
- Present study design diagrams.
- For each data source, provide the information needed to understand data provenance and fitness for purpose (see the section on assessing data suitability).
- Justify the use of statistical method for addressing confounding and report methods clearly (see appendix 3).
- Clearly describe the exclusion of patients from the original data to the final analysis, including reasons for exclusion using patient flow (or attrition) diagrams.
- Present characteristics of patients across treatment groups, before and after statistical adjustment if possible. For external control studies, differences in variable definitions and data collection should be clearly described.
- Present results for adjusted and unadjusted analyses and for all subgroup and sensitivity and bias analyses.

Quality appraisal

Evidence developers should identify risks of bias at the study planning stage. These should be described alongside how design and analytical methods have been used to address them, and how robust the results are to deviations from assumptions in the main analysis using sensitivity or bias analysis. This can be done for specific domains of bias using the reporting methods in appendix 2. This information will help those completing (or critically appraising) risk of bias tools. The preferred risk of bias tool for non-randomised studies is the ROBINS-I (Sterne et al. 2016) but it should be recognised that it may not cover all risks of bias (D'Andrea et al. 2021). It should be recognised that the uncertainty in non-randomised studies will not typically be fully captured by the statistical uncertainty in

the estimated intervention effect (Deeks et al. 2003).

Developers should comment on the generalisability of study results to the target population in the NHS. This may draw on differences in patients, care settings, treatment pathways or time and is supported by information provided from the data suitability assessment. Developers should also discuss any methods used to address external validity bias, with the results of adjusted and unadjusted analysis presented.

Appendix 1 – Data Suitability Assessment Tool (DataSAT)

See tools and resources for a downloadable DataSAT assessment template.

DataSAT assessment template

Research question

Add the research question here.

Data provenance

Item	Response
Data sources	For each contributing data source provide the name, version and date of data cut. Provide links to their websites, if available.
Data linkage and data pooling	Report which datasets were linked, how these were linked, and performance characteristics of the linkage. Note whether linkage was done by a third party (such as NHS Digital). Clearly describe which data sources were pooled.
Type of data source	Describe the types of data source (for example, electronic health record, registry, audit, survey).
Purpose of data collection	Describe the main purpose of data collection (for example, clinical care, reimbursement, device safety, research study).

Item	Response
Data collection	Describe the main types of data collected (for example, clinical diagnoses, prescriptions, procedures, patient experience data), how data was recorded (for example, clinical coding systems, free text, remote monitoring, survey response), and who collects the data (for example, healthcare professional, self-reported, digital health technology). If the nature of data collection has changed during the data period (for instance, change in coding system or practices, data capture systems) describe the changes clearly. Any differences between data providers in how and what data were collected and its quality should be described. If additional data collection was done for a research study please describe, including how the validity and consistency of data collection was assured (for example, training).
Care setting	State the setting of care for each dataset used (for example, primary care, secondary care, specialist health centres, social services, home use [for wearable devices, or self-reported data on apps or websites]).
Geographical setting	State the geographical coverage of the data sources.
Population coverage	State how much of the target population is represented by the dataset (for example, population representativeness or patient accrual).
Time period of data	State the time period covered by the data.
Data preparation	Provide details of whether raw data were accessed for analysis, or whether the data owner had undertaken any data preparation steps such as cleansing or transformation. Mention whether centralised transformation to a common data model was undertaken. Include links to any relevant information including common data model type and version number and details of mapping. Full details of data preparation specific to addressing the research question is covered in the section on reporting on data curation.
Data governance	Provide the details of the data controller and funding for each source. Describe the information governance processes for data access and use.
Data specification	Note whether a data specification document is available. This may include a data model, data dictionary, or both.

Item	Response
Data management plan and quality assurance methods	Note whether a data management plan, documentation of source quality assurance methods is available with links to relevant documents.
Other documents	Note whether any other documentation is available. Provide hyperlinks or citations to key publications, if available. If the dataset is available from the Health Data Research UK (HDRUK) innovation gateway, provide the hyperlink to its profile on the HDRUK website.

Data quality

Details of data quality should be provided for key study variables including population eligibility criteria, outcomes, interventions or exposures, and covariates.

Study variable	Target concept	Operational definition	Quality dimension	How assessed	Assessment result
What type of variable (for example, population eligibility, outcome)	Define the target concept (for example, myocardial infarction [MI])	Define operational definition. For example, MI defined by an ICD-10 code of I21 in the primary diagnosis position	Choose: accuracy or completeness	Describe how quality was assessed. Provide reference to previous validation studies if applicable.	Provide quantitative assessment of quality if available. For example, 'positive predictive value 85% (75% to 95%)'

Data relevance

Please see recommendations for reporting data relevance.

Item	Response
Population	Describe the extent to which the analytical sample reflects the target population. This should consider any data exclusions (for example, because of missing data on key prognostic variables).

Item	Response
Care setting	Describe how well the care settings reflect routine care in the NHS.
Treatment pathway	Describe how the treatment pathways experienced by people in the data reflects routine care pathways in the NHS (including any diagnostic tests).
Availability of key study elements	Note how the dataset met the requirements of the research question in terms of availability of the necessary data variables including key population eligibility criteria, outcomes, intervention and covariates (including confounders and effect modifiers).
Study period	State the extent to which the time period covered by the data provides relevant information to decisions. This should cover any important changes to care pathways (including tests) or background changes in outcome rates.
Timing of measurements	Describe whether the timing of measurements meet the needs of the research question.
Follow up	Note how the follow-up period available in the dataset is sufficient for assessing the outcomes.
Sample size	Provide the sample size of the target population in the dataset and demonstrate that it is adequate to generate robust results.

DataSAT – case study

Please note that the reporting for this case study is based on publicly available information in Wing et al. 2021.

Research question

What is the effect of the long-acting beta-2 agonist and inhaled corticosteroid combination product fluticasone propionate plus salmeterol compared with no exposure or exposure to salmeterol only in people with chronic obstructive pulmonary disease (COPD)?

Data provenance

Item	Response
Data sources	Clinical Practice Research Datalink (CPRD) GOLD Hospital episode statistics (HES) Admitted Patient Care data.
Data linkage and data pooling	CPRD and HES are linked. Patients are identified in a centralised linkage algorithm done by NHS digital. This uses an 8-step deterministic linkage algorithm based on 4 identifiers: NHS number, sex, date of birth and postcode. Linkage to HES data is possible for 75% of enrolled patients. See information on linked data for CPRD.
Type of data source	HES = administrative records CPRD = electronic health records
Purpose of data collection	Hospital Episode Statistics (HES) is derived from the Secondary Uses Service (SUS) data based on information submitted to NHS digital by healthcare providers. Data collection is primarily intended to support the reimbursement of hospitals for the provision of services in England. CPRD collects anonymised patient data from a network of GP practices across the UK. Initially this data is collected during a patient's time in primary care services.
Data collection	CPRD = demographics, clinical diagnoses (Read v2 or SNOMED-CT), tests (medcode or SNOMED-CT), prescriptions (prodcode) including dose, route of administration and duration. CPRD GOLD collects fully coded patient electronic health records from GP practices using the Vision software system. Data are recorded by health and care staff working within the Vision software. HES = diagnoses (ICD-10), procedures (OPCS-4), admission, discharge, type of care, basic demographics. HES data are collected during a patient's time at hospital and may be recorded during their interactions with health and care staff in the hospital and assembled by teams of clinical coders.
Care setting	HES = secondary care CPRD = primary care

Item	Response
Geographical setting	HES = England CPRD = a representative sample of UK general practices using Vision software. HES-linked CPRD data is available for England only.
Population coverage	CPRD GOLD has data for about 3 million currently registered people (around 4.74% of UK population). See CPRD data highlights HES data covers all NHS Clinical Commissioning Groups in England.
Time period of data	The CPRD-linked HES dataset covers from January 2000 to January 2017.
Data preparation	No details available for CPRD. However, general practices are included only after demonstrating their records are of research quality. HES applies centralised processing before the data are released for research: The rules that run during the processing of the HES data set. These are in place to improve the value and quality of the data and include rules that validate the data within certain fields, derive additional fields and values, remove records that are invalid or out of scope for the HES data set.
Data governance	CPRD is a centre of the MHRA, which is an executive agency of the Department of Health & Social Care (DHSC). DHSC is therefore the data controller for CPRD data. HES data is controlled by the Health and Social Care Information Centre (also known as NHS Digital). CPRD has received funding from the MHRA, Wellcome Trust, Medical Research Council, NIHR Health Technology Assessment programme, Innovative Medicines Initiative, UK Department of Health, Technology Strategy Board, Seventh Framework Programme EU, and various universities, contract research organisations and pharmaceutical companies. HES data collection is mandated and funded by the UK Government. Data protection and processing notice for CPRD. Hospital episode statistics GDPR webpage.

Item	Response
Data specification	Fields in HES are derived from the NHS data model and the NHS data dictionary. CPRD GOLD data specification document.
Data management plan and quality assurance methods	HES undertakes processing and data quality checks: The processing cycle and HES data quality. No data quality assurance information was identified for CPRD GOLD. However, records from individual general practices are assessed and only included in CPRD after being deemed of research quality.
Other documents	None.

Data quality

Study variable	Target concept	Operational definition	Quality dimension	How assessed	Assessment result
Population	COPD	CPRD diagnostic (Read v2) codes for COPD (see codelist in supplementary material of Quint et al. 2014)	Accuracy	Previously published validation study comparing algorithms for identifying people with COPD with physician review questionnaire as gold standard (Quint et al. 2014)	Positive predictive value (PPV): 87% (95% Confidence interval [CI] 78% to 92%)

Study variable	Target concept	Operational definition	Quality dimension	How assessed	Assessment result
Population	Disease severity	Global Initiative for Chronic Obstructive Lung Disease (GOLD) stage derived from spirometry measurements (see codelist)	Completeness	Proportion of patients with missing spirometry data	20%
Intervention	Fluticasone propionate + salmeterol	CPRD prescribing record matching definition of drug treatment determined by codelist	Accuracy	CPRD prescribing data is expected to be highly accurate	n/a

Study variable	Target concept	Operational definition	Quality dimension	How assessed	Assessment result
Outcome	COPD exacerbation	Any of the following: CPRD diagnostic (Read) code for lower respiratory tract infection or acute exacerbation of COPD A prescription of a COPD-specific antibiotic combined with oral corticosteroid (OCS) for 5 to 14 days A record (Read code) of 2 or more respiratory symptoms of AECOPD with a prescription of COPD-specific antibiotics and/or OCS on the same day. See codelist	Accuracy	Previously published validation study comparing algorithms for identifying people with COPD exacerbations with physician review questionnaire as gold standard (Rothnie et al. 2016)	PPV: 86% (95% CI 83% to 88%) Sensitivity: 63% (95% CI 55% to 70%)

Study variable	Target concept	Operational definition	Quality dimension	How assessed	Assessment result
Outcome	All-cause mortality	Record in Office for National Statistics (ONS) mortality statistics (centrally linked to CPRD data)	Accuracy	ONS mortality records are the gold standard data for deaths	n/a
Covariate (confounder)	Alcohol intake	Reported directly in CPRD (closest to index date)	Completeness	Proportion of patients with missing data on alcohol intake	30%

Data relevance

Item	Response
Population	Patients in CPRD have similar demographic characteristics to the wider UK population. Results from CPRD are generally expected to generalise to the wider eligible population. Complete records analysis was done excluding records with missing data on socioeconomic status, alcohol consumption and BMI. All these variables had less than 5% of the data missing. Around one-fifth of patients were excluded because they did not have spirometry measurements recorded in the CPRD. Those without measurements tend to have less contact with health services, which could impact on the generalisability of results.
Care setting	Appropriate. COPD drugs are typically administered in primary care (CPRD) while relevant events may be observed in primary or secondary care (CPRD or HES).
Treatment pathway	The data represents routine practice in the NHS.

Item	Response
Availability of key study elements	Sufficient data on exposures and outcomes are available. Although only prescribing and not dispensing data is available from CPRD this is expected to be a good proxy for dispensing. No information was available on negative reversibility spirometry results which may be a key confounder. Dosage information is limited in CPRD.
Study period	There have been no major changes to UK clinical practice for the management of COPD since the study period.
Timing of measurements	The longitudinal nature of the analysis allows for the research question to be answered. The date of entry is expected to reflect the actual timing of clinical events well.
Follow up	The average follow up of 2 years is sufficient for the primary outcome of COPD exacerbations to have occurred.
Sample size	The needed sample size for COPD exacerbations was estimated to be 600 per arm at 80% and 5% significance (see Wing et al. 2021 for details). The actual sample size of about 2,500 per arm far exceeds this.

Appendix 2 – Reporting on methods used to minimise risk of bias

See tools and resources for a downloadable methods to address bias reporting template.

Methods reporting template

Form for reporting on methods used to minimise risk of bias

Type of bias	How bias was addressed or assessed
Selection bias at study entry	Selection bias at study entry can arise for several reasons including selection of patients based on eligibility criteria related to the exposure and outcome, or from deviations between the date the patient meets eligibility criteria, the date treatment is assigned, and the start of follow up. Common types of time-related bias are prevalent-user bias, lead time bias, immortal time bias and depletion of susceptibles. Discuss the potential for selection bias at study entry and how this was addressed or investigated through study design, statistical analysis or sensitivity analysis.
Selection bias at study exit	A common cause of selection bias because of how individuals exit a study is informative censoring. This may be because of loss to follow up or the occurrence of censoring events. Discuss the possibility of informative censoring and how this was addressed in the analysis.

Type of bias	How bias was addressed or assessed
Addressing confounding	Describe the risk of confounding from unmeasured (or unknown) confounders, poorly measured confounders, or time-varying confounding. This should be informed by a systematic identification of potential confounders, clear causal assumptions including the possibility of time-varying confounding, and differences in baseline characteristics between comparison groups. Show how you dealt with any identified risk of confounding through study design (such as selection of a suitable active comparator) and analysis (using an appropriate statistical model, accounting for time-varying confounding). If possible, provide empirical data on the balance of baseline characteristics after adjustment. If concerns remain about residual confounding, show its impact on results has been assessed using sensitivity or bias analysis. Confirm that no covariates were inappropriately adjusted to induce bias. For example, show that no covariates on the causal pathway between interventions and outcomes were adjusted for (overadjustment). This may result from the use of covariates measured after the index date. Avoid adjustment for colliders or instruments. This can be informed by causal diagrams.
Detection bias	Describe the potential for detection bias resulting from differences in healthcare practices across comparison groups (for example, because of differential frequency or intensity of follow up, or different tests) or length of follow up. Describe how these have been dealt with through study design (for example, use of comparator with similar follow up) or analysis (for example, adjustment for healthcare use before index date).
Measurement error and misclassification	Describe the potential for bias from measurement error or misclassification (this should be informed by assessment of data suitability). Consider which variables are inaccurate, whether this is random or systematic, and how it differs across comparison groups. Show you addressed risks of bias through statistical analysis (for example, by incorporating external data or calibration) or assessed its impact on results using sensitivity or bias analysis.

Type of bias	How bias was addressed or assessed
Missing data	Describe the potential for bias from missing data (this should be informed by assessment of data suitability). Consider which variables have missing data, whether this is random or systematic, and how it differs across comparison groups. Show how you have addressed risks of bias using statistical methods (such as multiple imputation) and demonstrating their validity. If missingness may not be explainable by observed variables or has unknown mechanisms, sensitivity or bias analysis can be used to explore the impact of different missing 'not at random' assumptions.
Reverse causation	Describe the risk of reverse causation between the intervention and the outcome arising from causal relationships between variables, time lag between recording of data on interventions and outcomes, or care pathways. Describe how risk of reverse causation was addressed through study design (for example, induction periods or longitudinal follow up), analysis (for example, instrumental variables), or assessed through sensitivity analysis.

Methods reporting – case study 1

Please note that the reporting for this case study is based on publicly available information in Fu et al. 2021.

The study assesses the impact of initiating dialysis at different estimated glomerular filtration rates (eGFR) on cardiovascular events and survival in people with advanced chronic kidney disease. The study used data from the Swedish Renal Registry.

Example of completed methods reporting tool based on Fu et al. 2021.

Type of bias	How bias was addressed or assessed
Selection bias at study entry	Previous observational studies of the effects of the timing of dialysis initiation are at high risk of lead time and immortal time bias resulting from non-alignment of the time at which eligibility criteria were met, treatment assignment, and start of follow up. The study emulated a target trial informed by the IDEAL trial. To avoid issues with misspecification of time zero, the study used the cloning, censoring, and weighting method. Patients are cloned and assigned to each treatment according to eGFR (one of 15 treatment strategies in the base case) and are censored once they deviated from a given treatment strategy. The approach was validated by replicating results from the IDEAL trial over the range of eGFR values seen in the trial. Selection bias due to the choice of population was not an issue in this population-based study.
Selection bias at study exit	Selection bias can be induced by the censoring when patients stop adhering to the 'treatment strategy' if this is related to patient characteristics. Inverse probability of censoring weights were estimated using baseline and time-varying confounders to address censoring-induced selection bias. Loss to follow up is very low.

Type of bias	How bias was addressed or assessed
Addressing confounding	The outcome model adjusted for baseline measurements including demographics, laboratory measurements, prior treatment and hospitalisations. Time-varying confounders were adjusted for in censoring weights including current and previous measurements of eGFR. Data was not available on other potentially important confounders including muscle mass stores, uraemic symptoms, volume status, quality of life, or physical activity, and data was only available for subset of the cohort on urine albumin-creatinine ratio and plasma potassium. To assess the possibility of residual confounding, the study did the following sensitivity analyses: • adjusted for urine albumin-creatinine ratio and plasma potassium in the subset of patients with measurements and observed no impact on results • replicated the results of the IDEAL trial over the eGFR separation observed in the trial.
Detection bias	Outcomes included 5-year all-cause mortality and major adverse cardiovascular events (composite of cardiovascular death, non-fatal myocardial infarction, or non-fatal stroke). These are likely to be accurately observed regardless of small differences in level of surveillance, for example, resulting from earlier dialysis treatment.
Measurement error and misclassification	Timeliness and accuracy of variables extracted from the Swedish Renal Registry have previously been demonstrated. In particular, cardiovascular comorbidities have a very high positive predictive value, generally between 85% to 95%. eGFR was calculated with the Chronic Kidney Disease Epidemiology equation from routine plasma creatinine measurements. This has been shown to be accurate to within 30% of measured glomerular filtration rate 85% of the time.

Type of bias	How bias was addressed or assessed
Missing data	Data on initiation of dialysis and key outcomes are thought to be complete. Data on mandatory items such as eGFR is also very high. For non-mandatory data items in the registry, missingness was greater. For example, body mass index was missing in 26% of patients, urinary albumin to creatinine ratio in 44%, and potassium in 29%. This was assumed to be missing completely at random and determined by the preferences of the attending physician. Sensitivity analysis in the subset of people with data available showed had no impact on results.
Reverse causation	Reverse causation is not expected to be a problem in this analysis.

Methods reporting – case study 2

Please note that the reporting for this case study is based on publicly available information in Wilkinson et al. 2021.

The study estimates the comparative effectiveness of alectinib versus ceritinib on survival in people with ALK-positive non-small-cell lung cancer. The study uses real-world data on ceritinib from Flatiron Health to form an external control to patients having alectinib in phase 2 trials.

Example of completed methods reporting tool based on Wilkinson et al. 2021.

Type of bias	How bias was addressed or assessed
Selection bias at study entry	The study compared people enrolled in phase 2 trials assigned alectinib against patients from routine care in the US initiating ceritinib. Several steps were taken to minimise the risk of selection bias: Matching inclusion criteria in the real-world data to the population included in the trial Excluding additional patients from the trial with prior lines of therapy not observed in the real-world data Using real-world data over a similar time period to that covered in the trial Using a new-user, active comparator design To help demonstrate the validity of the approach, the comparison was repeated using only real-world data and similar results were found.
Selection bias at study exit	This was an as-started analysis with limited loss to follow up. Censoring is not thought to be informative.

Type of bias	How bias was addressed or assessed
Addressing confounding	Key prognostic variables were prospectively identified by a systematic review. Key known confounders were captured in the data albeit with limitations. See below for information on addressing missing data and misclassification of key confounders. Observed confounders measured at or before baseline were used to estimate propensity scores. Estimation used the inverse probability of treatment weights method. There was no evidence of large differences in covariate patterns between treatment groups after adjustment (standardised mean difference was less than 0.1 for all variables). In sensitivity analysis, adjustment for additional variables did not change results. Quantitative bias analysis was used to assess how strong a confounding effect an unknown confounder would need to have to eliminate the estimated treatment effect. The estimated e-value was 2.4 which would require a level of confounder-mortality and confounder-treatment association substantially higher than seen for any measured confounders.
Detection bias	The outcome of mortality was not thought to be subject to detection bias.
Measurement error and misclassification	Data on mortality is sufficiently well captured in the real-world data with sensitivity of 91% and specificity of 96%. There were concerns that central nervous system metastases were misclassified (underreported) in the real-world data due to limited surveillance. A sensitivity analysis found that the prevalence in the real-world data would have to be 40% larger to eliminate the estimated treatment effect.

Type of bias	How bias was addressed or assessed
Missing data	Missing data on baseline performance status (European Cooperative Oncology Group [ECOG] score) was high in the real-world data (32%) and this is a key prognostic variable. The main analysis assumed data was missing completely at random in a complete case analysis. Because this assumption was expected to be invalid, sensitivity analysis was performed using multiple imputation assuming data was missing at random. Results were consistent with the complete case analysis. Quantitative bias analysis was performed to address remaining concerns about missing not at random data, when ECOG scores are worse than expected by the imputation model. Using threshold analysis the study conclusions remained similar under any reasonable assumptions about the ECOG scores in those with missing values.
Reverse causation	Reverse causation is not expected to be a problem in this analysis.

Appendix 3 – Reporting information for selected analytical methods

Guide to reporting on selected analytical methods

Method	Description	Reporting information
Direct or indirect standardisation	Methods to increase comparability of exposure groups in terms of selected covariates	• Standard reference population (description) • Covariates used for standardisation
Stratification	Dividing the data into subsets, or strata for analysis	• Covariate definition of strata • Number of observations in each stratum • Descriptive statistics and results within each stratum
Matching	Matching individuals with the same or similar characteristics	• Variables used for matching • Matching algorithm • Matching caliper (if relevant) • Matching ratio • Number matched and number excluded

Method	Description	Reporting information
Propensity score (general)	Estimate of probability of receiving a particular intervention; range of methods available (below)	• Model used to estimate propensity scores (such as logistic or multinomial) • Covariates used and how they were included in the model • Propensity score distribution before and after adjustments (for example, pre- and post-matching) • N/% contributing to matched, trimmed, truncated or weighted analyses • Diagnostic checks for any statistical analysis done • See Tazare et al. 2022 for reporting of high-dimensional propensity score models
Propensity score (stratification)	Patients grouped into strata (for example, deciles) based on propensity score and stratum-specific effects aggregated	• How strata are defined • Trimming and whether applied before or after strata defined • Tables for stratified population characteristics

Method	Description	Reporting information
Propensity score (weighting)	Weights attached to individuals based on inverse of propensity scores	• How weights are calculated • Whether and how weights are trimmed, truncated or stabilised • Tables for unweighted and weighted population characteristics • Mean and distribution of weights
Propensity score (matching)	Matches individuals with similar propensity scores	• Matching algorithm used including caliper and scale • Matching ratio (such as fixed 1:1 or variable 1:5) • Tables for unmatched and matched population characteristics
Multivariable regression adjustment (includes using propensity scores)	Statistical models comparing outcomes as a function of the intervention and covariates	• Type of model (such as linear regression or Poisson) • Covariates used and how they were included • Diagnostic checks

Method	Description	Reporting information
Instrumental variable analysis	Exploits external variation in exposure across people or over time using an 'instrument'. An instrumental variable is associated with the intervention but is otherwise unrelated to the outcome.	• Type of model (such as 2-stage least squares) and diagnostic checks • Strength of association between instrument and intervention (for example, odds ratio, risk difference) • Theoretical justification that the instrument does not affect the outcome except through the intervention and that the instrument does not share any causes with the outcome • Tables with distribution of population characteristics across levels of the instrument and intervention • For binary outcomes, exposures and instruments, table of the frequencies of each combination of instrument, treatment, and outcome • See Swanson and Hernán 2013 for reporting by specific causal effects in instrumental variable analysis and their dependent assumptions (for example, monotonicity) • The results of falsification tests: see Labrecque and Swanson 2018 for specific

Method	Description	Reporting information
		examples

Method	Description	Reporting information
Interrupted time series	Individuals or groups are used as their own controls and observed over multiple time points. Effects are observed by comparing outcome trends in the time period before and after intervention.	• Type of model (such as segmented linear regression with ordinary least squares regression) • Study time period and time intervals • Pre-specification of point of intervention effect (for example, explanation needed if point of analysis is not point of intervention delivery) • Number of pre-intervention, post-intervention, and between-intervention data (time) points, and the data points contributing to forecasting • Table comparing participant characteristics and missing data across each group analysed (for example, before and after intervention and for defined subgroups) • Table and graph showing outcomes across time (that is, pre- and post-intervention trend) • Results of diagnostic checks (for example, for autocorrelation, stationarity, seasonality, model specification checks) and any

Method	Description	Reporting information
		adjustments made • Results of falsification tests (for example, the use of pseudo start periods before intervention delivery)

How the framework was developed

Background

We developed the framework by collating research and existing best-practice guidance from research or professional organisations and other regulatory or health technology assessment bodies.

We sought feedback on the framework during its development through a series of workshops and through an open public consultation.

We engaged with and received feedback from many stakeholders including:

- patients and patient organisations
- health charities
- healthcare professionals
- the pharmaceutical and medical technologies industries
- data controllers and contract research organisations
- academia
- international health technology assessment bodies
- UK health system partners
- NICE committee members.

We revised the framework based on the feedback we received.

We would like to thank everyone who took part in the development and review of the real-world evidence framework.

NICE development team

Seamus Kent, Lynne Kincaid, Manuj Sharma, Shaun Rowark, Stephen Duffield, Vandana Ayyar Gupta, Joanne Glossop, Pall Jonsson

Update information

Minor changes since publication

March 2024:

- We added information about external validity bias to the section on risk of bias.
- We added a passage describing approaches for sampling in new studies requiring primary data collection.
- We added a new section on assessing external validity bias and adjusting for differences between the study sample and the target population.

July 2023:

- We updated links to recent case studies in the section on use of real-world evidence in NICE guidance.
- We added a link to the HARmonized Protocol Template to Enhance Reproducibility (HARPER) tool for protocol design in the section on conduct of quantitative real-world evidence studies.
- We added clarifying information on:
  - searching for fit-for-purpose data in the section on conduct of quantitative real-world evidence studies
  - federated data networks and how they relate to common data models in the section on assessing data suitability
  - describing the uses of instrument-based approaches and quasi-experimental studies and a correction around measurement error for risk ratios and rate ratios in the section on methods for real-world studies of comparative effects.
- We updated links to NICE Decision Support Unit reports.
- We updated the link to the hospital episode statistics GDPR webpage in appendix 1.
- We updated appendix 3 to provide additional information on reporting quasi-

experimental studies.

We made minor changes to style and language throughout without changing the meaning.

ISBN: 978-1-4731-4640-2