CONFIDENTIAL UNTIL PUBLISHED

MRI fusion biopsy in people with suspected prostate cancer

Addendum to the External Assessment Group's Report

Produced by Centre for Reviews and Dissemination (CRD) and Centre for Health Economics (CHE)

Date 18/11/22

The evidence synthesis results presented in the External Assessment Group report (EAR), and used in the cost-effectiveness analyses, must be interpreted with caution due to the high risk of bias in the evidence base and wide uncertainty in the results. The Diagnostic Assessment Committee (DAC) will likely need to consider how the uncertainties and limitations of the clinical evidence may impact on the cost-effectiveness of software fusion compared to cognitive fusion.

This document is an addendum to the EAR. In this document: 1) we provide additional interpretation of the results of the evidence synthesis to clarify the consistency between the approaches used and, 2) given the complexity of the treatment choices informed by MRI influenced biopsies, we will present disaggregated cost-effectiveness results by true Cambridge Prognostic Group (CPG) status, so that the effective net health contribution of each group (given the evidence synthesis results) for overall cost-effectiveness is known. While the DAC is likely to be considering the plausibility of the estimated effects by CPG, this addendum will clarify their likely impact on cost-effectiveness.

1 SUMMARY

The clinical evidence presented in the EAR (section 4.4.2) suggests that, compared to software fusion biopsy, cognitive fusion biopsy may be associated with a higher probability of being classified as not having cancer, and a lower probability of being classified at higher ISUP grades or CPGs (CPG used as interchangeable with ISUP grade henceforth), suggesting that software fusion may be better at detecting higher grade cancers. Similar results were obtained when adding a systematic component to software and cognitive fusion biopsies.

Key issues to consider include the sparsity of evidence, particularly of comparative diagnostic accuracy at CPG >2, potential biases, and the lack of ability to make comparisons across all software fusion devices in the scope. In addition, the lack of a gold standard comparison and a lack of evidence on prevalence of prostate cancer by CPG across the population of interest, hinder interpretation and confidence in the results.

The key findings on the economic value of software fusion compared to cognitive fusion are:

1. The costs (and harms) of software fusion biopsy in the diagnostic model component can only be offset in the long-term model component, which will only arise from differences in diagnostic accuracy between software and cognitive fusion.
2. The value gains for software fusion appear to stem from increased detection at CPG ≥2 and, once we adjust for prevalence by CPG category, the greatest contribution to the cost-effectiveness of software fusion compared to cognitive results from increased correct detection at CPG 2.
3. Increased detection at CPG 1 due to reduced detection of 'no cancer' results in value losses at all cancer grades (i.e., there are net losses from shifting classification from 'no cancer' to clinically non-significant cancer (CPG 1)).
4. The magnitude of value realised for software fusion vs. cognitive fusion from the balance between different degrees of misclassification and correct classification with the two technologies also depends on the prevalence at each cancer grade.

Therefore, the value of software fusion is driven by i) comparative diagnostic accuracy derived where evidence is particularly sparse (cancer grades above 2), and by prevalence, which is also affected by evidence sparsity.

In light of this, judgements on the economic value of software fusion, require integration of the uncertainties over the clinical evidence with the overall cost-effectiveness.

We provide more detailed commentary in the two subsequent sections to support our findings. In section 3.3 we provide a framework to help with the translation of different judgements on the clinical evidence into impacts on the cost-effectiveness of software fusion.

2 INTERPRETATION OF THE EVIDENCE SYNTHESIS RESULTS

Two types of model were fitted to the diagnostic accuracy data: a multinomial model which estimated the differential ORs for classification in each of 4 cancer grade categories (CPG 1, 2, 3, or 4-5) compared to the reference category 'no cancer' for software fusion compared to cognitive fusion – Model 1a; and models estimating the ORs of classifying an individual as having any cancer (i.e. CPG 1 to 5 versus no cancer – Model 2a) or clinically significant cancer (CPG 2 to 5 versus no cancer or CPG 1 – Model 3a) for software fusion compared to cognitive fusion. The models use essentially the same data (apart from studies not reporting clinically significant cancer for Model 3a), and retrieve concordant results, but because the interpretation of the odds ratios differs across the two modelling approaches, careful consideration is required.

The multinomial model (Model 1a, EAR Table 10) suggests that compared to software fusion biopsy, patients undergoing cognitive fusion biopsy may show:

i) a higher probability of being classified as not having cancer,

ii) similar probability of being classified as having non-clinically significant cancer (CPG 1), and

iii) lower probability of being classified at higher CPGs, particularly CPG 2.

Model 2a suggests software fusion biopsy classifies more patients as having cancer (any CPG), than cognitive fusion biopsy (OR 1.30 95% CrI 1.06, 1.61; EAR Figure 5), thus agreeing with point i) above. This does not contradict point ii), since Model 2a is comparing detection of any cancer, i.e., all CPG 1 to 5 combined, and not only the detection of non-clinically significant cancer. The increase in the ORs for detection of any cancer is driven by the increase in the probability of categorisation at CPG > 1, which in this case is driven by increases at CPG 2 as noted in point iii).

We illustrate this point using data from the PAIREDCAP study which compared cognitive fusion biopsy to software fusion biopsy (EAR Table 8). The probabilities of being classified as having no cancer, non-clinically significant cancer (CPG 1) or clinically significant cancer (CPG 2-5) are summarised in Table 1. This shows that the probabilities of being classified in CPG 1 are very similar whereas software fusion classifies a greater proportion of patients in CPG grades 2 and above. This leads to an OR for detection of any cancer vs. no cancer that is large (OR 1.52 (95%CI 1.04, 2.22) with a 95% CI not including the null effect but to a smaller OR (1.34 (0.94, 1.90)) (with 95% CI including the null effect) for classifying patients as clinically significant vs. clinically non-significant or no cancer.

Table 1 Example of different absolute probabilities and odds ratios depending on the type of comparison, observed in the PAIREDCAP study.

PAIREDCAP	Cognitive	Software
Probabilities observed in study		
no cancer	0.38	0.29
CPG 1	0.15	0.17
CPG 2-5	0.47	0.54
Probabilities: any cancer vs no cancer		
no cancer	0.38	0.29
any cancer (CPG 1-5)	0.62	0.71
OR (95% CI)	1.52 (1.04, 2.22)	
Probabilities: clinically significant cancer vs non clinically significant cancer		
No cancer or CPG 1	0.53	0.46
CPG 2-5	0.47	0.54
OR (95% CI)	1.34 (0.94, 1.90)	

While Model 2a suggests software fusion biopsy classifies more patients as having cancer than cognitive fusion biopsy (OR 1.30 95% CrI 1.06, 1.61; EAR Figure 5), Model 3a also suggests software fusion biopsy classifies more patients as having clinically significant cancer than cognitive fusion biopsy, but with a wider confidence interval that includes the null effect (OR 1.35 95% CrI 0.86, 2.10; EAR Figure 5). The example above using PAIREDCAP illustrates that while the increased ORs are driven by increases in the probability of detecting CPG 2-5 (and not by increases in detection of CPG 1), by using odds and collapsing CPGs, the statistical model has higher power to detect statistically significant differences against no cancer than against non-clinically significant cancer (which pools no cancer with CPG 1). These results are consistent with findings i) to iii) above.

In conclusion, the results of the synthesis models require careful interpretation as they refer to comparisons between different cancer grades. The interpretation of the multinomial models on the absolute probability scale results is more intuitive and directly relevant to clinical practice. Overall, results are concordant across analyses and concordant with the data. Only the multinomial results are used in the economic model as the value of diagnostic information provided by each test is dependent on the subsequent clinical decisions based on test results, and clinical management is conditional on cancer grades (jointly with other prognostic information).

3 COST-EFFECTIVENESS DRIVERS

This section reports further results from the base-case model on the comparison of targeted strategies (see EAR, Section 6.4.6.), that aim to aid the DAC's decision making when trying to integrate the uncertainties over the clinical evidence with the overall cost-effectiveness.

This is important because the classification of suspected prostate cancer is complex, resulting in 15 final combinations of diagnosed category and true disease status in the decision model. The complexity of the classification of disease, treatment allocation rules (comprising of different distributions of active surveillance, radiotherapy, and radical prostatectomy in each diagnosed category), combined with the impacts of the different treatments, makes it difficult to establish how the misclassification of suspected prostate cancer lesions across different categories drives the value of software fusion compared to cognitive fusion. To clarify this issue we present: the final diagnostic accuracy for the strategies as implied by the test sequences modelled (section 3.1), the disaggregated cost-effectiveness results (corresponding to aggregated results Section 6.4.6. of EAR) by true CPG (section 3.2), and the trade-offs between different degrees of misclassification and correct classification (section 3.3). The results presented do not reflect parameter uncertainty.

3.1 Diagnostic accuracy of the test sequences in the decision model

Table 2 illustrates the distribution of test results and conditional accuracy probabilities at final classification for cognitive and software fusion biopsies, which includes first biopsy (see EAR Table 35 for corresponding data for first biopsy), and repeat biopsy for a proportion of individuals. The difference in diagnostic accuracy between software fusion and cognitive fusion by each classification is shown in brackets in Table 2 (green for increased detection and red for reduced detection).

Table 2 Final classification: Distribution of test results, conditional diagnostic accuracy and prevalence probabilities

		distribution test results					distribution of test results				
		50.5%	18.3%	14.4%	10.2%	6.6%	44.6%	17.2%	20.3%	11.1%	6.7%
		cognitive fusion biopsy					**software fusion biopsy**				
		accuracy matrix					accuracy matrix				
prev	**CPG**	**No cancer**	**1**	**2**	**3**	**4 or 5**	**No cancer**	**1**	**2**	**3**	**4 or 5**
12%	**No cancer**	100%					100% (0%)				
32%	**1**	82%	18%				66% (-16%)	34% (+16%)			
26%	**2**	29%	35%	36%			24% (-5%)	19% (-16%)	57% (+21%)		
18%	**3**	18%	13%	20%	49%		22% (+3%)	5% (-7%)	21% (1%)	52% (+3%)	
12%	**4 or 5**	12%	10%	11%	10%	56%	11% (-1%)	4% (-6%)	13% (+2%)	14% (+4%)	58% (+2%)

prev, prevalence. Results do not consider biopsy related mortality.

The diagnostic accuracy at final classification is consistent with the results for the first biopsy for both strategies (targeted cognitive fusion and software fusion), suggesting that software fusion increases the correct classification across all CPGs (cells along the diagonal line) compared to cognitive fusion, particularly for CPG 2 (21% more) and CPG 1 (16% more). For those with true CPG 2, the greatest reduction in misclassification is observed at diagnosed CPG 1 (16% less).

Given these findings, we expect results in the cost-effectiveness model to be mainly driven by the trade-offs associated with misclassification at CPG 1 and CPG 2.

3.2 Estimation of disaggregated base-case results

Table 3 presents the base-case analysis incremental NHB (INHB) at £20,000 per additional QALY of software fusion compared to cognitive fusion as a total estimate (0.01 QALY [0.00810 with further decimal cases], as in EAR, table 54) and disaggregated by true disease category (where the prevalence for each true disease category is set to 100%). The total INHB corresponds to the sum of the INHB by true disease category weighted by its corresponding prevalence.

Table 3 Base-case analysis results by CPG category

CPG	Prevalence (weights)	INHB by CPG	INHB by CPG x prevalence*
No cancer	12.1%	-0.00500	-0.00061
1	31.8%	-0.01631	-0.00519
2	26.2%	0.02890	0.00757
3	18.3%	0.01907	0.00349
4 or 5	11.6%	0.02435	0.00283
Total INHB*			0.00810

*estimates weighed by the prevalence for each true disease category

The results suggest:

- The disaggregated INHB estimates are negative for the 'no cancer' and CPG 1 categories, which suggests that the increased correct detection of CPG 1 with software fusion does not result in net health gains in relation to cognitive fusion.
- The disaggregated INHB estimates suggest higher net health gains for software fusion compared to cognitive fusion for CPG ≥2. Once prevalence is considered (column 3), the largest effective contribution to the total INHB arises from CPG 2.

To aid interpretation of these results, in Table 4 further disaggregates the base-case results by model component, diagnostic or long-term, and within the long-term model component by health-state (localised or metastatic). As in Table 3, the grey shading highlights estimates unweighted by prevalence, whereas the totals correspond to prevalence weighted values.

Table 4 Base-case analysis results by CPG category for the diagnostic and long-term model results

	Diagnostic model	Long-term model				
	INHB	INHB	Inc	Inc	INHB	INHB
CPG	Total	Total	QALYs	Costs	Localised	Metastatic+ EoL
No cancer	-0.0050	-	-	-	-	-
1	-0.0056	-0.0107	-0.0017	£180	-0.0075	-0.0032
2	-0.0043	0.0332	0.0198	-£268	0.0636	-0.0304
3	-0.0046	0.0237	0.0178	-£117	0.0327	-0.0090
4 or 5	-0.0046	0.0289	0.0290	£2	0.0392	-0.0103
Total*	-0.0049	0.0130	0.0113	-£34	0.0248	-0.0118

*estimates weighed by the prevalence for each true disease category; EoL, end of life; Inc, incremental.

INHB for the localised disease health states included costs and disutilities of localised disease monitoring, treatment, and associated adverse events. INHB for the metastatic disease health states included costs and disutilities of metastatic disease monitoring, treatment, and associated adverse events. For simplicity, end of life costs were also included in the metastatic INHB.

In the diagnostic model, the INHB of software fusion is negative across all CPGs. Since the INHB of software fusion in the model overall (diagnostic + long-term) is positive, this suggests that the costs and harms of software fusion in the diagnostic model are only offset by long-term costs and health-related quality of life (HRQoL) outcomes that result from the subsequent clinical management of individual conditional on final biopsy classification. The different diagnostic INHB for each category reflect only differences in the proportion of repeat biopsies across true disease categories (which is conditional on the diagnosed category at first biopsy).

The long-term model INHB estimates follow the same pattern across true disease categories as observed for the full model results (Table 3). For true disease categories CPG 2 and above, the INHB for software fusion vs. cognitive fusion is positive; the greater contribution to the INHB stems from the CPG 2 category. Compared to cognitive fusion, QALY gains occur for CPG 2 and above with software fusion, and these are accompanied by cost savings for CPG 2 and 3.

The INHB for CPG 1 is negative in the long-term model due to higher costs and lower QALYs compared to cognitive fusion; this is due to the increased correct detection of CPG 1 leading to more individuals receiving immediate (conservative or radical) treatment of localised disease with associated costs and adverse events (if they had been misclassified as no cancer they would have received only monitoring in the first two years in the model) for software fusion compared to cognitive fusion, which are not offset by the benefits of early treatment. The annual probability of progression from localised to metastatic disease is similar for CPG 1 misclassified compared to correctly classified CPG 1 (0.14 vs 0.13), so the benefits from increased correct detection at this category are limited.

The localised and metastatic INHB estimates also suggest that the increased correct detection in category CPG 2 with software fusion is contributing more to the total long-term model INHB. We note that while the metastatic INHB is negative across all cancer categories, this does not mean that there are higher net health losses with software fusion compared to cognitive fusion in the metastatic health states because the INHBs are not estimated by individual in the model. Since individuals spend less time in the metastatic health states with software fusion compared to cognitive fusion due to overall slower progression to metastatic disease with software fusion (e.g., for CPG 2, 3.55 and 3.70 undiscounted life-years are accrued in the metastatic health states for software fusion and cognitive fusion, respectively), software fusion accrues overall fewer QALYs than cognitive fusion in the metastatic health states. Despite the lower costs accrued with software fusion in the metastatic states, the metastatic INHB is always negative.

In the next section, we examine the absolute NHB by final classification in the model to understand how shifts in classification may impact on cost-effectiveness estimates.

3.3 Disaggregated estimates of cost-effectiveness by final classification category

We estimated the NHB that could be achieved in the long-term model if all individuals were identified in a particular final classification category; Table 5 report the results for the 15 possible classifications. Results in Table 5 are not specific to software fusion or cognitive fusion.

Table 5 Long-term model NHB at each final classification category

CPG	**No cancer**	**1**	**2**	**3**	**4 or 5**
No cancer	8.966 (9.435,-0.469)				
1	7.996 (8.121,-0.125)	7.930 (8.074,-0.145)			
2	7.072 (6.756,0.316)	6.855 (6.578,0.277)	7.066 (6.927,0.139)		
3	5.215 (4.077,1.139)	5.117 (4.027,1.090)	5.468 (4.571,0.897)	5.707 (4.925,0.782)	
4 or 5	3.476 (1.740,1.737)	3.408 (1.689,1.719)	3.642 (2.007,1.635)	3.816 (2.245,1.571)	3.912 (2.385,1.527)

The NHB by health state are reported between brackets (localised disease NHB, metastatic disease NHB).

Results suggest that there will be more (long-term) NHB loss in misclassifying CPG ≥2 as CPG 1 than as 'no cancer'. The highest increase in NHB for CPG ≥2 can be achieved with technologies that shift misclassification from CPG 1 to the correct classification. When shifting between adjacent categories, the highest NHB gain can be generated when someone with CPG 3 misclassified as CPG 1 with one technology is identified as CPG 3 with the alternative (+0.351 QALYs). The lowest NHB gain between adjacent categories is generated for those with true CPG 4-5, when they 'move' from a CPG 3 to a correct diagnosis (+0.096 QALY).

The incremental value of one technology will depend on how it changes the distribution across classification categories for each true disease category compared to the alternative technology, and on the prevalence per true disease category. For software fusion compared to cognitive fusion, the INHB will be positive for the classification categories where it increases detection and negative for those where detection is decreased (see Table 2 for differences between diagnostic accuracy matrices).

This can be illustrated with an example for true disease category CPG 2. For software fusion vs. cognitive fusion:

- the reduction in detection of CPG 2 as 'no cancer', the category with highest NHB for CPG 2, (7.072 QALYs) is small (-5%), so the INHB is -0.339 QALYs;
- the reduction in detection of CPG 2 as CPG 1 (NHB = 6.885 QALYs) is -16% resulting in an INHB of -1.086 QALYs;
- the increased correct detection of CPG 2 (+21%, NHB = 7.066 QALYs) is sufficient to offset the negative INHBs from the alternative classifications (as 'no cancer' and CPG 1), and the total INHB across this disease category is 0.033 QALYs.
- Since the prevalence of CPG 2 is 26% the relative contribution from changes in detection rates for this category is 0.009.

As highlighted at the beginning of this addendum, the uncertainties and limitations of the clinical evidence reduce the robustness of the evidence synthesis results. The same uncertainties also impact on the cost-effectiveness results. The information in Tables 3, 4 and 5 can be used by the DAC to consider how their judgements on what are the plausible differences in the prevalence and relative diagnostic accuracy between software fusion and cognitive fusion (particularly above CPG 2, where disaggregated evidence is sparser) can be translated into cost-effectiveness impacts.