

ABSTRACT (max. 240 words)

Objectives

To investigate the importance of accounting for potential performance dependency when evaluating the cost-effectiveness of two diagnostic tests used in combination.

Methods

Two meta-analysis models were fitted to estimate the diagnostic accuracy of Wells score and Ddimer in combination. The first model assumes that the two tests perform independently of one another, thus two separate meta-analyses were fitted to the Ddimer and Wells score data and then combined. The second model allows for any performance dependency of the two tests by incorporating published data on the accuracy of Ddimer stratified by Wells score, as well as studies of Ddimer alone and Wells score alone. The results from the two meta-analysis models were input into a decision model to assess the impact that assumptions regarding performance dependency have on the overall cost-effectiveness of the tests.

Results

The results highlight the importance of accounting for potential performance dependency when evaluating the cost-effectiveness of diagnostic tests used in combination. In our example, assuming the diagnostic performance of the two tests to be independent resulted in the strategy 'Wells score moderate/high risk treated for DVT and Wells score low risk tested further with Ddimer' being identified as the most cost-effective at the £20,000 willingness-to-pay threshold (probability cost-effective 0.8). However, when performance dependency is modelled the most cost-effective strategies were 'Ddimer alone' and 'Wells score low/moderate risk discharged and Wells score high risk further tested with Ddimer' (probability cost-effective 0.4).

Conclusions

When evaluating the effectiveness and cost-effectiveness of diagnostic tests used in combination, failure to account for diagnostic performance dependency may lead to erroneous results, and non-optimal decision-making.

Introduction

In the area of diagnostic test performance, evidence-based evaluations are crucial to the decision making process as early diagnosis can lead to diseases being treated more successfully than if treatment were delayed. Often evaluations are performed focusing on the accuracy of a single test to diagnose a particular condition(1), however, in routine clinical practice a diagnosis is usually based on the results obtained from multiple tests.

A recent review of NIHR Health Technology Assessment reports of decision models for diagnostic tests containing meta-analysis results from 1997 to 2009(1) found that 6 of the 14 (43%) reports included in the review considered a combination of diagnostic tests strategy in the economic decision modelling part of the report. In these 6 reports the accuracy of each combination of diagnostic tests was calculated either assuming i) conditional independence between tests or ii) the accuracy of the second test to be perfect (which may be reasonable to assume in some contexts). However, where multiple tests are used for diagnosis it is highly likely that the tests will not perform independently (that is, in the case of 2 tests, the performance of the second test may differ depending on the results of the first test) and therefore it is important to allow for this in the analysis. In fact, there is evidence that when the assumption of dependence between tests is ignored, then this may lead to erroneous disease probability estimates(2) which, if input into an economic decision model, will carry forward into the cost-effectiveness analysis results.

In this paper we investigate the importance of allowing for potential performance dependency when evaluating the cost-effectiveness of two diagnostic tests used in combination (which uses recent advances in meta-analysis methodology outlined in our companion paper(3)). This is assessed by observing the impact on the cost-effectiveness results, and subsequent conclusions reached, when performance dependency is first ignored and then incorporated. The paper focuses on the example of diagnosing deep vein thrombosis (DVT) using Wells score and Ddimer.

Motivating example: Ddimer and Wells score tests for the diagnosis of deep vein thrombosis

DVT is a blood clot in a deep vein (lower limb) that is usually treated with anticoagulants. An accurate diagnosis of DVT is crucial in order to lower mortality due to Venous Thromboembolism related adverse events and also to reduce the impact of side effects from anticoagulant treatment given to patients wrongly diagnosed with DVT.

DVT may be diagnosed by reference tests such as Ultrasound and Venography which have high diagnostic accuracy but such tests are expensive to perform. Therefore, cheaper, quicker but less accurate tests are often used for diagnosis of DVT. Two of these tests, which will be considered in this paper, are Ddimer (i.e. measures the concentration of an enzyme in the blood, the higher the measurement the more likely DVT) and Wells score (devised from an assessment of the clinical features of DVT such as clinical history, symptoms and signs(4, 5)). For use in diagnosis the latter test is usually categorised into *low* (score <1), *moderate* (score 1 or 2) and *high* (score >2) risk of having DVT. For more details about the diagnostic performance data of Ddimer, Wells score and Ddimer given Wells score used in the analysis see Novielli et al.(3).

In a recent review, Goodacre et al(6) found Ddimer and Wells score to not be accurate enough as stand-alone diagnostic tools but that algorithms containing both Wells score and Ddimer were potentially valuable for diagnosis.

Methods

Diagnostic accuracy meta-analysis models

To investigate the impact the assumption of test performance dependency has on the cost-effectiveness results, two statistical analyses were undertaken to obtain the joint diagnostic accuracy for Wells score and Ddimer when used in combination.

The first analysis assumed the diagnostic performance of the two tests to be independent and therefore used the results from two separate meta-analyses to inform the cost-effectiveness decision model. As diagnostic accuracy is usually measured in terms of both sensitivity (i.e. the proportion of actual positives which are correctly identified) and

specificity (i.e. the proportion of negatives which are correctly identified), bivariate meta-analyses(7) (which allow for the between-study correlation of sensitivity and specificity potentially induced through varying test thresholds used in the different studies) were fitted to the Wells score data, and the Ddimer data separately.

The second analysis used the meta-analytic modelling framework developed by Novielli et al.(3), to account for test performance dependency in the estimation of the diagnostic accuracy of Wells score and Ddimer used in combination. This analysis used a multi-component meta-analysis framework(8) to incorporate data from studies reporting the accuracy of Ddimer stratified by Wells score, as well as studies of Ddimer alone and Wells score alone. Random effects models were used with different likelihoods required for the different data types but linked together through the use of shared parameters(9, 10). For more details about the data (together with references) and the analysis, see Novielli et al.(3).

Decision model

The comprehensive cost-effectiveness decision model (i.e. integrating the meta-analysis and decision model into a single coherent framework(11)) for evaluating a single diagnostic test used by Sutton et al.(12) (adapted from Goodacre et al(6)) was modified to allow for the incorporation of two tests in combination(Figure 1). This decision model assumed a simplified diagnosis-to-treatment pathway for DVT whereby patients who were diagnosed as positive (based on one of the strategies defined in the next section) were treated with anticoagulants which potentially may cause harmful side effects such as bleeding at different intensities (i.e. false and true positive patients may be subject to non-fatal bleeding, fatal intracranial bleeding, non-fatal intracranial bleeding or no bleeding when treated with anticoagulant). The accuracy parameters (i.e. false positive, false negative, true positive, true negative for Wells score and Ddimer) were informed by the meta-analysis models discussed above. All other model parameter values and sources (i.e. prevalence of DVT, risk of pulmonary embolism, quality of life adjusted life years per each possible health status, costs, etc) are reported in Goodacre et al.(6).

Strategies

As mentioned previously, for use in diagnosis Wells score is usually categorised into *low* (score <1), *moderate* (score 1 or 2) and *high* (score >2) risk of having DVT. For the analyses presented in this paper 3 different classifications of Wells score were used; i.e. i) WS_1 - low (score < 1) and moderate / high (score ≥ 1), ii) WS_2 - low / moderate (score ≤ 2) and high (score > 2) or iii) WS_3 – low (score <1), moderate ($1 \leq \text{score} \leq 2$) and high (score > 2).

In our cost-effectiveness analyses 10 diagnostic strategies were considered as outlined in Table 1. For each dichotomy of two diagnostic tests, two possible strategies can be defined(13): 1) ‘believe the negatives’ - only patients diagnosed as positive by the first test received the second test (i.e. $(WS_1 \& DD)_{BN}$ and $(WS_2 \& DD)_{BN}$)- and 2) ‘believe the positives’ - only patients diagnosed as negative by the first test received the second test (i.e. $(WS_1 \& DD)_{BP}$ and $(WS_2 \& DD)_{BP}$).

Note that for every pair of dichotomous (or dichotomised) tests combined according to one of the strategies “believe the negative” or “believe the positive”, the order of the tests does not affect the diagnostic accuracy of the strategy(3) (though which test is conditioned on may affect the estimation of effectiveness parameters in the synthesis model) but may affect the costs incurred. For example, Wells score is usually given first (i.e. to everyone with suspected DVT) because it is less expensive than Ddimer and does not require any specialist technology, and can be carried out by an experienced doctor quickly at initial presentation. Therefore, given that Wells score is less expensive than Ddimer, any sequence of the two tests (i.e. diagnostic strategies 1, 2, 3 and 4 listed above) where Wells score is dichotomised and given first will be dominant from an economic point of view compared to the same equally accurate strategy where Ddimer is given first. Note that for strategy number 8, where the first test, Wells score, is not dichotomised, this property does not hold.

Modelling framework

All analyses were conducted using a comprehensive decision modelling framework(11) which evaluates both the evidence synthesis models and decision model within a single coherent framework. The modelling framework was implemented using Markov Chain Monte Carlo (MCMC) simulation in WinBUGS software(14). Non-informative (vague) prior

distributions were used for all parameters estimated by the statistical model. Graphical tools were used to assess convergence of the MCMC chains and sensitivity analyses were performed to assess the influence of the initial values and prior distributions on the results. The WinBUGS code (including the specific prior distributions used) is provided in the Appendix.

Results

Meta-analyses

Table 2 presents the results from the 2 different meta-analysis models (i.e. independent and dependent). The results are presented in terms of the sensitivity and specificity. A good diagnostic strategy, in terms of diagnostic accuracy, is defined as one where the sensitivity is high and specificity is high. In this case, the strategies with the highest sensitivity and specificity are strategy 1) (i.e. Wells score WS_1 followed by Ddimer, 'believe the negatives') and strategy 8) (i.e. Wells score WS_3) regardless of the meta-analysis model(s) used. Note that strategies 9) and 10), where no diagnostic test is undertaken, result in zero specificity and zero sensitivity respectively.

The modelling assumptions of performance independence and dependence of tests only affect the strategies which contain both Wells score and Ddimer used in combination (i.e. strategies 1) to 4)). Comparing the sensitivity estimates for these strategies, it can be observed that different estimates are obtained for the two different synthesis models considered (i.e. independent or dependent), except for strategy 2 where the mean estimates are identical. Similar results are observed for specificity, except for strategy 3 where the credible intervals for the mean estimates obtained from the two different synthesis models overlap. For both sensitivity and specificity it can also be observed that the uncertainty around the estimates is greater when the performance dependence of tests is assumed. The impact these differences in sensitivity and specificity obtained from the two different synthesis models, and their associated uncertainty, had on the cost-effectiveness results is reported below.

Cost-effectiveness

The accuracy estimates reported in Table 2 were input into the decision model to evaluate the cost-effectiveness of the 10 different diagnostic strategies outlined above. Figure 2 presents the cost-effectiveness acceptability curves (CEACs) for each diagnostic strategy using the independent and the dependent meta-analysis results. When the test performance of Wells score and Ddimer were assumed to be independent, strategy 1) (i.e. Wells score WS_1 followed by Ddimer 'believe the negatives') was identified as the most cost-effective at the £20,000 willingness-to-pay threshold (probability cost-effective 0.8) (Figure 2 panel (a)). When this assumption was relaxed, and diagnostic performance was assumed to be dependent between the two tests, the most cost-effective strategies at the £20,000 threshold were strategy 5 (i.e. Ddimer alone) and strategy 3 (i.e. Wells score WS_2 followed by Ddimer 'believe the negatives') with a probability of 0.4 (Figure 2, panel (b)). Thus, in this example, the assumption of performance dependence altered the cost-effectiveness results and the conclusions drawn.

The final panel (c) in Figure 2 presents a further set of CEACs. These are, in fact, generated using the results from the dependent meta-analysis (as for panel (b)) but here estimation of test performance is based on the predictive posterior distribution rather than the posterior distribution for the mean effect. Although currently this approach is not routinely taken, there are strong arguments, in some contexts, for the use of predictive distributions when the target population can be viewed as a further 'sample' from the distribution of potential patient populations – see Ades et al for a full consideration of this topic(15). In this example, using the predictive distribution did not change the overall rankings of the front-running strategies but the extra variability induced in the modelling results in a diminished certainty in which strategies are considered cost-effective. For example, when assuming diagnostic performance dependency between the two tests and using the distribution of mean effects, at the £20,000 threshold, Ddimer was estimated to be most cost-effective with a probability of 0.4; this diminished to 0.3 when predictive distributions of test performance were used instead.

Discussion and conclusions

Traditionally, as for primary studies, systematic reviews and meta-analysis of diagnostic test accuracy studies have focused on the performance of individual tests despite the use of

multiple tests being commonplace in routine clinical practice. A recent review(1) showed that, where the performance accuracy and cost-effectiveness of multiple diagnostic tests had been evaluated in NIHR health technology assessments, the strong (and we suspect often incorrect) assumption of independence between tests was usually made. This is likely, at least in part, to be due to such evaluations being informed by meta-analyses of individual tests and also due to fewer studies reporting multiple tests in the literature. However, our companion paper(3) outlines a methodology for combining disparate evidence in order to evaluate sequences of tests while taking account of any correlation in the performance of the tests. In this paper we have shown the importance of allowing for test performance dependency (utilising the aforementioned methodology(3)) when evaluating the diagnostic accuracy, and subsequent cost-effectiveness, of tests used in combination. The importance of modelling test dependency is brought home in the DVT example considered here where the conclusions regarding the most cost-effective strategy for diagnosing DVT changed compared to the analysis when any such dependency is ignored. We appreciate that no data at all may be available on dependency between tests in some contexts but we do not consider that situation in the present paper. However, we do believe steps should be taken to model dependency even in the absence of data since assuming independence is almost certainly incorrect. This could take the form of a sensitivity analysis exploring the impact of specifying different values for the correlation or placing a probability distribution on the parameter derived from related data (i.e. from other diagnostic tests) and/or expert opinion. The latter would be necessary if one wished to establish the value of conducting studies estimating the degree of dependency via a value of information analysis(16).

For detailed discussions of the limitations of the primary study data we refer the reader to our companion paper(3). Issues raised there include heterogeneity in study results and known variable study quality, which were not accounted for in the modelling but potentially could have been addressed through the use of study level covariates. There are also limitations in the economic modelling including the restriction to the evaluation of a maximum of two tests in any one clinical pathway. Although in principle both the synthesis and decision modelling could be extended to combinations of 3 or more tests we leave an implementation of this as a topic for further work. Here, the number of parameters in the evidence synthesis modelling would increase sharply, and we suspect there would usually

be a greater paucity of direct evidence evaluating 3 or more tests on the same population. This may actually lead to a necessary rethinking of how effectiveness data on sequences of tests is obtained; for example, would a single database of individual patient data on which all relevant tests were used on a cohort of patients contain more information than a formal systematic review of the whole published literature (in which no data comparing all tests is available)? Another issue is that we only considered the use of the Ddimer test at the (mean) threshold reported in the literature; that is to say, consideration was not given to alternative operating loci along the summary receiver operating characteristic (sROC) curve estimated by the synthesis model. However, further modelling could be carried out to identify the loci on the sROC curve which maximises cost-effectiveness as we have described elsewhere(12).

One assumption of the cost-effective evaluation presented in this paper is that the diagnostic strategies that were evaluated were “definitive”; that is, they represented an exhaustive range of diagnostic test strategies to DVT, where the diagnosis leads either to discharge (when negative diagnostic result) or treatment (when positive diagnostic result), and there is no further testing before or after each strategy. This was a simplification of the problem of diagnosing DVT in order to evaluate the effect of considering or ignoring dependency between tests when performing cost-effective evaluations. It can also be considered that there are more tests that may be combined in a strategy for DVT (e.g. ultrasound, Venography).

In our companion paper(3) we noted that the studies evaluating single tests had little impact on the estimates of test performance and that this had important implications for those designing and conducting studies of test accuracy in the future. The impact of this observation could be sharpened and further quantified using value of information methods(17) in an economic framework such as the one presented. For example, such modelling will identify the strategies which are potentially most cost-effective and uncertainty in the effectiveness of such strategies could be reduced by the conduct of randomised controlled trials evaluating them. Which strategies and the sample size of trial arms could both be determined using value of information methods within the modelling framework presented here.

In conclusion, reliable clinical and economic evaluations of appropriate/optimal use of diagnostic test pathways are challenging. Data are often far from ideal (due to studies typically only considering the performance of individual tests rather than multiple tests often requirement for evaluating full diagnostic pathways), which, as we have illustrated, further complicates modelling. However, we believe models which allow for test dependency should routinely be used since ignoring dependency will be naïve in many contexts.

References

1. Novielli N, Cooper NJ, Sutton AJ, Abrams KR. How is evidence on test performance synthesised in economic decision models of diagnostic tests? A systematic appraisal of Health Technology Assessments in the UK since 1997. *Value in Health* 2010;13(8):952 -7.
2. van Walraven C, Austin PC, Jennings A, Forster AJ. Correlation between serial tests made disease probability estimates erroneous. . *Journal of Clinical Epidemiology*. 2009;62(12):1301-5.
3. Novielli N, Cooper NJ, Sutton AJ. Meta-analysis of the accuracy of combinations of two diagnostic tests used in combination: Application to Ddimer and Wells score for the diagnosis of Deep Vein Thrombosis Value in Health. Submitted.
4. Wells PS, Anderson DR, Bormanis J, Fred G, Mitchell M, Gray L, et al. Application of a diagnostic clinical model for the management of hospitalized patients with suspected deep-vein thrombosis. *Thrombosis and Haemostasis*. 1999;81(4):493-7.
5. Wells PS, Hirsh J, Anderson DR, Lensing AWA, Foster G, Kearon C, et al. Accuracy of clinical assessment of deep-vein thrombosis. *Lancet*. 1995;345(8961):1326-30.
6. Goodacre S, Sampson F, Stevenson M, Wailoo A, Sutton A, Thomas S, et al. Measurement of the clinical and cost-effectiveness of non-invasive diagnostic testing strategies for deep vein thrombosis Health Technology Assessment. 2006 10(15):1-168.
7. Arends LR, Hamza TH, van Houwelingen HC, Heijenbrok-Kal MH, Hunink MGM, Stijnen T. Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making*. 2008;28:621-38.
8. Dias S, Sutton AJ, Ades AE, Welton NJ. A Generalized Linear Modeling Framework for Pairwise and Network Meta-analysis of Randomized Controlled Trials. *Medical Decision Making*. In press.
9. Ades AE, Cliffe S. Markov Chain Monte Carlo Estimation of a Multiparameter Decision Model: Consistency of Evidence and the Accurate Assessment of Uncertainty. *Medical Decision Making*. 2002;22(4):359-71.
10. Knorr-Held R, Best N. A shared component model for detecting joint and selective clustering of two diseases. In: LTD BP, editor. *International Conference on the Analysis and Interpretation of Disease Clusters and Ecological Studies*1999. p. 73-85.
11. Cooper NJ, Sutton AJ, Abrams KR, Turner D, Wailoo A. Comprehensive decision analytical modelling in economic evaluation: a Bayesian approach. *Health Economics*. 2004;13:203-26.
12. Sutton AJ, Cooper NJ, Goodacre S, Stevenson M. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. *Medical Decision Making*. 2008;28(5):650-67.
13. Thompson ML. Assessing the diagnostic accuracy of a sequence of tests. *Biostat*. 2003;4(3):341-51.

14. Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS user manual: Version 1.4. Cambridge: MRC Biostatistics Unit; 2003.
15. Ades AE, Lu G, Higgins JPT. The interpretation of random-effects meta-analysis in decision models. *Medical Decision Making*. 2005;25:646-54.
16. Ades AE, Lu G, Claxton K. Expected value of sample information calculations in medical decision modelling. *Medical Decision Making*. 2004;24:207-27.
17. Claxton K. The irrelevance of inference: A decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics*. 1999;18:341-64.

