ABSTRACT [First-level Header]

Objectives: It is standard practice for diagnostic tests to be evaluated against gold standards in isolation. In routine clinical practice, however, it is commonplace for multiple tests to be used before making definitive diagnoses. This paper describes a meta-analytic modelling framework developed to estimate the accuracy of the combination of two diagnostic tests, accounting for the likely non-independence of the tests.

Methods: A novel multi-component framework was developed to synthesise information available on different parameters in the model. This allows data to be included from studies evaluating single tests or both tests. Different likelihoods were specified for the different sources of data and linked by means of common parameters. The framework was applied to evaluate the diagnostic accuracy of Ddimer and Wells score for Deep Vein Thrombosis, and the results compared to a model where independence of tests was assumed. All models were evaluated using Bayesian Markov Chain Monte Carlo simulation methods.

Results: The results showed the importance of allowing for the (likely) non-independence of tests in the meta-analysis model when evaluating a combination of diagnostic tests. The analysis also highlighted the relatively limited impact of those studies that evaluated only one of the two tests of interest.

Conclusions: The models developed allowed the assumption of independence between diagnostic tests to be relaxed while combining a broad array of relevant information from disparate studies. The framework also raises questions regarding the utility of studies limited to the evaluation of single diagnostic tests.

Introduction [First-level Header]

Accurate diagnosis is a prerequisite for the efficient allocation of treatments. Diagnostic tests with perfect or very high accuracy (reference tests) are often expensive and/or invasive; therefore, index tests, which are usually cheaper and less invasive but also less accurate, often play an important role in medical diagnosis. Rarely is the application of one index test sufficient to diagnose a particular condition, and diagnostic strategies involving multiple tests are often used in routine clinical practice. Where multiple tests are used for diagnosis, however, it is important to acknowledge that the diagnostic results from the different tests may not be independent of one another and therefore, when synthesising evidence to evaluate the accuracy of the combination of tests this interdependence needs to be taken into account, which is seldom done in practice.

Systematic reviews and, consequently, meta-analyses are routinely used to identify the evidence for medical decision making [1] and, more specifically, for clinical/economic decision analytic modelling [2]; since optimal decisions should not be based solely on single study results when multiple studies with relevant data exist [3,4]. Systematic reviews and meta-analyses of diagnostic test accuracy studies have focused on the performance of individual tests which, at least in part, is due to a large proportion of primary studies focusing on the evaluation of single tests. A recent systematic review of Health Technology Assessment reports [5] found that where economic decision models had been used to evaluate different combinations of tests, the accuracy of each combination was calculated based either 1) on the assumption of conditional independence between tests, or 2) by assuming the accuracy of the second test to be perfect. There is evidence that when the assumption of dependence between tests is not met, then both the meta-analysis (for the estimates of the accuracy rates) and the economic evaluation (informed by the meta-analysis results) have the potential to give

misleading conclusions [6]. In this paper we focus solely on clinical effectiveness with an associated paper [7] focusing on cost-effectiveness implications.
A number of approaches already exist to model test accuracy data allowing for conditional dependence between tests; however, these only consider data from a single study. These include 1) the estimation of the covariance between test results conditional on disease status [8-12]; 2) use of latent variable models [13-15]; 3) use of linear discriminant procedures to select the best combination of tests according to some maximising functions [16-22]; and 4) use of approaches based on distribution free statistics [22-24]. While the vast majority of the meta-analytic methodological literature focuses on estimating performance of individual tests, Siadaty et al. [25] do consider the simultaneous estimation of multiple tests allowing for dependency between patients for which multiple tests are available (i.e. where individual studies considered multiple tests). This approach, however, did not consider the estimation of the accuracy of *combinations of the tests* considered in their framework. Further, there is a growing methodological literature [26] on the estimation of multiple test performance in the absence of a gold reference standard which has some commonalties with the analyses presented here (though all studies included in our syntheses are assumed to have a gold standard reference test).

In this paper we propose, what we believe to be, the first modelling framework developed to estimate meta-analytically the accuracy of combinations of diagnostic tests, acknowledging the likely non-independence of the tests. The next section (Section 2) describes the motivating example of Ddimer test and Wells score for the diagnosis of Deep Vein Thrombosis (DVT). Section 3 describes the meta-analytical modelling framework developed to estimate the accuracy of combinations of diagnostic tests. The results from applying the framework developed to the motivating example are presented in Section 4, and Section 5, the discussion, concludes the paper.

Motivating example: Ddimer and Wells score tests for the diagnosis of Deep Vein Thrombosis [First-level Header]

Background [Second-level Header]

DVT is a blood clot in a deep vein (lower limb) that is usually treated with anticoagulants. Prompt treatment is essential in order to lower the risk of mortality due to Venous Thromboembolism related potential adverse events. Also, due to the potentially life-threatening side effects from anticoagulant treatment, the number of patients wrongly diagnosed as having DVT when they do not have the condition (i.e. false positives) needs to be kept to a minimum. Therefore, it is important that an accurate diagnosis of DVT is obtained quickly.

Reference tests with high diagnostic accuracy exist for DVT such as Ultrasound or Venography; however, several other index tests exist that are less accurate but cheaper, quicker and less invasive, such as Ddimer and Wells score [27,28]. Ddimer measures the concentration of an enzyme in the blood (i.e. the higher the measurement the more likely DVT) and Wells score is devised from an assessment of the clinical features of DVT (i.e. clinical history, symptoms and signs) [28,29]. A simplified and widely used version of the Wells score (as used in this paper) categorises patients into *low* (score <1), moderate (score 1 or 2) and *high* (score >2) risk of having DVT.

In a previous evaluation of the effectiveness and cost-effectiveness of different tests for DVT, Goodacre et al. [27] found that Ddimer and Wells score were not accurate enough as stand-alone diagnostic tools but there was evidence that test sequences containing both Wells score and Ddimer were potentially valuable; however, due to the limited methodology available at the time the approach taken to account for test dependency was limited.

The data [Second-level Header]

We carried out an initial systematic review (details available on request from the corresponding author) to identify publications reporting accuracy data of Ddimer stratified by Wells score (for the common threefold categorization) either for all Wells categories or only specific strata. The data identified from this systematic review is presented in Table 1. Eleven studies were identified that reported diagnostic performance of Ddimer for each of the 3 Wells score strata; these data are subsequently referred to as Type A. A further 3 studies reported on each on the 3 Wells strata but only had Ddimer results for one of the 3 strata; these data are subsequently referred to as Type B. Thirdly, for 4 further studies, Wells performance data were only available from a single strata but, for each of these reported strata, Ddimer data were also available; this data are subsequently referred to as Type C data.

In addition to these data, we include in our modelling framework, the considerable body of evidence on the diagnostic accuracy of Wells score alone and Ddimer alone. We identify this literature through published systematic reviews on the accuracy of Wells score [30] (updated with study T33 in Appendix A in Supplemental Materials at: XXX) and Ddimer alone [31]); subsequently referred to as Type D data (N=18 studies) and Type E data (N=97 studies) respectively (See Appendix A in Supplemental Materials at: XXX for inclusion criteria and references for all included and excluded studies. Note, some studies of Type D and E reported multiple different tests/patient groups for which data were analysed as separate observations. We subsequently refer to each set of observations from each study as an individual assay).

The diagnostic strategies of Wells score and Ddimer [Second-level Header]

In the framework developed, we follow the two possible schemes for combining two diagnostic tests outlined by Thompson [32]: 1) believe the negative result (i.e. only patients diagnosed as positive by the first test will be further tested), and 2) believe the positive result (i.e. only patients diagnosed as negative by the first test will be further tested). In this paper we will limit ourselves to evaluating the diagnostic accuracy of the 2 tests alone and 2 strategies evaluating the use of the 2 tests in combination as described below. Note, for simplicity, we have dichotomised Wells score into low versus moderate and high (though the approach would generalise to multiple categories and further categories are considered in the associated economic evaluation paper [7]).

1. Wells score only dichotomised as low versus moderate and high
2. Ddimer only at operative threshold as reported by the manufacturer
3. Wells score followed by Ddimer using the believe the negatives criterion
4. Wells score followed by Ddimer using the believe the positives criterion

In strategies 3 and 4, we have chosen to evaluate strategies where Wells score is used as the first diagnostic test (since it is the quickest and least invasive, although order does not affect overall test performance) followed by Ddimer.

Analysis framework [First-level Header]

Overview of analysis framework [Second-level Header]

Our overarching approach to analysis, which incorporates shared parameter modelling [2] is as follows:

1) Define the basic intermediate parameters (i.e. the probability of being diseased / healthy for each Wells strata, and the sensitivity and specificity of Ddimer stratified by Wells strata) that can be estimated using the data available;
2) Specify (different) likelihoods for each of the data types (A to E) in terms of these basic intermediate parameters; and finally;
3) Estimate the quantities of interest (i.e. the estimates of test accuracy for combinations of tests) from the basic intermediate parameters through functional transformations.

In Sections 3.3 and 3.4 which follow, a full description of the analysis framework is given.

Algebraic notation of data [Second-level Header]

Table 2 defines the algebraic notation used to describe the study data and presents this below the reproduced data for Type A, D and E. Type B and C data conform to the notation of Type A but with missing Ddimer values for some of the Wells score strata. For Type E studies (Ddimer data only) test accuracy data is only available for all patients and hence is notated as aggregated across Wells score strata.

In Table 2, d_{ki} and h_{ki} define the number of diseased and healthy individuals in the *k*th Wells strata (i.e. 1=low, 2=moderate, 3=high) of the *i*th study. The total diseased and the total healthy in study *i* is defined as N_{Di} and N_{Hi} respectively.

For the accuracy of Ddimer, tp_{ki} is the number of diseased patients that are correctly classified as positive by Ddimer (true positive (TP)) and tn_{ki} is the number of healthy patients that are correctly classified as negative by Ddimer (true negative (TN)) for the k^{th} Wells score strata for study *i*. Similarly, the number of healthy patients diagnosed as diseased (false positives (FP)) and the number of diseased patients diagnosed as healthy (false negatives (FN)) can be defined as stated in Table 2.

Sensitivity (i.e. the proportion of diseased patients which are correctly identified by the test) and specificity (i.e. the proportion of healthy patients which are correctly identified by the test) of Ddimer for the k^{th} Wells score strata for study *i* can be derived from the above quantities as follows:

$$sens_{ki} = \frac{tp_{ki}}{d_{ki}}$$

$$spec_{ki} = \frac{tn_{ki}}{h_{ki}}$$

The synthesis model used to combine the data available from each study is described in the next section.

Defining and estimating the basic intermediate parameters [Second-level Header]

As described in previous section, the relation between the data and the intermediate parameters is presented via the description of the likelihoods and using a multi-component model with shared parameters. In the account which follows, the number of studies of Type A, B, C, D and E are denoted as n_A, n_B, n_C, n_D and n_E respectively.

Wells score strata data (Type A, B, C & D): Multinomial random effect logistic meta-analysis model [Third-level Header]

Complete Wells score strata data (Type A, B and D) are modelled using a multinomial logistic regression model in the form presented by Ntzoufras [33] adapted to fit between-study random effects to account for heterogeneity. The likelihood for these data are specified via multinomial distributions (see Equation 1) with parameters p_{Dki} indicating the probability of being in the k^{th} Wells strata (k = 1, 2, 3) for a diseased patient in study i and similarly, p_{Hki} indicating the probability of being in the k^{th} Wells strata for a healthy patient.

$$(d_{1i}\,, d_{2i}\,, d_{3i}) \sim multinom((p_{D1i}, p_{D2i}, p_{D3i}); N_{Di})$$

$$(h_{1i}, h_{2i}, h_{3i}) \sim multinom\big((p_{H1i}, p_{H2i}, p_{H3i}); N_{Hi}\big) \qquad \text{Equation 1}$$

for i from 1 to $n_A + n_B$ (type A, B) and for i from $n_A + n_B + n_C + 1$ to $n_A + n_B + n_C + n_D$ (type D)

Where all other notation is as defined in the previous section.

For incomplete Wells score strata data (Type C) the multinomial likelihoods cannot be used due to the missing data. Note, Type C data will influence neither the estimate of the intermediate parameters nor the estimate of the final parameters for Wells score, but will contribute to the estimation of the conditional accuracy of Ddimer. A combination of binomial likelihoods is used instead of the multinomial likelihood. The assumption of exchangeability between studies [34] allows for the estimation of the missing strata data and the *p_{Dki}*'s and *P_{Hki}*'s when summed over *k* are constrained to equal 1. Further, to allow the estimation of the missing denominator data (as denoted by $\widehat{N}_{Di}$ and $\widehat{N}_{Hi}$), exchangeability between studies is assumed via the indirect estimation of the parameter for which there is no information (see below).

$$d_{ki} \sim binomial(p_{Dki}; \widehat{N}_{Di})$$
$$h_{ki} \sim binomial\left(p_{Hki}, \widehat{N}_{Hi}\right)$$
$$for\ i\ from\ \ n_A + n_B + 1 \quad to \quad n_A + n_B + n_c\ (type\ C)$$
$$for\ k = 1, 2\ and\ 3$$

Equation 2 [Second-level Header]

Type A, B, C and D data are then synthesised using a random effects structure with common between study variability (on a logit scale). The transformed parameters, ξ_{Dki} and ξ_{Hki}, are assumed to be exchangeable from distributions with mean parameters ξ_{Dk} and ξ_{Hk}, for Wells score level (k) 1=low, 2=moderate and 3=high. The degree of heterogeneity is assumed the same across the three diseased and healthy Wells score strata and the between study variance is represented by σ_D^2 and σ_H^2 respectively. Note, how the likelihoods in both equations 1 and 2 are linked by means of the parameters η_{Dk} and η_{Hk} in Equation 3.

$$p_{Dki} = \frac{\eta_{Dki}}{\sum_{k=1}^{3} \eta_{Dki}} \quad , \quad \xi_{Dki} = \ln(\eta_{Dki})$$

$$p_{Hki} = \frac{\eta_{Hki}}{\sum_{k=1}^{3} \eta_{Hki}} \quad , \quad \xi_{Hki} = \ln(\eta_{Hki})$$

Equation 3

$$\xi_{Dki} \sim Normal(\xi_{Dk}, \sigma_D^2) \; ; \; \xi_{Hki} \sim Normal(\xi_{Hk}, \sigma_H^2)$$

$$for\; i\; from \quad 1 \quad to \quad n_A + n_B + n_C + n_D \; (type\; A, B, C\; and\; D)$$

$$for\; WS\; level\; (k) \quad 1 = low, \;\; 2 = moderate \quad and \quad 3 = high$$

The overall meta-analysed proportions of diseased and healthy patients in each of the *k* Wells score strata (p_{Dk}^{pooled} and p_{Hk}^{pooled} respectively) are the basic intermediate parameters of interest.. These are obtained by the following back-transformations:

$$p_{Dk}^{pooled} = \exp(\xi_{Dk}) \Big/ \sum_{k=1}^{3} \exp(\xi_{Dk})$$ Equation 4

$$p_{Hk}^{pooled} = \exp(\xi_{Hk}) \Big/ \sum_{k=1}^{3} \exp(\xi_{Hk})$$

for *k* from 1 to 3 strata.

Ddimer data stratified by Wells score (Type A, B, C data): Bivariate random effect logistic meta-analysis model [Third-level Header]

Type A, B and C data which provide Wells score strata specific Ddimer accuracy data is modelled as three separate bivariate random effect models [35] – one for each strata. The formulation of the model allows for missing strata data, which exists for both Type B and C studies. Additionally, for Type C data the total number of diseased (d_{ki}) and healthy (h_{ki}) is also missing; however, this is estimated by the model for the meta-analysis of Wells score data as specified in Section 3.3.1. Algebraically,

$$tp_{ki} \sim binomial(sens_{ki}, d_{ki}) \qquad tn_{ki} \sim binomial(spec_{ki}, h_{ki})$$

$$logit(sens_{ki}) = \mu_{Dki} \qquad logit(spec_{ki}) = \mu_{Hki}$$ Equation 5

$$\binom{\mu_{Dki}}{\mu_{Hki}} \sim MultivariateNormal\left[M = \binom{\mu_{Dk}}{\mu_{Hk}}, \Sigma = \begin{pmatrix} \sigma_{Dk}^2 & \sigma_{DHk} \\ \sigma_{HDk} & \sigma_{Hk}^2 \end{pmatrix}\right]$$

$for\ i\ in\ 1\ to\ \ n_A + n_B + n_C\ \ (Type\ A, B\ and\ C)$ and

$for\ WS\ level\ (k)\ \ \ 1 = low,\ \ 2 = moderate\ \ \ and\ \ \ 3 = high$

Where μ_{Dki} and μ_{Hki} are the logit sensitivity and specificity respectively in the *k*th Wells score strata of the *i*th study; *M* is the vector of mean logit responses containing μ_{Dk} and μ_{Hk} (i.e. the mean logit sensitivity and specificity in the *k*th Wells score strata respectively); and *Σ* is the between-study covariance matrix containing the between study variances $(\sigma_{Dk}^2,\ \sigma_{Hk}^2)$ for logit sensitivity and specificity respectively and the covariance (σ_{DHk}^2) in the *k*th Wells score strata.

A back-transformation is required to estimate sensitivity and specificity of Ddimer for each of the *k* = 1 to 3 Wells score strata as presented below:

Equation 6

$$spec_k^{pooled} = \frac{\exp(\mu_{Dk})}{1 + \exp(\mu_{Dk})}$$

$$spec_k^{pooled} = \frac{\exp(\mu_{Hk})}{1 + \exp(\mu_{Hk})}$$

Ddimer data alone (Type E data) aggregated across Wells strata: Bivariate random effect logistic meta-analysis model [Third-level Header]

In order to include Type E data, which provides information on accuracy of Ddimer aggregated across Wells strata, it is assumed that the overall accuracy of Ddimer is a function of the proportion of diseased and healthy patients in each Wells score category and the Wells score strata specific accuracy performance parameters of Ddimer. The model is fitted as follows: A bivariate logit model, of the form outlined in Equation 5, is used to meta-analyse the overall accuracy of Ddimer aggregated across Wells strata (i.e. strata specific indexing dropped from the summary parameters).

$$TP.agg_i \sim binomial(sens.agg_i, N_{Di}) \;\; ; \; TN.agg_i \sim binomial(spec.agg_i, N_{Hi})$$

$$logit(sens.agg_i) = \mu.agg_{Di} \, ; \; logit(spec.agg_i) = \mu.agg_{Hi} \qquad \text{Equation } \mathbf{7}$$

$$\binom{\mu.agg_{Di}}{\mu.agg_{Hi}} = MVN\left[M = \binom{\mu.agg_D}{\mu.agg_H}, \Sigma = \begin{pmatrix} \sigma.agg_D^2 & \sigma.agg_{DH} \\ \sigma.agg_{HD} & \sigma.agg_H^2 \end{pmatrix}\right]$$

$$for\; i\; in\; n_A + n_B + n_C + n_D + 1 \quad to \quad n_A + n_B + n_C + n_D + n_E \;\; (type\; E)$$

Where parameters are of the form described in Equation 5 but *'.agg'* is added to parameters to indicate aggregation across Wells strata.

The aggregate Ddimer accuracy parameters, estimated in Equation 7, are now expressed as functions of the parameters used to define the strata specific accuracy model. Explicitly, since we do not know the proportion of diseased and healthy patients in each Wells score strata for these studies, we predict these proportions stochastically assuming them to be exchangeable with the studies we do know the proportions for (i.e. data types A, B, C and D). We then express the aggregate Ddimer accuracies as “weighted” averages of the Wells score strata (i.e. weighted by the predicted proportions in each Wells strata described above). Algebraically, the (logit) predicted proportions of patients in each Wells score category for diseased and healthy are defined as ${\xi^{new}}_{Dki}$ and ${\xi^{new}}_{Hki}$ respectively, where each is assumed to be exchangeable with the logit proportions estimated from the previous studies, vis.

$${\xi^{new}}_{Dki} \sim Normal(\xi_{Dk}\,, \sigma_D^2)$$

$${\xi^{new}}_{Hki} \sim Normal(\xi_{Hk}\,, \sigma_H^2) \qquad \text{Equation 8}$$

$$for\; i\; in\; n_A + n_B + n_C + n_D + 1 \quad to \quad n_A + n_B + n_C + n_D + n_E \;\; (type\; E)$$

where all other parameters are as defined in Equation 3. Equation 9 below defines the proportion of individuals predicted to be diseased or healthy on the natural scale for each Wells strata *k* (${p^{new}}_{Dki}$ and ${p^{new}}_{Hki}$ respectively).

$${p^{new}}_{Dki} = \frac{\exp({\xi^{new}}_{Dki})}{\sum_{k=1}^{3}\exp({\xi^{new}}_{Dki})} \qquad \text{Equation 9}$$

$${p^{new}}_{Hki} = \frac{\exp({\xi^{new}}_{Hki})}{\sum_{k=1}^{3}\exp({\xi^{new}}_{Hki})}$$

$$for\ i\ in\ n_A + n_B + n_C + n_D + 1 \quad to \quad n_A + n_B + n_C + n_D + n_E \quad (Type\ E)$$

Next, we express overall sensitivity and specificity as the “weighted” average of the Wells strata specific values of interest.

$$sens.agg_i = \sum_{k=1}^{3} sens_{ki} \times {p^{new}}_{Dki} \qquad \text{Equation 10}$$

$$spec.agg_i = \sum_{k=1}^{3} spec_{ki} \times {p^{new}}_{Hki}$$

$$for\ i\ in\ n_A + n_B + n_C + n_D + 1 \quad to \quad n_A + n_B + n_C + n_D + n_E \quad (Type\ E)$$

By doing this, Type E data will contribute information to the Wells strata specific estimate of Ddimer performance.

Estimating the accuracy parameters of interest from the basic intermediate parameters [Third-level Header]

The parameters of ultimate interest are the accuracies of the different diagnostic strategies (sensitivities and specificities) as outlined in Section 2.3. How these are estimated from the intermediate parameters, as described in the previous section, is presented below.

Strategy 1 - Wells score only dichotomised as low versus moderate and high [Fourth-level Header]

The parameters estimated in the previous section can be used to estimate the accuracy of Wells score dichotomised as low versus moderate and high. That is, the sensitivity (*$sens_{WS}$*) is $p_{D2} + p_{D3}$ and specificity (*$spec_{WS}$*) is p_{H1}.

Strategy 2 - Ddimer only at operative threshold as reported by the manufacturer [Fourth-level Header]

The accuracy of Ddimer on its own is not obtained from the multi-component model described previously since a standard synthesis model for a single test is all that is required. Ddimer data from Type A studies are aggregated over Wells score categories. This is combined with Type E data and then analysed using a bivariate random effect logit model [41].

Strategy 3 - Wells score followed by Ddimer using the believe the negatives criteria [Fourth-level Header]

The accuracy of Wells score dichotomised as low versus moderate and high under the believe the negatives strategy (i.e. a patient is considered healthy if either or both test results are negative) can be derived using the following formulae is:

$$sens_{(WS\ and\ DD)_{BN}} = sens_{WS} \times sens_{DD/WS=2,3}$$

$$where\ sens_{DD/WS=2,3} = \left(w_{D1} \times sens_2^{pooled}\right) + \left(w_{D2} \times sens_3^{pooled}\right)$$

$$and\ w_{D1} = \frac{p_{D2}}{(p_{D2}+p_{D3})}\ , w_{D2} = 1 - w_{D1} \qquad \text{Equation 11}$$

$$spec_{(WS\ and\ DD)_{BP}} = 1 - \left[(1 - spec_{WS}) \times \left(1 - spec_{DD/WS=2,3}\right)\right]$$

$$where\ spec_{DD/WS=2,3} = \left(w_{H1} \times spec_{2}^{pooled}\right) + \left(w_{H2} \times spec_{3}^{pooled}\right)$$

$$and\ w_{H1} = \frac{p_{H2}}{(p_{H2} + p_{H3})},\ \ w_{H2} = 1 - w_{H1}$$

where $sens_{DD/WS=2,3}$ and $spec_{DD/WS=2,3}$ are the sensitivity and specificity of Ddimer for the combined Wells score moderate (*k*=2) and high (*k*=3) strata.

Strategy 4 Wells score followed by Ddimer using the believe the positives criteria [Fourth-level Header]

The accuracy of Wells score dichotomised as low versus moderate and high, under the believe the positives (*(WS and DD)_{BP}*) strategy can be derived using the following formulae:

$$sens_{(WS\ and\ DD)_{BP}} = 1 - [(1 - sens_{WS}\ \ ) \times (1 - sens_{1}^{pooled}\ \ )]$$

$$spec_{(WS\ and\ DD)_{BP}} = spec_{WS}\ \ \times spec_{1}^{pooled} \qquad \text{Equation 12}$$

Analysis plan and approach to model fitting [Second-level Header]

The modelling framework is implemented using Markov chain Monte Carlo (MCMC) simulation [36] in WinBUGS software [37] for Bayesian modelling. Non-informative (vague) prior distributions are used for all parameters. An initial run of 5,000 iterations of the MCMC sampler were discarded as a ‘burn-in’(37), with inferences based on a further 20,000 sample iterations. Convergence of the MCMC chains was assessed and sensitivity analyses showed no influence of the initial values and prior distributions on the posterior distributions obtained. The WinBUGS code (including the specific prior distributions used) is provided in Appendix B in Supplemental Materials at: XXX.

The model described above was fitted to all data types (A, B, C, D and E) and is compared to an analysis in which the performance of both tests is assumed independent of one another. Additionally, in order to assess the impact of the different data types on the analysis, a further analysis was conducted in which each data type was sequentially added (i.e. A, AB, ABC, ABCD, ABCDE).

Results of the data analysis [First-level Header]

Estimates of basic intermediate parameters [Second-level Header]

Table 3 presents and compares the estimates of the basic intermediate parameters for the models assuming independence and dependence between Wells score and Ddimer. From Table 3 it can be observed that the proportions of diseased patients per Wells score category are similar regardless of the dependency assumption. Reassuringly, in both cases the proportion diseased increases and the proportion healthy decreases with Wells score category. In Table 3, the performance accuracy of Ddimer for each Wells score strata is reported. While sensitivity does vary across Wells score strata, it is specificity for which the biggest differences are observed (albeit with considerable uncertainty). That is, for the model that assumes dependence between tests, specificity for the low risk Wells score strata is estimated to be 0.699 (0.598 to 0.797), for moderate 0.390 (0.212 to 0.561) and for high 0.433 (0.300 to 0.566). In Table 3, all heterogeneity parameters are non-negligible suggesting variability between study results is greater than would be expected by chance. Such heterogeneity could be explored by adding covariates to the models presented, although, to keep this paper’s methodological innovations focused, this is not investigated further here.

Figure 1 displays the 95% credible regions [38,39] for the overall accuracy of Ddimer (across all Wells score strata) compared with the Wells score strata specific accuracy estimates. This plot highlights the potentially important differences in the performance of the Ddimer test for the different Wells score strata. If the assumption of independence between Wells score and Ddimer was true, then we would expect the four different credible regions displayed to be overlaid on top of one another but considerable divergence is observed.

Estimates of final parameters [Second-level Header]

Table 4 presents the results of the 4 different strategies of Wells score and Ddimer described in section 2.3 for the models assuming independence and dependence between tests. When comparing these two modelling approaches it can be observed that the sensitivities of the different strategies are similar for both the independent and dependent models but differences are observed for the specificities although all the credible intervals overlap. The implications of the performance of these different strategies is considered further in a full economic analysis elsewhere [7].

Table 5 presents the results of the 4 different strategies sequentially adding the different data types (i.e. A, AB, ABC, ABCD, ABCDE). It can be observed that there is relatively limited impact, both in terms of the point estimate and uncertainty, of adding data from the studies which evaluated only one of the 2 tests of interest (i.e. Types D and E) despite the fact that there are relatively higher numbers of these types of studies compared to the other data types (Types A, B and C).

Discussion [First-level Header]

This paper presents a meta-analytic framework which allows for the fact that the performance of multiple diagnostic tests, when used in combination, may not be independent from one another. This is in contrast to current practice in HTA where independence between tests is commonly assumed [5] (which is potentially misleading and can lead to over-estimation of the performance of combinations of tests). We believe this is the first published methodological research in this aspect of evidence synthesis methodology.

Our overall approach, in which the relationship to disparate data sources is expressed using multiple likelihood functions sharing common parameters, has much in common with other recent developments in evidence synthesis methodology in other contexts [40-42]. Such an approach allows the use of data from studies reporting on the accuracy of individual tests or multiple tests given to the same patients (completely or incompletely reported). In this way the amount of data that can be incorporated from the literature is maximised. The approach described could be adapted to the case where both tests are dichotomous and extensions to incorporate 3 or more tests, or situations where the gold standard is imperfect could be developed. Further, our interest was evaluating the overall performance of a sequence of tests; a similar approach may be used to evaluate the performance of tests which compete for the same location in a given diagnostic pathway. We believe this work has an important message for those funding and conducting new studies estimating the diagnostic accuracy of tests. In the DVT example presented here, the majority of research had been carried out in studies evaluating only a single test (i.e. data types D and E). This, perhaps, is at discord with clinical practice where we believe it is commonplace for multiple tests to be used to diagnose patients. In the motivating example, at least, the literature on the individual studies had minimum impact on the estimation of the sequences of tests of interest. Although this finding needs further investigation (including application in other clinical contexts) the potential implication is that studies of individual tests are highly inefficient if test sequences are of ultimate interest and therefore our research would suggest that studies evaluating multiple tests should replace many of the studies of individual tests currently performed carefully ensuring appropriate clinical context. On completion of our work, we were heartened to see a new prospective cohort study evaluating the combined performance of the Wells score and Ddimer test in a primary care setting(43). Indeed, the findings reported here lead us to question, when carrying out such syntheses, whether identification and synthesis of studies reporting individual tests is even justifiable - given the resource implications - when evaluating test combinations.

As this was a methodological paper, we relied on previous meta-analyses of the performance of Ddimer and Wells score and we acknowledge these have limitations including 1) studies of different Ddimer test products were combined together; 2) considerable heterogeneity in study results was not explored by the incorporation of covariates (although the framework presented would allow this extension); and 3) issues of variable study quality are largely ignored.

We appreciate the interpretation of the findings of the type of analyses presented here need careful consideration. For the motivating example (Table 4), identifying an optimal strategy is not straightforward since an explicit trade-off between sensitivity and specificity is required for decision making. To do this, among other things, the relative impact of a false positive compared to a false negative diagnosis will need consideration. Further, economic considerations are increasingly relevant and thus the economic consequences of the alternative clinical pathways will also be necessary and we extend our research in a further paper to demonstrate how this can be achieved [7].

Acknowledgments [First-level Header]

The authors would like to thank Professor Steve Goodacre for useful discussions regarding previous evaluations of Ddimer and Wells score tests for DVT, and Professor Keith Abrams for interesting discussions regarding the analysis.

References [First-level Header]

1. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.1: The Cochrane Collaboration; 2008.

2. Welton NJ, Sutton AJ, Cooper NJ, Abrams KR, Ades AE. Evidence synthesis for decision making in healthcare. Chichester: Wiley; 2012.

3. George FB, Oscar L, Jaap F, Rogier D. The evidence provided by a single trial is less reliable than its statistical analysis suggests. J Clin Epidemiol 2009;62:711-5.e1.

4. Lau J, Ioannidis JPA, Schmid CH. Summing up evidence: one answer is not always enough. Lancet 1998;351:123-7.

5. Novielli N, Cooper NJ, Abrams KR, Sutton AJ. How Is Evidence on Test Performance Synthesized for Economic Decision Models of Diagnostic Tests? A Systematic Appraisal of Health Technology Assessments in the UK Since 1997. Value Health 2010;13:952-7.

6. van Walraven C, Austin PC, Jennings A, Forster AJ. Correlation between serial tests made disease probability estimates erroneous. J Clin Epidemiol 2009;62:1301-5.

7. Novielli N, Cooper NJ, Sutton AJ. Evaluating the cost-effectiveness of diagnostic tests in combination: Is it important to allow for performance dependency? Value Health Submitted. year;volume:page range <<This is a companion paper that has also been accepted for publication and therefore has not yet been assigned a year, volume, pages>>.

8. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics 1980;36:167-71.

9. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics 1985;41:959-68.

10. Enoe C, Georgiadis MP, Johnson WO. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Prev Vet Med 2000;45:61-81.

11. Dendukuri N, Lawrence J. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics 2001;57:158-67.

12. Georgiadis MP, Johnson WO, Gardner IA, Singh R. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. J R Stat Soc Series C 2003;52:63-76.

13. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: A multiple latent variable model. Stat Med 2009;28:441-61.

14. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol 1995;141:263-72.

15. Principato F, Vullo A, Matranga D. On implementation of the Gibbs sampler for estimating the accuracy of multiple diagnostic tests. J Appl Stat 2010;37:1335-54.

16. Shen Y, Wu D, Zelen M. Testing the independence of two diagnostic tests. Biometrics 2001;57:1009-17.

17. Jin H, Lu Y. A Procedure for determining whether a simple combination of diagnostic tests may be noninferior to the theoretical optimum combination. Med Decis Making 2008;28:909-16.

18. Liu A, Schisterman EF, Zhu Y. On linear combinations of biomarkers to improve diagnostic accuracy. Stat Med 2005;24:37-47.

19. Macaskill P, Walter SD, Irwig L, Franco EL. Assessing the gain in diagnostic performance when combining two diagnostic tests. Stat Med 2002;21:2527-46.

20. McIntosh MW, Pepe MS. Combining several screening tests: Optimality of the risk score. Biometrics 2002;58:657-64.

21. Qin J, Zhang B. Best combination of multiple test for screening purposes. Stat Med 2010;29:2905-19.

22. Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. J Am Stat Assoc 2010;88:1350-5.

23. Huang X, Qin G, Fang Y. Optimal combinations of diagnostic tests based on AUC. Biometrics 2011;67:568-76.

24. Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. Biostatistics 2000;1:123-40.

25. Siadaty MS, Philbrick JT, Heim SW, Schectman JM. Repeated-measures modeling improved comparison of diagnostic tests in meta-analysis of dependent studies. J Clin Epidemiol 2004;57:698-711.

26. Dendukuri N, Schiller I, Joseph L, Pai M. Bayesian meta-analysis of the accuracy of a test for Tuberculous Pleuritis in the absence of a gold standard reference. Biometrics 2012; 68 (4): 1285-1293

27. Goodacre S, Sampson F, Stevenson M, et al. Measurement of the clinical and cost-effectiveness of non-invasive diagnostic testing strategies for deep vein thrombosis Health Technol Assess 2006;10:168.

28. Wells PS, Hirsh J, Anderson DR, et al. Accuracy of clinical assessment of deep-vein thrombosis. Lancet 1995;345:1326-30.

29. Wells PS, Anderson DR, Bormanis J, et al. Application of a diagnostic clinical model for the management of hospitalized patients with suspected deep-vein thrombosis. Thromb Haemost 1999;81:493-7.

30. Goodacre S, Sutton AJ, Sampson FC. Meta-analysis: The value of clinical assessment in the diagnosis of deep venous thrombosis. Ann Intern Med 2005;143:129-39+I-40.

31. Goodacre S, Sampson FC, Sutton AJ, et al. Variation in the diagnostic performance of D-dimer for suspected deep vein thrombosis. QJM 2005;98:513-27.

32. Thompson ML. Assessing the diagnostic accuracy of a sequence of tests. Biostatistics 2003;4:341-51.

33. Ntzoufras I. Bayesian modeling using WinBUGS. Chichester: Wiley; 2010.

34. Bernardo J, Smith A. Bayesian theory. Chichester: Wiley; 1994.

35. Arends LR, Hamza TH, van Houwelingen HC, et al. Bivariate random effects meta-analysis of ROC curves. Med Decis Making 2008;28:621-38.

36. Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics. Chapman and Hall: London; 1996.

37. Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS user manual: Version 1.4. Cambridge: MRC Biostatistics Unit; 2003.

38. Harbord RM, Deeks JJ, Egger M, et al. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 2007;8:1 - 21.

39. Novielli N, Cooper NJ, Sutton AJ, Abrams KR. Bayesian model selection for meta-analysis of diagnostic test accuracy data: Application to Ddimer for deep vein thrombosis. Res Synth Methods 2010;1:226-38.

40. Ades AE, Cliffe S. Markov Chain Monte Carlo estimation of a multiparameter decision model: Consistency of evidence and the accurate assessment of uncertainty. Med Decis Making 2002;22:359-71.

41. Sutton AJ, Kendrick D, Coupland CAC. Meta-analysis of individual- and aggregate-level data. Stat Med 2008;27:651-69.

42. Jackson CH, Best NG, Richardson S. Bayesian graphical models for regression on multiple data sets with different variables. Biostatistics 2009;10:335-51.

43. Geersing G, Erkens PM, Lucassen WAM, et al. Safe exclusion of pulmonary embolism using the Wells rule and qualitative D-dimer testing in primary care: prospective cohort study. BMJ 2012;345:e6564.