Systematic review and meta-analysis of the diagnostic accuracy of ultrasonography for deep vein thrombosis

Steve Goodacre*1, Fiona Sampson1, Steve Thomas2, Edwin van Beek3 and Alex Sutton4

Address: 1School of Health, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK, 2Academic Vascular Unit, University of Sheffield, Coleridge House, Northern General Hospital, Herries Road, Sheffield, S5 7AU, UK, 3Carver College of Medicine, University of Iowa Hospitals and Clinics, Department of Radiology, 3895 JFP, 200 Hawkins Drive, Iowa City, IA 52242-1077, USA and 4Department of Health Sciences, University of Leicester, 22-28 Princess Road West, Leicester, LE1 6TP, UK

Email: Steve Goodacre* - s.goodacre@sheffield.ac.uk; Fiona Sampson - f.c.sampson@sheffield.ac.uk; Steve Thomas - s.m.thomas@sheffield.ac.uk; Edwin van Beek - edwin-vanbeek@uiowa.edu; Alex Sutton - ajs22@leicester.ac.uk

* Corresponding author

Abstract

Background: Ultrasound (US) has largely replaced contrast venography as the definitive diagnostic test for deep vein thrombosis (DVT). We aimed to derive a definitive estimate of the diagnostic accuracy of US for clinically suspected DVT and identify study-level factors that might predict accuracy.

Methods: We undertook a systematic review, meta-analysis and meta-regression of diagnostic cohort studies that compared US to contrast venography in patients with suspected DVT. We searched Medline, EMBASE, CINAHL, Web of Science, Cochrane Database of Systematic Reviews, Cochrane Controlled Trials Register, Database of Reviews of Effectiveness, the ACP Journal Club, and citation lists (1966 to April 2004). Random effects meta-analysis was used to derive pooled estimates of sensitivity and specificity. Random effects meta-regression was used to identify study-level covariates that predicted diagnostic performance.

Results: We identified 100 cohorts comparing US to venography in patients with suspected DVT. Overall sensitivity for proximal DVT (95% confidence interval) was 94.2% (93.2 to 95.0), for distal DVT was 63.5% (59.8 to 67.0), and specificity was 93.8% (93.1 to 94.4). Duplex US had pooled sensitivity of 96.5% (95.1 to 97.6) for proximal DVT, 71.2% (64.6 to 77.2) for distal DVT and specificity of 94.0% (92.8 to 95.1). Triplex US had pooled sensitivity of 96.4% (94.4 to 97.1%) for proximal DVT, 75.2% (67.7 to 81.6) for distal DVT and specificity of 94.3% (92.5 to 95.8). Compression US alone had pooled sensitivity of 93.8% (92.0 to 95.3%) for proximal DVT, 56.8% (49.0 to 66.4) for distal DVT and specificity of 97.8% (97.0 to 98.4). Sensitivity was higher in more recently published studies and in cohorts with higher prevalence of DVT and more proximal DVT, and was lower in cohorts that reported interpretation by a radiologist. Specificity was higher in cohorts that excluded patients with previous DVT. No studies were identified that compared repeat US to venography in all patients. Repeat US appears to have a positive yield of 1.3%, with 89% of these being confirmed by venography.

Conclusion: Combined colour-doppler US techniques have optimal sensitivity, while compression US has optimal specificity for DVT. However, all estimates are subject to substantial unexplained heterogeneity. The role of repeat scanning is very uncertain and based upon limited data.
Background

Deep vein thrombosis (DVT) is an important cause of mortality and morbidity that requires accurate diagnosis. Ultrasound (US) examination has now largely replaced contrast venography as the standard test for diagnosing clinically suspected DVT [1]. Numerous studies have compared US to contrast venography in patients with clinically suspected DVT. These were most recently summarised by Kearon in 1998 who concluded that US had a sensitivity of 97% for proximal DVT, 72% for distal DVT and a specificity of 94% [2].

Meta-analytic techniques have developed rapidly in recent years. There is increasing recognition that the results of individual studies of a diagnostic test are often subject to substantial heterogeneity and that methodological factors may influence the results of studies [3,4]. Statistical techniques, such as meta-regression, allow researchers to explore data from systematic reviews for evidence that study-level covariates may influence diagnostic accuracy. There is also an increasing recognition that systematic reviews of diagnostic test data may be subject to publication bias, [4] although solutions to this problem, such as registries of studies, have yet to be developed.

Since US is now established as a definitive diagnostic test for DVT it is unlikely that many new studies evaluating the diagnostic accuracy of US will be forthcoming. This therefore represents an opportune time to undertake a definitive systematic review, meta-analysis and meta-regression of the diagnostic accuracy of US for clinically suspected DVT. We aimed to estimate the sensitivity and specificity of US for DVT, identify study-level covariates that are associated with variation in sensitivity and specificity, and seek evidence of publication bias in diagnostic studies of US for DVT.

Methods

We sought to identify all diagnostic cohort studies of patients with clinically suspected DVT who underwent testing with US followed by a reference standard of contrast venography. We searched Medline, EMBASE, CINAHL, Web of Science, Cochrane Database of Systematic Reviews, Cochrane Controlled Trials Register, Database of Reviews of Effectiveness, and ACP Journal Club (1966 to April 2004). The bibliographies of all articles selected for the review were scanned for potentially relevant articles that were not identified by the original search.

Two independent reviewers (ST and EVB) extracted the following data from the selected studies onto a standardised proforma: the setting for patient recruitment, any exclusion criteria, population demographics, whether recruitment was consecutive and/or data collection prospective, which US technique was used, the US operator, and the number of true positives (proximal and distal), true negatives, false positives and false negatives (proximal and distal), either as reported or calculated from the reported data. The same two reviewers also independently determined whether US was interpreted by observers blind to the venogram result, and whether venography was interpreted by observers blind to the results of US. Discrepancies were checked and resolved by an independent reviewer (FS). If it was not possible to extract the necessary data from the published report we contacted the authors for clarification. We reviewed the data reported by each study and removed studies that contained duplicated data.

Statistical analysis

Random effects models were used to estimate overall sensitivity and specificity, and a Chi-square test for heterogeneity between studies. Where 0 counts occurred for study data, a continuity correction of 0.5 was added to every value for that study in order to make the calculation of sensitivity and specificity defined. These analyses were undertaken using MetaDiSc statistical software [6] and further details of the models fitted is given elsewhere [7]. Initially all studies were analysed together and random effects meta-regression undertaken to identify potential causes of heterogeneity for sensitivity and specificity separately [8] (analysis carried out in STATA). Any covariate that showed an association with sensitivity or specificity
Characteristics of the included cohorts

The studies reported a total of 10,323 patients, with cohorts varying in size from 11 to 847 patients (median N = 72). The studies varied in the way they reported their findings: 53 reported proximal and distal DVT separately, 19 only reported proximal DVT, three only reported distal DVT, and 25 were unclear or reported proximal and distal DVT together. DVT prevalence varied from 20% to 94% (median 48%). The proportion of proximal DVT (of all DVT detected) ranged from 48% to 100% (median 78%). The mean or median age was reported by 60 studies, and ranged from 39 to 68 (median 57). The male to female ratio was reported by 65 studies, with the proportion of males ranging from 15% to 95% (median 45%).

Cohorts were recruited from the following settings: outpatient clinic-11, inpatients-12, emergency department-4, mixed-18, and not stated-55. Recruitment was reported to be consecutive in 48, and prospective in 67. Twelve cohorts excluded patients with previous DVT, while 45 papers did not report any exclusion criteria. The following techniques were used: 22 used compression ultrasonography alone, five used Colour Doppler alone, 16 used continuous wave Doppler alone, 25 used triplex, 28 used duplex, and four used other techniques. Ultrasound was interpreted blind to the results of venography in 62 cohorts and was unclear in 38. Venography was interpreted blind to the ultrasound result in 56 cohorts, was interpreted by observers aware of ultrasonad result in two, and was unclear in 42.

Results of meta-analysis

Figures 2 and 3 show the Forest plots of sensitivity and specificity respectively. Point estimates of sensitivity and specificity are plotted, with 95% confidence intervals, for each cohort. Pooled sensitivity (95% CI, p-value for heterogeneity) for detecting any DVT was 89.7% (88.8 to 90.5, p < 0.001). Pooled sensitivity for detecting proximal DVT was 94.2% (93.2 to 95.0, p < 0.001) and for distal DVT was 63.5% (59.8 to 67.0, p < 0.001). Pooled specificity, calculated using data from all the studies, was 93.8% (93.1 to 94.4, p < 0.001). When restricted to the 53 studies reporting full data specificity was 94.2% (93.4 to 95.0, p < 0.001). Great care should be taken when interpreting these estimates because of the substantial heterogeneity. It may be argued that calculating summary estimates in these circumstances is inappropriate. However, it does provide a useful baseline from which to explore heterogeneity.

Results of meta-regression

We undertook random effects meta-regression to identify possible causes for the heterogeneity. The results of meta-regression are outlined in Table 1. Using a threshold of p < 0.1 for statistical significance, interpretation by a radiologist, prevalence of DVT, the proportion of proximal DVT and date of publication were all significant predictors of sensitivity. The only significant predictor of specificity was exclusion of patients with a previous history of DVT.

More recently published studies, those with a higher prevalence of DVT and those with a higher proportion of proximal DVT tended to have higher sensitivity. There were 33 studies in which the operator was reported as being a
Potentially relevant studies identified and screened for retrieval, n=3992

Studies excluded, n=3592

Studies retrieved for more detailed evaluation, n=400

Studies excluded, n=249

Potentially suitable studies to be included in the meta-analysis, n=151
With six additional studies identified from citation lists

Studies excluded from the meta-analysis:
Duplicated data, n=6
Unable to extract data, n=9
Asymptomatic or mixed cohorts, n=43

Studies included in the meta-analysis, n=99

Figure 1
Flow diagram of studies considered for the review.
Figure 2
Forest plot of sensitivity.
Figure 3
Forest plot of specificity.
radiologist. Meta-analysis showed that that diagnostic performance was generally slightly worse among these studies. Overall sensitivity (95% CI) was 86.1% (83.8 to 88.3), sensitivity for proximal DVT was 94.4% (92.3 to 96.1), sensitivity for distal DVT was 62.6% (55.4 to 69.4), and specificity was 92.4% (90.9 to 93.7). Twelve cohorts reported excluding patients with previous DVT. Meta-analysis showed that that specificity was higher amongst these cohorts: 97.6% (96.6 to 98.3).

Table 2 shows pooled estimates of sensitivity and specificity stratified by US technique used. Optimal sensitivity is achieved by using duplex or triplex, while optimal specificity is achieved by using compression alone.

**Funnel plots**

These are shown in figure 4 (sensitivity) and figure 5 (specificity). Both plots are asymmetrical, suggesting that smaller studies tend to report higher sensitivity and specificity. One possible explanation of this is publication bias. Smaller studies reporting lower sensitivity or specificity may be less likely to be submitted or accepted for publication.

**Repeat or serial US**

We did not identify any studies that compared the results of repeat or serial scanning to venography in a complete cohort of patients, so none were included in the meta-analysis. However, we did identify several studies that reported the results of repeat US: five studies used repeat scanning for unselected cohorts with suspected DVT; [109-113] while four used repeat scanning for selected groups, based on the results of clinical risk scoring or D-dimer measurement [114-117]. Three studies used venography in some patients to confirm the results of positive repeat scanning [109,112,113]. Results from these studies are summarised in Table 3.

In unselected cohorts repeat scanning had a positive yield of zero to 2%. Where venography was used to confirm positive findings, the positive predictive value of ultrasound was 82 to 94%. Overall, our best estimate of the positive yield of repeat scanning in unselected patients is 35/2610 (1.34%; 95% CI 0.97 to 1.86%) with a positive predictive value of 146/164 (89.0%; 95% CI 83.3 to 92.9%).

When repeat scanning is restricted on the basis of clinical probability or D-dimer the results suggest a higher yield of positive scans, although none of the studies used venographic confirmation. Two studies of repeat ultrasound limited to patients with a positive D-dimer produced an overall positive scan yield of 22/606 (3.63%; 95% CI 2.42 to 5.44%) [116,117].

**Discussion**

The diagnostic accuracy of US for DVT varies according to the technique used. Optimal sensitivity is achieved by using duplex (proximal sensitivity 96%, distal sensitivity 71%, specificity 94%) or triplex US (proximal sensitivity 96%, distal sensitivity 75%, specificity 94%). Optimal specificity is achieved by using compression US alone (proximal sensitivity 94%, distal sensitivity 57%, specificity 98%). These findings suggest that compression US alone is probably the appropriate technique for most patients, if scanning is aimed simply at identifying proximal DVT. Most patients have a low probability of DVT, so optimal specificity is required to avoid generating
excessive false positive results. However, when evaluating patients at high risk of DVT, or if scanning aims to identify distal DVT, then duplex or triplex US will probably be the appropriate technique.

Beyond US technique we identified few study-level predictors of sensitivity or specificity. Sensitivity tended to be higher in more recent studies, probably reflecting developing technology and expertise. Sensitivity was surprisingly lower in studies where scans were interpreted by a radiologist. This may be because these studies were more likely to use techniques at an earlier stage in their development. Another cause could be that compression ultrasonography is the simplest technique, whereas Doppler and colour US techniques are more challenging and therefore more likely result in greater reporting variability. The association between proportion of proximal DVT and sensitivity is unsurprising as US has better sensitivity for proximal DVT. The association between DVT prevalence in the study cohort and sensitivity may be explained by a similar mechanism. Selection of a cohort with a higher prevalence of DVT is likely to involve selection of cases with more easily detectable (i.e. larger and more proximal) DVT. Prevalence has been shown to be associated with variation in the performance of other diagnostic tests for DVT. Heim et al [118] showed that D-dimer has poorer accuracy in cohorts with a higher prevalence of DVT, probably due to lower specificity.

We identified no studies to reliably estimate the diagnostic accuracy of repeat scanning in comparison to contrast venography. Our best estimate of the diagnostic value of repeat scanning is that, in unselected patients with suspected DVT, it will have a positive yield of 1.3%, of whom

---

Table 2: Pooled estimates of sensitivity and specificity stratified by US technique

<table>
<thead>
<tr>
<th>Technique</th>
<th>Sensitivity for all DVT</th>
<th>Sensitivity for proximal DVT</th>
<th>Sensitivity for distal DVT</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compression only, N = 22</td>
<td>90.3% (88.4 to 92.0) P &lt; 0.001</td>
<td>93.8% (92.0 to 95.3) P = 0.005</td>
<td>56.8% (49.0 to 66.4) P &lt; 0.001</td>
<td>97.8% (97.0 to 98.4) P = 0.01</td>
</tr>
<tr>
<td>Colour Doppler only, N = 5</td>
<td>81.7% (77.4 to 85.5) P &lt; 0.001</td>
<td>95.8% (85.7 to 99.5) P = 0.427</td>
<td>43.5% (23.2 to 66.5) P = 0.009</td>
<td>92.7% (89.7 to 95.1) P = 0.003</td>
</tr>
<tr>
<td>Continuous wave Doppler only, N = 16</td>
<td>81.1% (78.2 to 83.7) P &lt; 0.001</td>
<td>87.8% (84.7 to 90.5) P &lt; 0.001</td>
<td>41.8% (32.5 to 51.6) P = 0.015</td>
<td>84.0% (81.4 to 86.3) P &lt; 0.001</td>
</tr>
<tr>
<td>Triplex, N = 25</td>
<td>91.1% (89.0 to 93.0) P &lt; 0.001</td>
<td>96.4% (94.4 to 97.9) P &lt; 0.001</td>
<td>75.2% (67.7 to 81.6) P &lt; 0.001</td>
<td>94.3% (92.5 to 95.8) P &lt; 0.001</td>
</tr>
<tr>
<td>Duplex, N = 28</td>
<td>92.1% (90.7 to 93.5) P &lt; 0.001</td>
<td>96.5% (95.1 to 97.6) P &lt; 0.001</td>
<td>71.2% (64.6 to 77.2) P &lt; 0.001</td>
<td>94.0% (92.8 to 95.1) P &lt; 0.001</td>
</tr>
<tr>
<td>Others, N = 4</td>
<td>93.3% (88.8 to 96.4) P = 0.338</td>
<td>-</td>
<td>-</td>
<td>96.0% (92.2 to 98.2) P &lt; 0.001</td>
</tr>
</tbody>
</table>

95% CI in parentheses P-value = Chi-square test for heterogeneity

---

Figure 4
Funnel plot for sensitivity.

Figure 5
Funnel plot for specificity.
89% will be true positive and 11% false positive. A higher yield may be achieved by limiting repeat scanning to patients with a high clinical risk score and/or positive D-dimer. Whether these yields of positive scanning justify use of repeat scanning depend upon our estimates of the costs, benefits and risks of treating, or not treating, cases of DVT.

This study has some limitations that need to be considered. We did not search for unpublished data or studies published in languages other than English, French, Spanish, Italian or German. Studies of diagnostic tests are relatively easy to undertake, are often unfunded, and are not usually recorded on research registries. It is therefore unsurprising that systematic reviews of diagnostic test data rarely search for unpublished data [4] and that the potential effect of publication bias is unknown. Funnel plots for sensitivity and specificity were both asymmetrical. One possible explanation for this is that small studies reporting poor sensitivity or specificity may be less likely to be submitted or accepted for publication. If this is the case then the values for pooled sensitivity and specificity may represent over-estimates.

Despite undertaking meta-regression and stratifying results by US technique our findings were subject to significant unexplained heterogeneity. This heterogeneity is probably due to factors that were inadequately reported in the primary studies and therefore could not be explored in meta-regression. These factors include the characteristics of patients recruited (such as the prevalence of previous thromboembolism, obesity and co-morbidities), the training and experience of US operators, specific features of the US technique (such as the US frequency used), and any time delay between scanning and venography. These factors may have had a substantial influence upon sensitivity and specificity that will not have been identified in our analysis. Poor reporting also limited our ability to explore the effect of study design upon results. Use of blinding was often not described, studies rarely reported how uncertain or equivocal test results were handled, and the median prevalence of DVT in the cohorts (48%) suggests selective sampling of patients. These methodological weaknesses in the primary studies constitute a weakness in our meta-analysis.

The findings relating to repeat US scanning are subject to even greater limitations. Only a relatively small number of studies were identified and none compared repeat US to a venography in all cases. The potential benefit of repeat US is therefore very uncertain.

A potential clue to the influence of patient characteristics upon sensitivity and specificity is provided in a study by Wells et al [119], who reported their results stratified by the patient's clinical risk score into high, intermediate or low risk. Among patients with a high Wells score sensitivity (95% CI) was 91% (81 to 96) and specificity was 100% (77 to 100). Among patients with an intermediate Wells score sensitivity was 61% (46 to 74) and specificity was 99% (94 to 100). Among patients with a low Wells score sensitivity was 67% (42 to 85) and specificity was 98% (95 to 99). This suggests that US sensitivity may be dependent upon clinical probability of DVT and concurs with our finding that sensitivity was higher in cohorts with higher prevalence.

<table>
<thead>
<tr>
<th>Author &amp; date of publication</th>
<th>Group</th>
<th>Number (%) of initial scans positive</th>
<th>Number (%) of repeat scans positive</th>
<th>Number (%) of positive scans (initial or repeat) confirmed by venography</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heijboer, 1993 [109]</td>
<td>All patients</td>
<td>93/491 (19%)</td>
<td>7/397 (1.8%)</td>
<td>84/89 (94%)</td>
</tr>
<tr>
<td>Cogo, 1998 [110]</td>
<td>All patients</td>
<td>400/1702 (24%)</td>
<td>12/1252 (1.0%)</td>
<td>-</td>
</tr>
<tr>
<td>Sluzewski, 1991 [111]</td>
<td>All patients</td>
<td>67/174 (39%)</td>
<td>0/98 (0%)</td>
<td>-</td>
</tr>
<tr>
<td>Birdwell, 1998 [112]</td>
<td>All patients</td>
<td>63/405 (16%)</td>
<td>7/342 (2.0%)</td>
<td>23/28 (82%)</td>
</tr>
<tr>
<td>Birdwell, 2000 [113]</td>
<td>All patients</td>
<td>95/709 (13%)</td>
<td>9/521 (1.7%)</td>
<td>39/47 (83%)</td>
</tr>
<tr>
<td>Studies of unselected patients combined</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bernardi, 1998 [116]</td>
<td>Positive D-dimer</td>
<td>260/946 (27%)</td>
<td>5/88 (5.7%)</td>
<td>-</td>
</tr>
<tr>
<td>Kraaijenhagen, 2002 [117]</td>
<td>Positive D-dimer</td>
<td>391/1739 (22%)</td>
<td>17/518 (3.0%)</td>
<td>-</td>
</tr>
<tr>
<td>Studies of patients with a positive D-dimer combined</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wells, 1997 [114]</td>
<td>Intermediate Wells score</td>
<td>27/193 (14%)</td>
<td>3/166 (1.8%)</td>
<td>-</td>
</tr>
<tr>
<td>Tick, 2002 [115]</td>
<td>Intermediate or high Wells score &amp; positive D-dimer</td>
<td>300/531 (57%)</td>
<td>13/83 (15.7%)</td>
<td>-</td>
</tr>
</tbody>
</table>
The widespread current use of US to diagnose DVT is not based upon diagnostic cohort studies alone, but also upon management studies, in which cohorts of patients with negative US results are not treated, but followed up to identify evidence of missed thromboembolism. Studies of serial US [109-112,120], a single full-leg US [121-124], or US as part of a diagnostic algorithm [114,116,117,125-129] have shown low rates of thromboembolism during three to six month follow up. This suggests that, although our meta-analysis has shown that US does not have perfect sensitivity for DVT (especially distal thrombus), this does not translate into high rates of adverse outcome. This may be because application of a reasonably sensitive test to a population with low disease prevalence will result in a high negative predictive value, or it may be because DVT that are missed by ultrasound have a relatively benign natural history.

Conclusion
US has high sensitivity for proximal DVT, modest sensitivity for distal DVT and high specificity. Optimal sensitivity, particularly for distal DVT, is achieved by using duplex or triplex US, while optimal specificity is achieved by using compression US alone. US sensitivity appears to be higher in cohorts with higher DVT prevalence. However, these findings are subject to substantial unexplained heterogeneity and should be interpreted with caution. Evaluation of repeat US has been very limited and its potential benefit is very uncertain.

List of abbreviations
US = ultrasound
DVT = deep vein thrombosis
CI = confidence interval

Competing interests
The authors are not aware of any conflicts of interest relating to this article. The United Kingdom Health Technology Assessment R&D Programme funded this project (reference number 02/03/01). The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the UK Department of Health.

Authors’ contributions
SG designed the study, selected articles for inclusion, undertook meta-analysis, and wrote the paper. FS selected articles for inclusion, assisted with data extraction, and helped write the paper. ST and EvB helped to design the study, extracted data, assessed study quality, and helped write the paper. AS assisted with meta-analysis, undertook meta-regression, created the funnel plots, and helped write the paper. All the authors have seen and approved the final draft.

Acknowledgements
We thank Angie Ryan for her help with the literature searches and Kathryn Paulucy for clerical assistance.

References


Pre-publication history
The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2342/5/6/prepub