Clinical experience and pre-test probability scores in the diagnosis of pulmonary embolism


From the Canterbury Respiratory Research Group and 1Department of Medicine, Christchurch School of Medicine and Health Sciences, University of Otago, New Zealand

Received 16 October 2002 and in revised form 9 December 2002

Summary

Background: The Geneva and Wells pre-test probability scores are intended to replace empirical assessment of patients with suspected pulmonary embolism (PE). The effect of clinical experience on the inter-rater variability of these scores, and on empirical judgement, is unknown.

Aim: To determine whether medical staff appointment grade affects the inter-rater variability of these pre-test probability scores, or empirical assessment, in patients with suspected PE.

Design: Questionnaire survey.

Methods: Doctors were grouped by grade (mean number of years since graduation ± SEM): house officers 0.7 ± 0.2, registrars 6.3 ± 0.6, consultants 25 ± 4 and applied pre-test probability scores to actual case scenarios.

Results: The Geneva score was the most consistent method of determining pre-test probability and was unaffected by clinical experience (Geneva $\kappa = 0.73$, Wells $\kappa = 0.38$, empirical $\kappa = 0.23$, $p < 0.001$).

With empirical judgement, inter-rater variability was inversely proportional to clinical experience (house officers $\kappa = 0.37$, registrars $\kappa = 0.24$, consultants $\kappa = 0.16$, $p < 0.05$).

Discussion: The Geneva score was the least variable method and can be applied by junior or senior doctors. Using empirical judgement, junior doctors were more likely to agree on the pre-test probability of PE than were their more senior colleagues. This may imply that as physicians gain experience, they recognize that the diagnosis of PE can be difficult to assess and are reluctant to exclude it on clinical grounds.

Introduction

The use of history and clinical signs in the diagnosis of pulmonary embolism (PE) is notoriously inaccurate, and clinical judgement alone can rarely confirm or refute the diagnosis of PE with certainty. However, initial clinical assessment determines the need for imaging to objectively confirm or exclude PE.

The pre-test probability (PTP) of disease represents a formal assessment of the likelihood of disease before a confirmatory test, or investigation, is performed. The Prospective Investigation Of Pulmonary Embolism Diagnosis (PIOPED) investigators demonstrated that for any given Ventilation Perfusion (VQ) scan result, the higher the PTP of PE, the more likely it was that PE would be confirmed as being present.

Bayes’ theorem states that the probability of disease following the interpretation of a diagnostic investigation (the post-test probability) is determined by two factors. One, the probability of disease prior to carrying out the investigation (the pre-test probability) based on the information already available. Two, the characteristics of the test. For example, if in a patient with suspected PE,
the PTP of PE is low, estimated using either of
the three methods studied, and the D-dimer is not
 elevated, the post-test probability of PE is very low
 and PE is thus highly unlikely. Furthermore, sub-
 sequent investigation(s) are unlikely to change the
 management of this patient.

The Geneva7 and Wells8 scores are clinical pre-
 test probability scores derived from large trials that
 sought to determine the clinical signs and symp-
 toms that reliably predict the diagnosis of PE. The
 variables used in the calculation of these two scores
 are presented in Tables 1 and 2, respectively. These
 scores may be used to define the PTP of PE as low,
 moderate, or high,7,8 with the prevalence of PE
 increasing across the three groups. While similar in
 concept, some criteria used to calculate the Wells8
 score (e.g. ‘Alternative diagnosis less probable than
 PE’) are arguably more subjective than those used to
 calculate the Geneva7 score. Although the determi-
 nation of the PTP alone would never be enough to
 exclude or confirm PE, in combination with bedside
tests such as D-dimer, they may help triage those
 who do not need further testing or imaging, con-
ferring considerable resource savings. In order for
 such a strategy to be used in clinical practice, both
 the assessment of the PTP of PE and the measure-
 ment of D-dimer need to fulfil two criteria. Firstly,
 they must be accurate, and secondly, they must be
 reliable. One measure of reliability is inter-rater
 variability.

The need to define the inter-rater variability of
 clinical scoring methods, such as the Geneva and
 the Wells scores, before they are applied by large
 numbers of clinicians, has been previously identi-
fied.8 Also, it is often assumed that less experienced
doctors are not as capable of determining the pre-
test probability of PE9 We sought to determine
 whether the clinical experience of the assessing
 doctor affects the inter-rater variability of these
 clinical scores and empirical assessment.

Table 1 Criteria for the calculation of the Geneva score

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Points value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>60–79 years</td>
<td>+1</td>
</tr>
<tr>
<td>≥ 80 years</td>
<td>+2</td>
</tr>
<tr>
<td>Previous DVT or PE</td>
<td>+2</td>
</tr>
<tr>
<td>Recent surgery (&lt;4 weeks ago)</td>
<td>+3</td>
</tr>
<tr>
<td>Heart rate &gt; 100 bpm</td>
<td>+1</td>
</tr>
<tr>
<td>PaCO2 &lt; 35 mmHg</td>
<td>+2</td>
</tr>
<tr>
<td>35–39 mmHg</td>
<td>+1</td>
</tr>
<tr>
<td>PaO2 &lt; 49 mmHg</td>
<td>+4</td>
</tr>
<tr>
<td>49–59 mmHg</td>
<td>+3</td>
</tr>
<tr>
<td>60–71 mmHg</td>
<td>+2</td>
</tr>
<tr>
<td>72–82 mmHg</td>
<td>+1</td>
</tr>
<tr>
<td>Chest X-ray</td>
<td></td>
</tr>
<tr>
<td>Band atelectasis</td>
<td>+1</td>
</tr>
<tr>
<td>Elevation of hemidiaphragm</td>
<td>+1</td>
</tr>
</tbody>
</table>

Total score: <5 indicates a low probability of PE; 5–8 indicates a moderate probability of PE; >8 indicates a high probability of PE.

Table 2 Criteria for the calculation of the Wells score

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Points value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical signs of DVT</td>
<td>+3</td>
</tr>
<tr>
<td>Alternative diagnosis less probable than PE</td>
<td>+3</td>
</tr>
<tr>
<td>Heart rate &gt; 100 bpm</td>
<td>+1.5</td>
</tr>
<tr>
<td>Immobilization or surgery &lt;4 weeks ago</td>
<td>+1.5</td>
</tr>
<tr>
<td>Previous DVT or PE</td>
<td>+1.5</td>
</tr>
<tr>
<td>Haemoptysis</td>
<td>+1</td>
</tr>
<tr>
<td>Cancer</td>
<td>+1</td>
</tr>
</tbody>
</table>

Total score <2 indicates a low probability of PE; 2–6 indicates a moderate probability of PE; >6 indicates a high probability of PE.

Methods

Study setting and design

This study was performed at Christchurch Hospital,
 a 600-bed tertiary institution serving a population
 of 375 000. The first thirty case scenarios were
 selected from a database of all consecutive patients
 investigated for PE by VQ scan or Spiral Computer
 Tomography (Spiral CT) between August 2000 and
 August 2001 that were able to have PE diagnosed or
 excluded by objective criteria. PE was judged to have been objectively
diagnosed with a high probability VQ scan
 according to PIOPED criteria6 (11 cases), and/or a
 vascular filling defect on Spiral CT (3 cases). PE was
 excluded by a normal VQ scan (14 cases). PE also
 was considered excluded by a normal CT scan
 accompanied by a normal bilateral leg ultrasound
 (1 case), or normal D-dimer (1 case), and absence of
 a diagnosed thromboembolic event over a mini-
mum of 3 months (both cases). The prevalence of
 PE in this cohort was 47%. Of the 30, 16 were
 female and the mean ± SD age was 58 ± 15.0 years.

Thirty case scenarios were constructed from the
 information contained in the medical record for that
 admission. Each case scenario consisted of a brief
 history, examination findings, results of basic initial
investigations (including full blood count, biochemistry, and arterial blood gases), and included a copy of the formal chest X-ray report. If performed at admission, the electrocardiogram (28/30 cases) and chest X-ray films (23/30 cases) were available for inspection.

**Study participants**

We invited 20 consultant physicians, 31 medical registrars, and 33 medical house officers who participated in the acute medical rota to take part. We received positive replies from 12, 14, and 13 of them, respectively. The first ten from each medical staff grade who completed the questionnaire within the required timeframe were included.

**Format of questionnaire**

Participants were blinded to the results of the tests used to diagnose PE. After reviewing each case scenario, participants completed the Wells, and Geneva scores (Tables 1 and 2), and their empirical assessment of pre-test probability. The participants were unaware of what raw values for the respective scoring systems defined a patient’s PTP of PE as low, medium or high, so as not to influence their empirical assessment unduly. They indicated their empirical assessment of the likelihood of PE, similar to that used in the PIOPED study, by placing a cross on a Visual Analogue scale (VAS), where 0 = ‘PE is definitely not the cause of this patient’s symptoms’ and 100 = ‘PE is definitely the cause of this patient’s symptoms’. All participants reviewed all 30 cases, and listed their professional qualifications with the number of years since they graduated.

**Statistical analysis**

Raw scores from questionnaires were converted to low, medium, and high pre-test probability, according to the methods of the original authors (Tables 1 and 2). The distance from the point of origin (0) of the visual analogue scale was measured. Scores <20% were defined as low pre-test probability. Scores from 20–80% were defined as moderate pre-test probability. Scores >80% were defined as high pre-test probability.10

Inter-rater variability was assessed by calculating an extended form of the kappa (κ) statistic to allow for more than two observers11. Kappa statistic scores vary from +1.0 (perfect agreement) to 0.0 (no agreement) and test the null hypothesis of no agreement between raters. The standard errors calculated for each kappa were pooled to enable statistical comparisons between kappas.11

The negative predictive value (NPV) of a low pre-test probability score was calculated by aggregating the probability assessment for each clinician through the 30 cases and calculating the prevalence of PE in those cases that were labelled as low probability for each scoring method. The NPV of the PTP method was used (rather than positive predictive value) as it represents the original authors’ most useful finding, i.e. by indicating when a patient can safely have PE excluded by non-invasive means. The NPV between categories of doctor using each method of determining the PTP of PE was compared by one-way ANOVA, having confirmed the normality of these measures for parametric testing. The NPV between each of the methods used to determine the PTP of PE was compared by repeated measures ANOVA.

**Results**

**Study doctors**

Ten house officers, ten registrars, and ten consultants participated. The number of years since graduation was used as a surrogate for clinical experience (Table 3).

**Inter-observer variability**

The Geneva PTP score was the most consistent (i.e. the least variable) of the three methods used to determine the PTP of PE (κ for all experience groups: Geneva = 0.73, Wells = 0.38, empirical = 0.23). This finding was consistent across all three categories of doctors studied (p<0.001, Figure 1).

The more junior the clinician, the more likely they were to agree with their peers as to the PTP of

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Clinical experience and characteristics of study participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic</td>
<td>House officers (n = 10)</td>
</tr>
<tr>
<td>Years since graduation (mean ± SEM)</td>
<td>0.7 ± 0.2</td>
</tr>
<tr>
<td>Male/female</td>
<td>3/7</td>
</tr>
<tr>
<td>Number with postgraduate qualifications</td>
<td>0</td>
</tr>
</tbody>
</table>
PE when using their empirical judgement (house officers $\kappa = 0.37$, registrars $\kappa = 0.24$, consultants $\kappa = 0.16$, $p < 0.05$).

Negative predictive value

The seniority of the assessing clinician did not effect the NPV of a low PTP of PE, regardless of the method used. Empirical assessment of the PTP of PE had a significantly better NPV than either the Wells score ($p = 0.004$) or the Geneva score ($p = 0.001$) (Table 4). When deciding that a patient had a low PTP of PE using their empirical judgement, as opposed to one of the two clinical scoring methods studied, clinicians were more conservative (Figure 2).

Discussion

The Geneva score was the most consistent of the three methods studied at determining the pre-test probability of PE (Geneva $\kappa = 0.73$, Wells $\kappa = 0.38$, empirical $\kappa = 0.23$, $p < 0.001$, Figure 1). The inter-rater variabilities of both the Geneva and the Wells scores were not affected by the experience of the assessing clinician. With increasing experience, clinicians were less likely to agree with their peers when assessing the pre-test probability of PE empirically. This suggests that as their level of experience increases, clinicians are more aware that the diagnosis of PE is difficult to confirm or refute clinically.

Empirical assessment was the most accurate of the three methods studied, as determined by the NPV of a low PTP of PE. However, it was also the most conservative and least consistent method of assessing the PTP of PE. The advantage of using one of these clinical scoring methods for calculating the PTP of PE, compared to empirical assessment, is that they standardize the expression of clinical findings in an explicit manner. Theoretically therefore they should lead to fewer misunderstandings between colleagues when they report their findings to each other.

Wells et al. have shown that a low PTP in combination with a negative SimpliRED D-dimer test has a NPV when excluding PE of 98%. The advantage of using either the Wells or Geneva scores to assess the PTP of PE, as compared to empirical assessment, is that more patients would

<table>
<thead>
<tr>
<th>Table 4</th>
<th>The negative predictive value* of low probability scores according to the method used to calculate the PTP and clinical experience</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>House officers</td>
</tr>
<tr>
<td>Wells</td>
<td>109/143 (76%)</td>
</tr>
<tr>
<td>Geneva</td>
<td>96/117 (82%)</td>
</tr>
<tr>
<td>Empirical (VAS)</td>
<td>58/68 (85%)</td>
</tr>
</tbody>
</table>

*Calculated as the number of low probability scores when PE was excluded, divided by the total number of low probability scores.
be classified as having a low PTP of PE. As outlined, when combined with measurement of D-dimer, this would allow PE to be confidently excluded. Hence fewer patients would be subjected to the morbidity and mortality associated with investigations such as VQ scanning, Spiral CT, and pulmonary angiography.

The current study has several limitations. Firstly, the Geneva and Wells scores were designed to assess the PTP of PE before the lung scan was performed. In this study the case scenarios were presented retrospectively as written case scenarios, as opposed to the patient being in front of the clinician. However, the scenarios were consecutive actual cases, and care was taken to reproduce all the clinical information that would have been available at the time. Cases were selected only when enough information from the medical records could be obtained, and when PE could be ruled in or out by objective criteria. We do not believe a prospective study with so many observers from different groups of clinical experience would have been possible. Because all participants were presented with the same clinical information, our measures of inter-rater variability are greater than would be observed if the study were repeated prospectively. In addition although it could not be avoided, and participants were not given the Wells and Geneva scoring systems, the questions in the scores were asked before they marked their empirical assessment and this may have influenced their judgement of the likelihood of PE.

Finally, the NPV of a low PTP score is likely to be underestimated due to the high prevalence of PE (47%) in this cohort. Nonetheless, results were similar in another study that prospectively tested the Wells PTP score. The prevalence of PE in the database from which these cases were drawn is 24%. This is similar to most published series, and we believe that the high prevalence of PE in this sub-group was not attributable to any systematic selection bias; also the kappa values would not be influenced by the prevalence of PE. We agreed with the original authors of the scores that the most clinically relevant use of the PTP scoring methods were in order to identify patients that could potentially have PE excluded by non-invasive means, such as in combination with a negative D-dimer. The positive predictive value of clinical scores in PE have never been sufficient to diagnose PE, and this is illustrated by only 28% and 19% of PE positive patients scoring high probability in the original studies of the Wells and Geneva scoring systems, respectively.

In order for PTP scores to be widely adopted they need to be both accurate and reliable. It follows that they should be simple to use and objective. The ‘Alternative diagnosis less probable than PE’ used in calculating the Wells score is subjective, and this is reflected in the Wells score’s high inter-rater variability.

In conclusion, PTP scores in combination with non-invasive diagnostic methods may be useful in helping to exclude PE. Of the three methods studied, the Geneva score gave the most consistent results across participants. The inter-rater variability and accuracy of the Geneva PTP score was unaffected by clinical experience. From these results, the Geneva score would thus be the preferred method for both junior and senior doctors, in combination with non-invasive diagnostic methods, to exclude PE.

References