The PRONE score: an algorithm for predicting doctors' risks of formal patient complaints using routinely collected administrative data

Matthew J Spittal,[1] Marie M Bismark,[1] David M Studdert[2,3]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2014-003834).

[1]Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Victoria, Australia
[2]Center for Health Policy/PCOR, Stanford University Medical School, Stanford, California, USA
[3]Stanford Law School, Stanford, California, USA

Correspondence to
Dr David M Studdert, Center for Health Policy, Stanford Medical School, 117 Encina Commons, Stanford, CA 94305, USA; studdert@stanford.edu

Received 2 December 2014
Revised 12 March 2015
Accepted 20 March 2015
Published Online First
8 April 2015

Linked

▸ http://dx.doi.org/10.1136/bmjqs-2015-004337

CrossMark

To cite: Spittal MJ, Bismark MM, Studdert DM. *BMJ Qual Saf* 2015;**24**:360–368.

ABSTRACT

Background Medicolegal agencies—such as malpractice insurers, medical boards and complaints bodies—are mostly passive regulators; they react to episodes of substandard care, rather than intervening to prevent them. At least part of the explanation for this reactive role lies in the widely recognised difficulty of making robust predictions about medicolegal risk at the individual clinician level. We aimed to develop a simple, reliable scoring system for predicting Australian doctors' risks of becoming the subject of repeated patient complaints.

Methods Using routinely collected administrative data, we constructed a national sample of 13 849 formal complaints against 8424 doctors. The complaints were lodged by patients with state health service commissions in Australia over a 12-year period. We used multivariate logistic regression analysis to identify predictors of subsequent complaints, defined as another complaint occurring within 2 years of an index complaint. Model estimates were then used to derive a simple predictive algorithm, designed for application at the doctor level.

Results The PRONE (Predicted Risk Of New Event) score is a 22-point scoring system that indicates a doctor's future complaint risk based on four variables: a doctor's specialty and sex, the number of previous complaints and the time since the last complaint. The PRONE score performed well in predicting subsequent complaints, exhibiting strong validity and reliability and reasonable goodness of fit (*c*-statistic=0.70).

Conclusions The PRONE score appears to be a valid method for assessing individual doctors' risks of attracting recurrent complaints. Regulators could harness such information to target quality improvement interventions, and prevent substandard care and patient dissatisfaction. The approach we describe should be replicable in other agencies that handle large numbers of patient complaints or malpractice claims.

INTRODUCTION

Medicolegal agencies—such as malpractice insurers, medical boards and complaints handling bodies—are essentially reactive regulators: they deal with the aftermath of care that has gone badly. This posture has confined them largely to the sidelines of the patient safety movement. Clinicians eye medicolegal processes with scepticism and fear, and sometimes with outright disdain. The ex post nature of these processes, coupled with their focus on provider fault, is derided as antithetical to quality improvement efforts focused on prevention and systemic causes of harm.[1]

Part of the explanation for medicolegal agencies' limited role lies in their inability to make reliable predictions about which clinicians will experience complaints or claims. The conventional wisdom is that these events cannot be predicted at the individual practitioner level with acceptable levels of accuracy. Numerous studies have tried,[2–12] with limited success.

In a recent national study[13] of formal patient complaints against Australian doctors lodged with state regulators, we found that 3% of doctors (or 18% of doctors with at least one complaint) accounted for half of all complaints lodged over a 10 year period. We proposed and tested a new method for identifying doctors at high risk of incurring repeated medicolegal events. Among doctors who had already incurred one or more complaints, we found that risks of incurring more complaints in the near term varied dramatically—from less than 10% to more than 80%—depending on certain observable characteristics, or risk factors. However, one limitation of the method we deployed, recurrent event survival analysis, is its complexity. Few health regulators are likely to have the technical capacity to replicate this approach, much less incorporate it into their front-line case management activities. Relatedly, although survival analysis accounts well for changing baseline risk over time, it does not generally provide estimates of these risks that can be easily integrated into a risk algorithm.

In this study, we extended our analysis of patient complaints to create a simple predictive algorithm. Our immediate objective was to provide health complaints commissions in Australia with a tool that would allow them to reliably estimate practitioners' future risk of complaints in ways that could support proactive intervention. More generally, we sought to demonstrate the feasibility of using routinely collected administrative data to construct a risk calculator for predicting medicolegal events.

METHODS

Setting

Health complaints commissions (Commissions) are statutory agencies established in each of Australia's six states and two territories. Commissions have responsibility for receiving and resolving patient complaints about the quality of healthcare services. Patients or their advocates must initiate complaints in writing, but the process is free and legal representation is optional.[14]

Outside the clinic or hospital where care is rendered, Commissions are the primary avenue of redress for patients dissatisfied with the quality of care they have received. Plaintiffs' lawyers in Australia will rarely take cases unless they have first proceeded through Commission processes (although the vast majority of complaints do not go on to become malpractice claims). At least 10 other Organisation for Economic Co-operation and Development countries—including Austria, Finland, Israel and New Zealand—have similar bodies.[15] [16] In the UK, the closest analogue is the Parliamentary and Health Service Ombudsman.

Commissions in all Australian states except South Australia participated in the study. In 2011, these seven jurisdictions had 21 million residents and 90% of the nation's 70 200 practicing doctors.[17]

Data

Between May 2011 and February 2012 we collected data on-site at Commission offices in each participating state and territory. Complaints against doctors were identified by querying the Commissions' administrative data systems. The filing period of interest spanned 12 years and differed slightly by jurisdiction: 2000–2011 for the Australian Capital Territory, the Northern Territory, Queensland, Tasmania and Victoria; 2000–2010 for Western Australia; and 2006–2011 for New South Wales. (Changes to the data system in New South Wales necessitated a narrower sampling window there.) As described elsewhere,[13] we supplemented administrative data from the Commissions with additional doctor-level variables obtained from a database on Australian doctors held by AMPCo Direct, a subsidiary of the Australian Medical Association.

All data were combined in a complaint-level analytical data set, in which multiple complaints against a single doctor could be observed.

Variables

Our unit of analysis was the complaint. The primary outcome was the occurrence of a 'subsequent' complaint, defined as a complaint that occurred soon after a prior one. Specifically, a complaint against a doctor within 2 years of the most recent complaint against that doctor was coded '1'; a complaint that was not followed by another during the 2-year window, was coded '0'. To allow sufficient follow-up time, complaints that occurred within 24 months of the end of a jurisdiction's data period (2011 in all but one jurisdiction) were included in the analysis as subsequent complaints but not as index complaints. We chose a 2-year window based on our discussions with regulators in Australia and New Zealand, who indicated that this is generally the time horizon of interest to them in their efforts to separate complaint-prone practitioners from respondents whose appearance reflects merely 'baseline' risks of being complained against.

The set of predictors we examined consisted of variables measured at the doctor level (specialty, age, sex, location of practice) and the complaint level (number of previous complaints, time since previous complaint, complaint issue). Doctor age and all complaint level variables were coded as time-varying variables, meaning their values related to the most recent complaint. Specialty was classified into 13 categories, based on those promulgated by the Medical Board of Australia.[18] Doctor age was coded as <35 years or 35–65 years. (Doctors older than 65 years were excluded because of possible bias due to unobserved censoring: an absence of further complaints may have been due to retirement). Principal practice address was classified as urban or rural, based on the location of its postcode within a standard geographical classification system.[19] Subsequent complaints were categorised by time intervals (<6 months after the most recent complaint, 6 months to <1 year, 1–2 years).

Complaint issue in the most recent complaint was coded into two categories: complaints in which the primary issue related to clinical care (treatment; diagnosis; medication; hygiene or infection control; discharge or transfer; other clinical care issues) and complaints with other primary issues (communication;

costs or billing, medical records, certificates or reports; access and timelines, sexual contact or relationship; rough or painful treatment; confidentiality or information privacy; breach of conditions; grievance handling; discrimination; other issues).

Statistical model

We used multivariate logistic regression analysis to estimate odds of a subsequent complaint within 2 years. The predictors were number of previous complaints; time since previous complaint; complaint issue; and specialty, age, sex and practice location of the doctor against whom the complaint was made. Cluster-adjusted robust SEs were calculated to account for multiple observations pertaining to single doctors.

We estimated and compared three models: Model 1 used all available predictors; Model 2 focused on a subset of predictors that tend to be the most straightforward for regulators to obtain from routine operational data; and Model 3 had only one predictor, the variable most strongly associated with the outcome. Our goal was to identify the model that best balanced model fit against parsimony. The primary basis for comparing the models' performance was the *c*-statistic (or area under the receiver operating characteristic (ROC) curve). We computed *c*-statistics, adjusting them for optimism to guard against overfitting using a bootstrap sampling approach.[20] (Details of the adjustment-for-optimism method of cross validation are provided in the online supplementary appendix) We tested whether the unadjusted *c*-statistic for Model 1 was significantly different from Model 2, and whether Model 2 was significantly different from Model 3. Finally, as a sensitivity analysis, we repeated our analysis using only complaints over clinical care. All analyses were conducted using Stata V.13.1.[21]

Construction of the PRONE score

We sought to design a scoring system with the following features: (1) each risk factor is assigned a prespecified number of points, expressed in whole numbers; (2) the assigned points are proportionate to the ORs from the underlying model and discriminate well between differences in ORs; (3) points assigned to individual risk factors sum to produce a total risk score; (4) total scores are expressed across a range of approximately 20 values; and (5) the risk score is designed to be calculated anew at the doctor level each time a new complaint is lodged.

Once the preferred model was chosen, we assigned points to each predictor, indexing point values to the values of the model coefficients (the log ORs). The scoring system features we sought were best achieved by multiplying each coefficient by 3.7 and rounding to the closest integer. Summation of points across the risk factors produced a score ranging from 0 to 21. We dubbed this the PRONE (Predicted Risk Of New Event) score.

Analysis of performance of PRONE score

We assessed the performance of the PRONE score in three ways. First, to determine whether precision was lost in transforming coefficients from the multivariate model to crude integers for the PRONE score, we plotted ROC curves for the model coefficients and the scores and compared them. Second, to assess calibration of the PRONE score (ie, how closely the predicted probability of a subsequent complaint reflected actual risk[22]), we calculated and compared the observed and expected number of new complaints at each value of the PRONE score. The expected number of complaints was estimated by calculating P(Complaint)=exp($\beta_1+\beta_2$(PRONE score))/(1+exp($\beta_1+\beta_2$(PRONE score)) and multiplying these probabilities by the total number of doctors with each PRONE score. The Hosmer-Lemeshow χ^2 statistic was used to assess whether expected scores differed from observed scores.[23] Finally, we calculated sensitivity and specificity of the PRONE score at three levels that have policy relevance, in the sense that they represent potential thresholds for regulatory intervention.

RESULTS

There were 13 849 complaints in the analytical sample. A total of 8424 doctors were the subject of an initial complaint to Commissions and 31% (2586/8424) of them had subsequent complaints. There were a total of 6427 subsequent complaints, 70% (4488/6427) of which occurred within 2 years.

Sample characteristics

Table 1 describes characteristics of the doctors and complaints in our study sample. Sixty per cent of the complaints addressed clinical aspects of care, most commonly concerns with treatment (39%), diagnosis (16%) and medications (8%). About a fifth of complaints addressed communication issues, including concerns with the attitude or manner of doctors (13%), and the quality or amount of information provided (6%).

Nearly half of the doctors complained against were general practitioners and 15% were surgeons. Seventy-nine per cent were male and 80% were 35–65 years of age. On average, 398 days (SD 497 days) elapsed between the index and subsequent complaints. There was a trivial amount of missing data (<1%) for all variables except age (14%).

Choice of multivariate prediction model

All of the variables considered in Model 1, except the variable indicating the previous complaint related to clinical care, were statistically significant predictors of subsequent complaints, and this model had a *c*-statistic of 0.69 after adjusting for optimism (table 2). Model 2 dropped doctor age, the clinical care variable and practice location (a variable many regulators do not routinely collect). All four remaining variables were

Table 1 Characteristics of complaints and doctors who were the subject of the complaints

	n*	Per cent
Issue in complaints, n=13 849		
Clinical care	8352	60
Treatment	5407	39
Diagnosis	2251	16
Medication	1083	8
Hygiene/infection control	104	0.8
Discharge/transfer	53	0.4
Other clinical care	81	0.6
Communication	2909	21
Attitude or manner	1849	13
Information	790	6
Consent	416	3
Other communication	30	0.2
Costs or billing	970	7
Medical records, certificates or reports	891	6
Access and timeliness	854	6
Sexual contact or relationship	422	3
Rough or painful treatment	319	2
Confidentiality or information privacy	281	2
Breach of conditions	186	1
Grievance handling	129	0.9
Discrimination	54	0.4
Other	112	0.8
Characteristics of doctors, n=8424		
Gender		
Male	6667	79
Female	1676	20
Missing	71	0.8
Age		
22–34 years	464	6
35–65 years	6756	80
Missing	1204	14
Specialty		
General practice	3972	49
Surgery	1182	15
Orthopaedic surgery	329	4
General surgery	296	4
Plastic surgery	140	2
Other surgery	417	5
Internal medicine	934	11
Obstetrics and gynaecology	416	5
Psychiatry	504	6
Anaesthesia	314	4
Ophthalmology	188	2
Radiology	144	2
Dermatology	129	2
Other specialties	349	4
Location of practice		
Rural	1948	23
Urban	6378	77

*Complaint issues sum to more than 100% because some complaints involved multiple issues.

statistically significant predictors of subsequent complaints, and discrimination in this model (*c*-statistic=0.70) was very close to Model 1's, although there was a statistically significant difference between the *c*-statistics for the two models (p=0.018).

Model 3 consisted only of the number of prior complaints variable, which was clearly the strongest predictor of subsequent complaints in models 1 and 2. Although this predictor on its own showed reasonable discrimination (*c*-statistic=0.66), there was strong evidence that its fit was inferior to the other two models (p<0.0001 in both comparisons).

Therefore, considering parsimony, discriminative ability and potential practicality, we selected Model 2 as the basis for the PRONE score algorithm. The far right column in table 2 shows the PRONE score points assigned to each predictor, based on the corresponding coefficients from Model 2. Other commonly used approaches—for instance, assigning a point for every 100% increase in ORs—produced similar results (data not shown).

To test the robustness of our results, we refit our models using only those complaints that related to clinical care. The *c*-statistics for these three models were 0.69, 0.70 and 0.65, respectively, and the ORs were similar to those reported above (see Models 4, 5 and 6 in online supplementary appendix table S1).

Performance of PRONE score

A comparison of ROC curves for the PRONE score and the coefficients estimated in Model 2 showed almost perfect overlap, suggesting little if any discrimination was lost in transforming the model's coefficients to the scoring system (see online supplementary appendix figure S1).

Figure 1 shows, at each level of the PRONE score, the number of complaints predicted by the score alongside the number actually observed. The Hosmer-Lemeshaw χ^2 statistic was non-significant, $\chi^2(20)=21.83$, p=0.35, suggesting that PRONE scores corresponded closely to actual risk. The distribution of observed and expected complaints by PRONE score indicates some skewness, with scores centred around values of 4 and 5.

PRONE score predictions

Table 3 shows the frequency and risk of subsequent complaints within seven PRONE score groups. Risk increased monotonically with PRONE score. For example, among 2000 doctors whose index complaint scored between 0 and 2, 285 of them actually had a subsequent complaint, a 2-year risk of 14.2%. Among 221 doctors with PRONE scores between 15 and 17, 194 had subsequent complaints, a risk of 87.8%.

Thresholds for intervention

Decisions regarding suitable score thresholds for intervention involve a trade-off between sensitivity and

Table 2 Logistic regression models for risk of complaints within 2 years, and Predicted Risk Of New Event (PRONE) scoring system, derived from the ORs in model 2

	Model 1 OR (95% CI)	Model 2 OR (95% CI)	Model 3 OR (95% CI)	PRONE score
Complaint number				
1 (ref)	1.00	1.00	1.00	0
2	1.29 (1.11 to 1.49)	1.35 (1.20 to 1.51)	1.82 (1.66 to 2.00)	1
3	1.85 (1.56 to 2.20)	1.91 (1.65 to 2.22)	2.76 (2.43 to 3.14)	2
4	2.48 (2.01 to 3.07)	2.64 (2.18 to 3.20)	3.98 (3.35 to 4.72)	4
5	3.29 (2.51 to 4.31)	3.41 (2.67 to 4.35)	5.36 (4.27 to 6.73)	5
6	4.35 (3.11 to 6.10)	4.30 (3.15 to 5.87)	6.88 (5.11 to 9.25)	5
7	4.76 (3.08 to 7.34)	5.01 (3.35 to 7.49)	8.51 (5.85 to 12.4)	6
8	4.44 (2.78 to 7.08)	4.79 (3.08 to 7.43)	7.98 (5.23 to 12.2)	6
9	6.51 (3.38 to 12.53)	6.73 (3.68 to 12.3)	11.1 (6.19 to 19.8)	7
10+	18.89 (9.76 to 36.56)	18.3 (10.2 to 32.8)	33.8 (19.1 to 59.7)	11
Doctor's specialty				
Anaesthesia (ref)	1.00	1.00		0
Radiology	1.00 (0.47 to 2.12)	1.06 (0.51 to 2.22)		0
Other specialties	0.97 (0.63 to 1.49)	1.12 (0.76 to 1.64)		0
Internal medicine	1.40 (1.04 to 1.88)	1.50 (1.12 to 1.99)		1
Ophthalmology	1.36 (0.94 to 1.96)	1.58 (1.12 to 2.23)		2
General practice	1.61 (1.23 to 2.10)	1.75 (1.35 to 2.26)		2
Psychiatry	1.94 (1.44 to 2.62)	2.00 (1.49 to 2.68)		3
Orthopaedic surgery	2.02 (1.49 to 2.74)	2.26 (1.68 to 3.03)		3
Other surgery	2.11 (1.56 to 2.86)	2.30 (1.72 to 3.09)		3
General surgery	2.11 (1.51 to 2.95)	2.46 (1.79 to 3.38)		3
Obstetrics and gynaecology	2.36 (1.73 to 3.23)	2.51 (1.85 to 3.39)		3
Dermatology	2.73 (1.89 to 3.96)	3.15 (2.16 to 4.59)		4
Plastic surgery	3.98 (2.84 to 5.57)	4.44 (3.21 to 6.13)		6
Time since previous complaint				
1–2 years (ref)	1.00	1.00		0
6 months to 1 year	1.23 (1.04 to 1.47)	1.20 (1.02 to 1.43)		1
Less than 6 months	1.68 (1.44 to 1.95)	1.77 (1.53 to 2.04)		2
Doctor's sex				
Female (ref)	1.00	1.00		0
Male	1.45 (1.27 to 1.66)	1.51 (1.33 to 1.71)		2
Doctor's age				
22–34 years (ref)	1.00			
35–65 years	1.41 (1.10 to 1.82)			
Location of practice				
Urban (ref)	1.00			
Rural	1.12 (1.01 to 1.24)			
Complaint issue				
Other issue (ref.)	1.00			
Clinical care	1.02 (0.9 to 1.16)			
C-statistic (adjusted for optimism)	0.69	0.70	0.66	

specificity, and are likely to be influenced by the effectiveness, intrusiveness and cost of the intervention. While the exact nature of such interventions is beyond the scope of this paper, we use three hypothetical examples to illustrate the nature of these trade-offs. (We note that some of these interventions may fall outside current statutory powers of Australian Commissioners.)

Table 3 depicts three possible interventions: (1) advising doctors in writing that they are at risk of a future complaint; (2) compelling doctors to undertake a continuing medical education course on a topic that addresses issues commonly arising in their complaint profile; and (3) referral to a regulator (eg, medical board) for further action. For a low cost, relatively unintrusive intervention, such as an informational

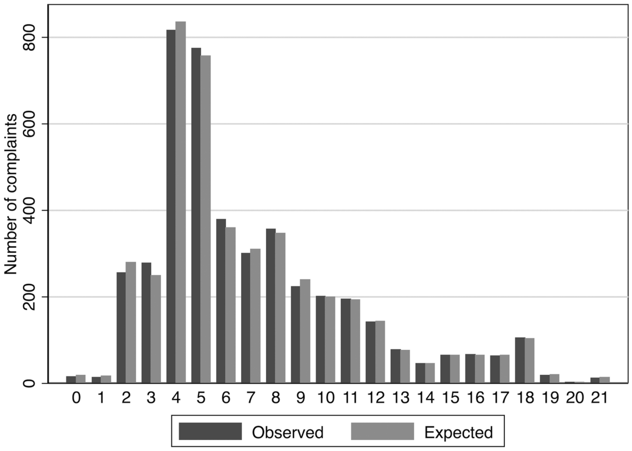


Figure 1 Calibration curves for 22-point PRONE (Predicted Risk Of New Event) score.

letter, a reasonably low threshold for intervention is appropriate. This will ensure that most doctors at risk of a subsequent complaint will receive the intervention (high sensitivity), although many letter recipients will not actually go on to incur further complaints (low specificity). A cut point of 3 or more on the PRONE score, which has sensitivity of 94% and a specificity of 19%, may be a reasonable threshold for this type of intervention.

A much sterner intervention, such as referral to a medical board for assessment and possible further action, would impose a significant burden on affected doctors, and would also generate substantial costs for regulators. For such interventions, there should be high confidence that the practitioner will in fact incur additional complaints in the near term (high specificity), even though setting tolerances in that way will mean that a non-trivial number of doctors who will incur additional complaints will fall below the threshold (low sensitivity). A PRONE score of 12 or more has specificity of 99% and sensitivity of 13%.

Requiring doctors at relatively high risk of additional complaints to undertake a relevant continuing medical education course is an example of intervention that is moderately intrusive. The PRONE score threshold triggering such an intervention should seek to strike a balance between sensitivity and specificity. A cut point of 5 or more, for example, would achieve sensitivity of 69% and specificity of 58%.

Table 3 Frequency and risk of complaint within 2 years, by Predicted Risk Of New Event (PRONE) score groups

PRONE Groups	Number of doctors in group	Total complaints in group	Risk of subsequent complaint (95% CI)
0–2	2000	285	14.2 (12.7 to 15.9)
3–5	7144	1868	26.1 (25.1 to 27.2)
6–8	2474	1034	41.8 (39.8 to 43.8)
9–11	1057	620	58.7 (55.6 to 61.6)
12–14	353	263	74.5 (69.6 to 79.0)
15–17	221	194	87.8 (82.7 to 91.8)
18–21	149	137	91.9 (86.4 to 95.8)

DISCUSSION

The PRONE score exhibited strong predictive properties and has considerable potential as a tool for determining the likelihood that doctors named in complaints will reappear as the subjects of additional complaints in the near future. The precise contours of the algorithm we report map directly to complaints lodged with health service commissions in Australia, not other settings. However, the evidence presented of the feasibility and potential value of this approach has wider implications. It is ripe for testing and possible replication in other places and agencies, including liability insurers and medical boards.

Previous attempts[4–8 10 24 25] to predict medicolegal risk have substantial limitations and, to the best of our knowledge, none, with the exception of the Patient at Risk Score (PARS) score developed by Hickson *et al*,[8] has attempted to convert findings into an user-friendly scoring system to guide interventions. An impressive feature of the PARS system is that it links scores to a suite of doctor-focused interventions aimed at preventing recurrence. However, the PARS predictive algorithm is based on patient complaints accumulated in a handful of hospitals, not a population-wide sample of complaints. Three of the four variables used in PARS are similar to PRONE—namely, doctor's sex, specialty group (surgeons vs non-surgeons) and complaint count. PARS also has a variable for clinical activity, which PRONE does not have. Beyond that, it is very difficult to compare PRONE with PARS because details of the structure and performance of the identification algorithm used in PARS have not been published.

This study extends a recent analysis[13] of patient complaints against Australian doctors. Our earlier analysis showed that the incidence of complaints was extremely skewed, and 'frequent flyers' accounted for a very substantial proportion of all complaints; it also demonstrated the feasibility of identifying those frequent flyers early in the trajectory of their complaints profile. This study takes an important next step: converting such predictive modelling into a simple risk scoring system that is amenable to routine use by regulators.

Our approach differs from previous attempts to predict medicolegal risk in two main respects: inclusion of time-varying predictors (such as a continuously changing measure of prior events) and a focus on practitioners who have already experienced at least one complaint. These distinctive aspects of our analytical design boost our ability to make stronger risk predictions than earlier studies have.

How might the PRONE score be used in practice? We envision two ways it could be incorporated into the complaint-handling process. First, because the score can be recalculated each time a new complaint is lodged, it could be useful for 'red-flagging' cases for a deeper file review—for example, reviewing all previous complaints against a practitioner to ascertain if there are any troubling patterns. Second, the score could be useful for 'tiering' interventions. For example, a low PRONE score may suggest that minimal action is required beyond resolution of the immediate complaint, while a high score may prompt a regulator to consider whether more active intervention is needed to guard against the risk of future harm.

The strengths of the PRONE score are that it is simple and relies on information that most complaint-handling or claim-handling bodies collect routinely. Its predictive properties, based on the risk of a subsequent complaint within each score band, are fairly good. Although the *c*-statistic (0.70) for the multivariate model from which the scoring system is derived indicates only moderate discrimination, our various tests of PRONE score performance were encouraging; in particular, we observed few 'false positives' among doctors who score in the upper reaches of the scale (>15), which is where the scoring system performs best.

However, our study also has several limitations. First, factors others than those we considered predict complaints. For example, patient and doctor characteristics,[26 27] aspects of the doctor-patient relationship,[28 29] and the system in which the doctor works[30 31] are all known to affect patient dissatisfaction and subsequent complaining and claiming behaviour. However, the difficulty regulators face in measuring such factors across an entire case load makes them poor candidates for a risk calculator intended for routine use.

Nonetheless, recognition that unobserved variables may influence a practitioner's complaint risk is crucial at the point of intervention. Environmental or system-related factors—such as solo practice, poor information systems or especially challenging patient populations—are particularly powerful examples of confounding factors. The relationship between individual-related and system-related causes of poor quality care is complex.[32] But the coexistence of observed individual factors and unobserved systemic factors does not negate the value of the PRONE score. Even in situations where a practitioner's outlier status is explained primarily by systemic causes, the score may be an effective way of spotlighting practice environments of concern. Thus, the score is best understood as a method of identifying practitioners whose behaviour and event history warrants special attention and further investigation, rather than a direct determinant of the type of action needed to improve quality.

Third, risk factors acting in concert (ie, interaction effects) may increase or decrease risk of subsequent complaints. Inclusion of interactions may increase the score's predictive power, but at the expense of simplicity, which could act as a barrier to adoption by regulators. In exploratory analyses we found only one significant interaction (between complaint number and specialty). It was a poor candidate for inclusion in the score however, because it would substantially increase the number of parameters without improving discrimination.

Third, we used head counts of practitioners, not more sophisticated measures of doctors' exposure to complaint risk, such as volume of patients treated or procedures conducted. Fourth, we used logistic regression to develop our predictive model. This approach allowed us to directly estimate time parameters (eg, time since last complaint). Since the risk of a new complaint decreases over time, this allowed the resultant PRONE score to increase, decrease or remain unchanged at each new calculation of a doctor's score. Disadvantages of this approach are that it does not fully account for changing baseline risk and the subjectivity associated with choice of time cut points. Survival analysis, which we used in our previous analysis of complaints,[13] deals better with these issues, but has its own disadvantages: it cannot handle time-related predictor variables, risk scores can only increase, and it is a difficult approach to explain to regulators.

Fifth, although our methodology should be generalisable to other medicolegal settings, the extent to which it actually is is unknown. The potential gains from replication in other settings depend on three key factors: (1) the proportion of all events attributable to multievent practitioners; (2) the capacity of available predictors to reliably estimate those events; and (3) the size of the reference population. If the proportion of all events attributable to multievent practitioners is small, or data catchment relates to a relatively small population (eg, a single hospital), this would undermine the usefulness and feasibility of risk prediction. Health regulators and liability insurers interested in developing an approach like the PRONE score should be attentive to these factors.

Finally, even if the PRONE score were to be adopted and deployed effectively, it is insufficient, on its own, to improve the quality and safety of care. It is merely the 'front end' of a quality improvement strategy. The scoring system must connect to interventions that work. Exactly what those interventions are, and how they interlock with the scoring system, lie beyond the scope of this study, although table 4 exemplified the types of approaches that regulators are likely to find attractive.

A risk calculator, like the PRONE score, could be deployed retrospectively or prospectively. As part of a general case load review, such an algorithm could be applied to identify practitioners at highest risk of further events and in need of prompt intervention. Another approach would be to incorporate the tool

Table 4 Precision and rationale of thresholds on the Predicted Risk Of New Event (PRONE) score for three indicative interventions

Intervention	Tolerance for missing practitioners who will recur in short-term	Tolerance for netting practitioners who will not recur in short-term	Rationale	Implications for threshold on PRONE score	Nominal thresholds on PRONE score	Sensitivity/specificity (%)
Advise doctor of future complaint risk	Low	High	Prefer lower false negative rate (quality-of-care considerations); false positives not problematic (cheap to implement, minimally confronting)	High sensitivity desirable, specificity level subordinate	≥3	94/19
Exhortation or compulsion to undertake CME	Moderate	Moderate	Prefer lower false negative rate (quality-of- care) and lower false positive rate (absorbs resources, may waste practitioners' time)	Balance between sensitivity and specificity	≥5	69/58
Refer to regulator for further action	High	Low	Must minimise false positive rate (natural justice); false negative rate undesirable (quality-of-care) but trumped	High specificity essential, sensitivity level subordinate	≥12	13/99

CME, continuing medical education.

into day-to-day handling of complaints or claims, giving regulators an ability to observe ascending levels of risk and tailor responses accordingly. The potential for prospective use is particularly novel and exciting because it holds the promise of ushering medicolegal agencies into the prevention business. However, such uses would inevitably raise ethical and legal challenges. The best way to deflect those challenges may be to ensure that, in any attempt to make levels of predicted risk trigger points for intervention, the intrusiveness of the intervention is well matched to the confidence of the prediction.

Twitter Follow Marie Bismark at @mbismark

Acknowledgements The authors thank the health services commissioners and their staff in the Australian Capital Territory, New South Wales, the Northern Territory, Queensland, South Australia, Tasmania, Victoria and Western Australia; their support, assistance and advice made this study possible.

Contributors MJS, MB, and DMS devised the study. MJS and MB collected the data. MJS undertook the analysis and wrote the first draft. MB and DMS contributed to the interpretation of the results and manuscript revision.

Funding The study was funded by an ARC Laureate Fellowship (FL110100102 to DMS) from the Australian Research Council. The research was conducted independently from the funder.

Competing interests None.

Ethics approval Human Research Ethics Committee, University of Melbourne.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The study data are owned by state health complaint commissions. Permission was granted to use it for purposes of the study, but not for any other use.

REFERENCES

1 Studdert DM, Brennan TA. No-fault compensation for medical injuries: the prospect for error prevention. *JAMA* 2001;286:217–23.

2 Venezian EC, Nye BF, Hofflander AE. The distribution of claims for professional malpractice: some statistical and public policy aspects. *J Risk Insur* 1989;56:686–701.

3 Rolph JE, Kravitz RL, McGuigan K. Malpractice claims data as a quality improvement tool. II. Is targeting effective? *JAMA* 1991;266:2093–7.

4 Cooil B. Using medical malpractice data to predict the frequency of claims: a study of Poisson process models with random effects. *J Am Stat Assoc* 1991;86:285–95.

5 Bovbjerg RR, Petronis KR. The relationship between physicians' malpractice claims history and later claims. *JAMA* 1994;272:1421–6.

6 Gibbons R, Hedeker D, Charles S. A random-effects probit model for predicting medicl malpractice claims. *J Am Stat Assoc* 1994;89:760–7.

7 Weycker DA, Jensen GA. Medical malpractice among physicians: Who will be sued and who will pay? *Health Care Manag Sci* 2000;3:269–77.

8 Hickson GB, Federspiel CF, Pichert JW, *et al*. Patient complaints and malpractice risk. *JAMA* 2002;287:2951–7.

9 Khaliq AA, Dimassi H, Huang C-Y, *et al*. Disciplinary action against physicians: Who is likely to get disciplined? *Am J Med* 2005;118:773–7.

10 Rolph JE, Adams JL, McGuigan KA. Identifying malpractice-prone physicians. *J Emp Leg Stud* 2007;4:125–53.

11 Hickson GB, Federspiel CF, Blackford J, *et al*. Patient complaints and malpractice risk in a regional healthcare center. *South Med J* 2007;100:791–6.

12 Tamblyn R, Abrahamowicz M, Dauphinee D, *et al*. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA* 2007;298:993–1001.

13 Bismark MM, Spittal MJ, Gurrin LC, *et al*. Identification of doctors at risk of recurrent complaints: a national study of healthcare complaints in Australia. *BMJ Qual Saf* 2013;22:532–40.

14 Bismark MM, Spittal MJ, Gogos AJ, *et al*. Remedies sought and obtained in healthcare complaints. *BMJ Qual Saf* 2011;20:806–10.

15 Fallberg L, Mackenney S. Patient Ombudsmen in seven European countries: an effective way to implement patients' rights? *Eur J Health Law* 2003;10:343–57.

16 Paterson R. The patients' complaints system in New Zealand. *Health Aff (Millwood)* 2002;21:70–9.

17 Australian Bureau of Statistics. *Australian Social Trends, April 2013*. Australian Bureau of Statistics, 2014.

18 Australian Health Practitioner Regulation Agency—Registers of Practitioners. AHPRA. http://www.ahpra.gov.au/Registration/Registers-of-Practitioners.aspx (accessed Feb 2012).

19 Australian Bureau of Statistics. *Australian Standard Geographical Classification (ASGC)*. Canberra: Australian Bureau of Statistics, 2005. http://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/DAF9F28078CD196CCA257089008041E4/$File/12160_jul%202005.pdf

20 Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

21 StataCorp. *Stata: Release 13.1*. College Station, TX: StataCorp LP, 2013.

22 Wilson PWF. Challenges to improve coronary heart disease risk assessment. *JAMA* 2009;302:2369–70.

23 Hosmer DW, Lemeshow S. *Applied logistic regression (Wiley Series in probability and statistics)*. 2nd edn. Wiley-Interscience Publication, 2000.

24 Nye BF, Hofflander AE. Experience rating in medical professional liability insurance. *J Risk Insur* 1988;55:150–7.

25 Morrison J, Wickersham P. Physicians disciplined by a state medical board. *JAMA* 1998;279:1889–93.

26 Elkin K, Spittal MJ, Studdert DM. Risks of complaints and adverse disciplinary findings against international medical graduates in Victoria and Western Australia. *Med J Aust* 2012;197:448–52.

27 Bismark MM, Spittal MJ, Studdert DM. Prevalence and characteristics of complaint-prone doctors in private practice in Victoria. *Med J Aust* 2011;195:25–8.

28 Beckman HB, Markakis KM, Suchman AL, *et al*. The doctor-patient relationship and malpractice. Lessons from plaintiff depositions. *Arch Intern Med* 1994;154:1365–70.

29 Chen J, Fang H, Rizzo JA. Physician-patient Language Concordance and Malpractice Concerns. *Med Care* 2011;49:1040–4.

30 Kessler DP. Evaluating the medical malpractice system and options for reform. *J Econ Perspect* 2011;25:93–110.

31 Kilgore ML, Morrisey MA, Nelson LJ. Tort law and medical malpractice insurance Premiums. *Inquiry* 2006;43:255–70.

32 Mello MM, Studdert DM. Deconstructing negligence: the role of individual and system factors in causing medical injuries. *Georgetown Law J* 2008;96:599–623.

Supplementary Appendix

Cross-validation method

A common method of cross-validation is to split the study data into a training and validation samples. In this approach, a statistical model is developed on the training sample and its performance is then assessed against the validation sample. One limitation of this approach is that only a subset of the data is used for model building, and when sample size is limited (as is often the case) this can lead to models that are more unstable than they would have been had the whole sample been used to generate them.

Various alternative methods have been proposed to address this problem, including *K*-fold cross-validation, leave-one-out cross-validation, and bootstrapping methods that adjust for optimism.[1,2] We chose the latter approach because it is relatively simple to implement using standard statistical software and because it has been in other studies that have developed similar kinds of risk scores (see, for example, Cook at al [3]).

Cross-validating a model using bootstrapping methods that adjust for optimism involves three main steps. First, a statistical model is developed using the entire dataset (typically using logistic or linear regression analyses). A measure of fit is estimated for this model (e.g. R-squared or *c*-statistic), which may be called "the overall fit statistic". Second, samples are drawn with replacement from the entire dataset ("bootstrap samples") and the model developed in the previous step is re-estimated for each of these samples, with fit statistics calculated for each one ("bootstrap fit statistics"). Importantly, the coefficient values estimated from each bootstrap sample are also applied to the entire dataset and overall fit statistic is re-calculated ("population fit statistics"). The average of the difference between the population fit statistics and the bootstrap fit statistics represents the "optimism" that biases the overall fit statistic. Therefore, the final step is to subtract the value of the optimism statistic from the value of the overall fit statistic. This value is reported as the measure of model fit, "adjusted for optimism".

Applying this methodology to our analysis, we considered three main multivariable logistic regression models as candidates for the predictive model on which to base our risk score. What follows is a specific description of the cross-validation of Model 1 (see Table 2 in the main paper); the same approach was used for the other two models.

The overall fit statistic for Model 1 was a *c*-statistic of 0.6934. After drawing 200 bootstrap samples, the average bootstrap fit statistic was a *c*-statistic of 0.6943. When the coefficients from each of the bootstrap models were applied to the entire dataset, the average population fit statistic was a *c*-statistic of 0.6935. The difference between the average population fit statistic and the average bootstrap fit statistic is 0.008 (0.6935 - 0.6943), which represents the measure of optimism in the overall *c*-statistic. Thus, after adjustment for optimism, the *c*-statistic for Model 1 is 0.6925 (0.6934 – 0.0008).

Using clinical care complaints only

In the paper we report the results of multivariate logistic regression using all complaints. Here we report the results using just those complaints that related to clinical care. The three models correspond to the three models described in the method and results section.

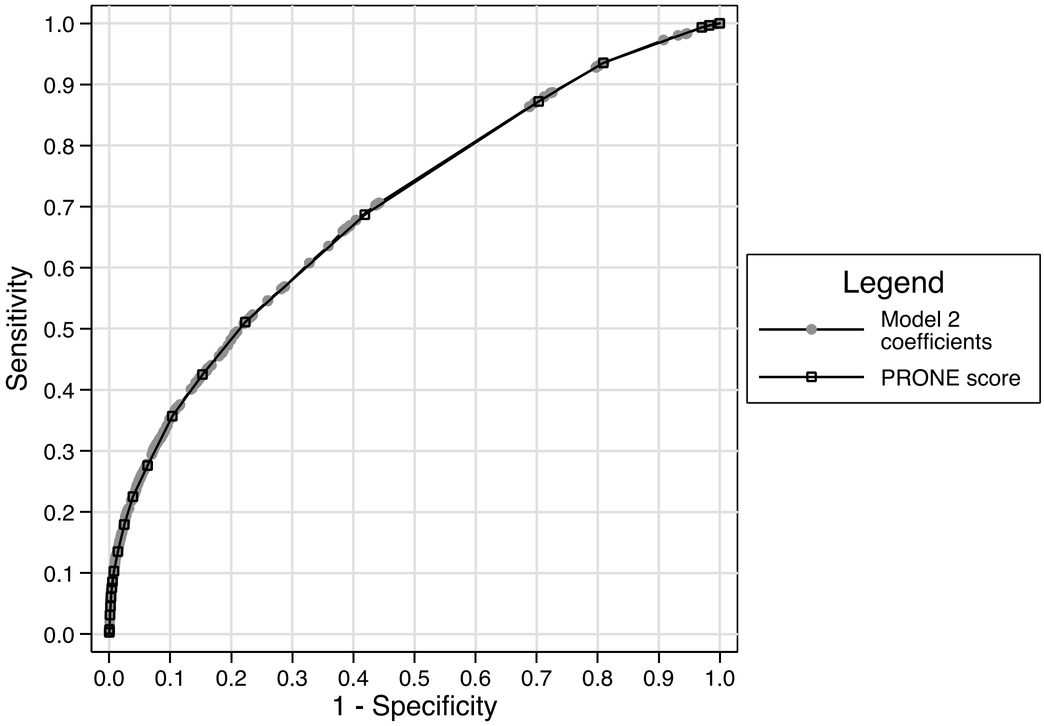
Table S1: Logistic regression models for risk of complaints within 2 years

	Model 4 **OR (95% CI)**	**Model 5** **OR (95% CI)**	**Model 6** **OR (95% CI)**
Complaint number			
1 (ref)	1.00	1.00	1.00
2	1.30 (1.09 to 1.55)	1.29 (1.09 to 1.53)	1.97 (1.74 to 2.23)
3	2.08 (1.64 to 2.63)	2.03 (1.63 t0 2.55)	3.43 (2.84 to 4.14)
4	2.73 (2.00 to 3.71)	2.83 (2.10 to 3.80)	4.95 (3.77 to 6.50)
5	5.04 (3.14 to 8.08)	4.74 (3.05 to 7.36)	8.87 (5.89 to 13.4)
6	6.46 (3.30 to 12.64)	7.01 (3.62 to 13.58)	15.68 (8.40 to 29.3)
7	5.12 (2.46 to 10.66)	4.97 (2.49 to 9.91)	10.37 (5.41 to 19.9)
8	5.34 (2.18 to 13.09)	5.87 (2.45 to 14.08)	13.22 (5.78 (30.2)
9	5.70 (1.78 to 18.26)	5.29 (1.83 to 15.31)	12.12 (4.52 to 32.5)
10+	33.82 (11.24 to 101)	28.90 (11.49 to 72)	60.6 (26.1 to 140)
Doctor's specialty			
Anaesthesia (ref)	1.00	1.00	
Radiology	0.89 (0.37 to 2.13)	0.99 (0.43 to 2.29)	
Other specialties	1.25 (0.73 to 2.15)	1.28 (0.78 to 2.11)	
Internal medicine	1.44 (0.96 to 2.15)	1.52 (1.02 to 2.26)	
Ophthalmology	1.89 (1.17 to 3.05)	2.10 (1.32 to 3.35)	
General practice	1.84 (1.26 to 2.68)	1.98 (1.37 to 2.85)	
Psychiatry	2.28 (1.48 to 3.52)	2.30 (1.51 to 3.51)	
Orthopaedic surgery	2.48 (1.63 to 3.77)	2.71 (1.81 to 4.08)	
Other surgery	2.42 (1.60 to 3.67)	2.62 (1.75 to 3.92)	
General surgery	2.32 (1.50 to 3.59)	2.70 (1.78 to 4.09)	
Obstetrics and gynaecology	2.76 (1.81 to 4.21)	2.93 (1.95 to 4.41)	
Dermatology	3.28 (1.99 to 5.43)	3.71 (2.24 to 6.15)	
Plastic surgery	4.59 (2.93 to 7.17)	5.06 (3.30 to 7.77)	
Time since previous complaint			
1 year or more (ref)	1.00	1.00	
6 months to 1 year	1.32 (1.03 to 1.70)	1.41 (1.11 to 1.80)	
Less than 6 months	1.82 (1.46 to 2.28)	1.94 (1.57 to 2.40)	
Doctor's sex			

Female (ref)	1.00	1.00	
Male	1.61 (1.36 to 1.90)	1.70 (1.45 to 2.00)	
Doctor's age			
22-34 years (ref)	1.00		
35-65 years	1.66 (1.21 to 2.28)		
Location of practice			
Rural (ref)	1.00		
Urban	1.13 (0.99 to 1.29)		
C-statistic (adjusted for optimism)	0.69	0.70	0.65

Performance of PRONE score

Figure S1: Receiver-operating characteristic curves showing the performance of the logistic regression model and the 22-point Complaint PRONE score in predicting risk of complaint within 2 years



References

1. Efron B, Tibshirani R (1993) An Introduction to the Bootstrap. Chapman & Hall/CRC. 1 pp.
2. Trevor H, Robert T, Jerome F (2001) The elements of statistical learning: data mining, inference and prediction. New York: Springer-Verlag.
3. Cook NR, Buring JE, Ridker PM (2006) The effect of including C-reactive protein in cardiovascular risk prediction models for women. Ann Intern Med 145: 21–29.