Two Critical Assessments (Predicting VTE in Acute Leukemia and the Ottawa Knee Rule), Verification Bias, and Tarnished Gold Standards.

Friday February 7, 2020

What Should We Do with This Manuscript?

- Accept the manuscript
- Reject the manuscript
- Ask the authors to revise the manuscript and submit the revised version for reconsideration
  - What revisions should we ask the authors to make?

Development of a Clinical Prediction Rule for Venous Thromboembolism in Patients with Acute Leukemia

Fatimah Al-Ani[1] Yimin Pearl Wang[1] Alejandro Lazo-Langner[1,2]

[1] Division of Hematology, Department of Medicine, Western University, London, Ontario, Canada
[2] Department of Epidemiology and Biostatistics, Western University, London, Ontario, Canada

Address for correspondence Alejandro Lazo-Langner, MD, MSc, FRCPC, Hematology Division, London Health Sciences Centre, Victoria Hospital, 800 Commissioners Road, E Room E6-216, London, Ontario N6A 5W9, Canada (e-mail: alejandro.lazolangner@lhsc.on.ca).

Thrombosis and Haemostasis. Published online: 01.01.2020
DOI https://doi.org/ 10.1055/s-0039-3400303.

Conflict of Interest
None declared.

Funding
This study was funded by the Canadian Venous Thromboembolism Clinical Trials and Outcomes Research (CanVECTOR) Network which receives grant funding from the Canadian Institutes for Health Research (Grant/Award number "CDT 142654"). FA-A was supported by a fellowship from the CanVECTOR Network. AL-L is an investigator of the CanVECTOR Network.

From Another Article

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic			Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	3
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	3
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	4
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	4
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	4, 8
	5b	D;V	Describe eligibility criteria for participants.	4, suppl.
	5c	D;V	Give details of treatments received, if relevant.	6, P

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement

Gary S. Collins, PhD; Johannes B. Reitsma, MD, PhD; Douglas G. Altman, DSc; Karel G.M. Moons, PhD

Ann Intern Med. 2015;162(1):55-63. DOI: 10.7326/M14-0697

The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Initiative developed a set of recommendations for the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. This article describes how the TRIPOD Statement was developed. . . . The resulting TRIPOD Statement is a checklist of 22 items, deemed essential for transparent reporting of a prediction model study.

Steps In Developing Prediction Rules

I. Hypothesis generation
II. Choice of gold standard
III. Choice of predictor variables
IV. Study Sample / Sample size
V. Data collection
VI. Construction of the rule
VII. Test characteristics / Incremental information and cost in different specifications of a rule
VIII. Assessment of the validity of the rule
IX. Provision of information that helps clinicians identify a course of action
X. Assessment of whether the rule affects practice

I. Hypothesis

Risk factors for venous thromboembolism in patients with solid tumors are well studied; however, studies in patients with acute leukemia [AL] are lacking. . . . [We think that] identifying risk factors for the development of VTE among patients with AL will enable clinicians to stratify their patients according to their VTE risk . . . for tailoring surveillance or prophylaxis strategies. . . .

II. Choice of Gold Standard

The main outcome of interest was the occurrence of an objectively documented VTE event including upper and lower deep venous thrombosis (DVT), pulmonary embolism (PE), or thrombosis of unusual sites such as cerebral or portal vein thrombosis.

III. Choice of Candidate Variables 1

Baseline characteristics of participants [Table 1]. . . .

Development of the Prediction Score
Of a total of 20 potential predictors. . . .

Table 1 Baseline population characteristics

Variable	Patients with thrombosis (n¼ 77)	Patients without thrombosis (n ¼ 424)	pValue
Female gender, n (%)	39 (50.6)	202 (47.6)	0.637
Age (y), mean (SD)	49 (18)	60 (17)	<0.001
Body mass index (kg/m²), mean (SD)	28.66 (6.69)	27.59 (6.64)	0.237
Laboratory parameters at presentation, mean (SD)			
Leukocyte count (x 10⁹/L)	44.7 (102.3)	40.4 (79.3)	0.652
Hemoglobin (g/L)	93.4 (24.3)	90.1 (22.3)	0.267
Platelet count (x 10⁹/L)	101.7 (87)	72.8 (80.8)	0.005
Lactate dehydrogenase (U/L)	614.7 (865.5)	639.6 (1030.5)	0.983
International normalized ratio	1.3 (0.5)	1.3 (0.6)	0.758
Partial thromboplastin time (s)	28. 5 (9.9)	28.4 (7.4)	0.891
Creatinine (g/dL) at presentation	94 (94.5)	95.1 (55.4)	0.734
Creatinine clearance (mL/min)	109.14 (39.92)	91.78 (53.27)	0.018
Leukemia lineage, n (%)			<0.001a
Acute myeloid leukemia	36 (46.8)	350 (82.5)	
Acute lymphoblastic leukemia	28 (36.4)	46 (10.8)	
Acute promyelocytic leukemia	13 (16.9)	28 (6.6)	
Comorbidities, n (%)			
Diabetes	11 (14.3)	64 (15.4)	0.805
Hypertension	18 (23.4)	140 (33.7)	0.076
Coronary artery disease	3 (3.9)	34 (8.2)	0.191
Previous stroke	1 (1.3)	13 (3.1)	0.375
Cancer	4(5.2)	97(23.3)	<0.001
History of venous thromboembolism	9 (11.7)	12 (2.8)	<0.001
Medications, n (%)			
Aspirin	9(12.0)	42(10.3)	0.659
Statins	15(19.7)	51(12.5)	0.091
Indwelling lines, n (%)			
Peripherally inserted central catheter	69 (90.8)	294 (69.8)	<0.001
Hickman's catheter	19 (25.0)	53 (12.6)	0.005
Subcutaneous port	5 (6.6)	17 (4.0)	0.322
Internal jugular catheter	5 (6.6)	23 (5.5)	0.698
Presence of any line	71 (92. 2)	300 (70.8)	<0.001
Venous thromboembolism, n (%)			
Any thrombosis	77 (15.3)	—	NA
Catheter-related thrombosis	42 (54.5)	—	NA
Lower extremity deep vein thrombosis and/or pulmonary embolism	28 (36.3)	—	NA
Cerebral thrombosis	5 (6.5)	—	NA
Other sites	2 (2.6)	—	NA

Abbreviations: NA, not applicable; SD, standard deviation.
a p-Value < 0.001 for acute myeloid leukemia versus acute lymphoblastic leukemia.

IV. Study Sample

We conducted a retrospective cohort study of adult patients (≥ 18 years of age) with a new diagnosis of AML or ALL between June 2006 and June 2017 at the London Health Sciences Centre, a tertiary care center in London, Ontario, Canada. The diagnosis of leukemia required confirmation by pathology and multi-parametric flow cytometry. . . . Patients were followed from diagnosis until either (1) the occurrence of VTE; (2) last follow-up; or (3) death. . . . we included all consecutive patients diagnosed at our center, thus reducing the risk of selection bias.

IV. Sample Size

Statistical Analysis

As per standard methodological criteria, a minimum of 5 to 10 events per predictor studied are required for the development of a clinical prediction model. . . . Therefore, based on our center's population we estimated that a sample size of 500 patients would be adequate to explore a predictive model including up to 5 variables.

V. Data Collection

Data collection included demographic data, leukemia lineage, comorbidities, initial laboratory at presentation, chemo-therapy used, the number of admissions to hospital, the presence, number, and type of a central catheter, and any VTE outcome during the follow-up period.

[How were data collected?]

VI. Construction of the Rule 1

Single variable logistic regression analyses were conducted to determine the strength of association between each potential predictor and the occurrence of VTE. . . . Potentially significant predictors ($p < 0.25$) [Table 1] were evaluated using multiple variable stepwise logistic regression analysis with VTE as the dependent variable. Statistically significant variables ($p < 0.05$) were included in the final model.

Table 1 Baseline population characteristics

Variable	Patients with thrombosis (n¼ 27)	Patients without thrombosis (n ¼ 424)	*p* Value
Female gender, n (%)	39 (50.6)	202 (47.6)	0.627
Age (y), mean (SD)	49 (18)	60 (17)	< 0.001
Body mass index (kg/m²), mean (SD)	28.66 (6.69)	27.59 (6.64)	0.237
Laboratory parameters at presentation, mean (SD)			
Leukocyte count (x 10^9/L)	44.7 (102.3)	40.4 (79.3)	0.652
Hemoglobin (g/L)	93 .4 (24.3)	90.1 (22.3)	0.267
Platelet count (x 10^9/L)	101.7 (87)	72.8 (85.8)	0.005
Lactate dehydrogenase (U/L)	634.7 (860.5)	639.6 (1030.5)	0.983
International normalized ratio	1.3 (0.5)	1.3 (0.6)	0.758
Partial thromboplastin time (s)	28. 5 (9.9)	28.4 (7.4)	0.891
Creatinine (g/dL) at presentation	94 (94.5)	95.1 (55.4)	0.734
Creatinine clearance (mL/min)	109.14 (39.92)	91 .78 (53.27)	0.018
Leukemia lineage, n (%)			< 0.001*
Acute myeloid leukemia	36 (46.8)	350 (82.5)	
Acute lymphoblastic leukemia	28 (36.4)	46 (10.8)	
Acute promyelocytic leukemia	13 (16.9)	28 (6.6)	
Comorbidities, n (%)			
Diabetes	11 (14.3)	64 (15.4)	0.805
Hypertension	18 (23.4)	140 (33.7)	0.076
Coronary artery disease	3 (3.9)	34 (8.2)	0.191
Previous stroke	1 (1.3)	13 (3.1)	0.375
Cancer	4(5.2)	97(23.3)	< 0.001
History of venous thromboembolism	9 (11.7)	12 (2.8)	< 0.001
Medications, n (%)			
Aspirin	9(12.0)	42(10.3)	0.659
Statins	15(19.7)	51(12.1)	0.091
Indwelling lines, n (%)			
Peripherally inserted central catheter	69 (90.8)	294 (69.8)	< 0.001
Hickman's catheter	19 (25.0)	53 (12.6)	0.005
Subcutaneous port	5 (6.6)	17 (4.0)	0.322
Internal jugular catheter	5 (6.6)	23 (5.5)	0.698
Presence of any line	71 (92.2)	300 (70.8)	< 0.001
Venous thromboembolism, n (%)			
Any thrombosis	77 (15.3)	—	NA
Catheter-related thrombosis	42 (54.5)	—	NA
Lower extremity deep vein thrombosis and/or pulmonary embolism	28 (36.3)	—	NA
Cerebral thrombosis	5 (6.5)	—	NA
Other sites	2 (2.6)	—	NA

Abbreviations: NA, not applicable; SD, standard deviation.
*p Value < 0.001 for acute myeloid leukemia versus acute lymphoblastic leukemia

VI. Construction of the Rule 2

Variables in the final score were chosen based on clinical applicability and reproducibility. The final prediction score was derived based on weighed variables in the final model. . . .

Of a total of 20 potential predictors, 3 were included in the final model. These included previous VTE, ALL lineage [Acute Lymphoblastic Leukemia], and platelet count > 50 x 10^9/L at the time of presentation. . . . A final prediction score was derived based on weighed variables in the final model as follows: previous history of VTE (3 points), ALL (2 points), and platelet count > 50 x 10^9/L (1 point). The score sum ranged between 0 and 6 score points.

Table 2 Final logistic regression model for occurrence of a venous thromboembolic event

Predictor	β-coefficient	95% confidence interval for the β-coefficient	*p*-Value	*p*-Value (bootstrapped)
History of venous thromboembolism	6.92	2.72–17.63	< 0.001	0.001
Acute lymphoblastic leukemia	4.88	2.73–8.75	< 0.001	0.001
Platelet count > 50 × 10/L at baseline	1.89	1.10–3.22	0.020	0.016

VI. Construction of the Rule 3

[Why did the authors dichotomize the prediction score?]

[How did they pick this particular cutoff value?]

VII. Test Characteristics, Incremental Information, and Cost 1

Of the 501 included patients, the overall cumulative incidence of VTE was 43.8% (n = 32) in the high-risk group, and 10.5% (n = 45) in the low-risk group. The 3-, 6-, and 12-month cumulative incidence of VTE according to risk category is shown in ▶Table 3. . . . Patients with a score of 0 to 2 points had lower risk for VTE compared with patients with a score 3 or more.

[Why are there differences between the values in the text and those for the 12-month values in Table 3?]

Table 3 Cumulative incidence of venous thromboembolism according to risk category

Risk category	*N*	Venous thromboembolism, *N* (%)		
		3 months	6 months	12 months
Low (0–2 points)	428	27 (6.3)	34 (7.9)	40 (9.3)
High (≥ 3 points)	73	21 (28.8)	30 (41.1)	31 (42.5)

VII. Test Characteristics, Incremental Information, and Cost 2

Discrimination

The C-statistic for the model was 0.664 (95% CI: 0.590–0.738) suggesting a good predictive accuracy.

[The C-statistic is the area under the ROC curve. Is this a good C-statistic?]

Calibration

The model was evaluated using Hosmer–Lemeshow tests and pseudo-R2 measures.

[What were the results of the Hosmer-Lemeshow test?]

VII. Test Characteristics, Incremental Information, and Cost 3

Kaplan–Meier survival analysis showed good discrimination between the two categories at all time points (▶ Fig. 1; log-rank p < 0.001).

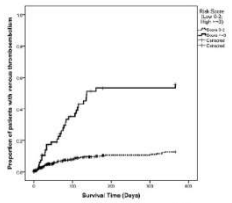


Fig. 1 Kaplan–Meier survival curve at 12 months of follow-up according to risk category.

VII. Test Characteristics, Incremental Information, and Cost 4

- What does the new predictor add to existing methods, for example, clinical intuition
- Cost

VIII. Assessment of Internal Validity

Internal validation was conducted using nonparametric bootstrapping methods. . . .

Table 2 Final logistic regression model for occurrence of a venous thromboembolic event

Predictor	β-coefficient	95% confidence interval for the β-coefficient	*p*-Value	*p*-Value (bootstrapped)
History of venous thromboembolism	6.92	2.72–17.63	< 0.001	0.001
Acute lymphoblastic leukemia	4.88	2.73–8.75	< 0.001	0.001
Platelet count > 50 × 10/L at baseline	1.89	1.10–3.22	0.020	0.016

[Is this the right way to report a bootstrap?]

VIII. Assessment of External Validity

Not done.

IX. Provision of Information That Helps Clinicians Identify a Course of Action

[We think that] identifying risk factors for the development of VTE among patients with AL will enable clinicians to stratify their patients according to their VTE risk . . . for tailoring surveillance or prophylaxis strategies. . . .

Table 3 Cumulative incidence of venous thromboembolism according to risk category

Risk category	*N*	Venous thromboembolism, *N* (%)		
		3 months	6 months	12 months
Low (0–2 points)	428	27 (6.3)	34 (7.9)	40 (9.3)
High (≥ 3 points)	73	21 (28.8)	30 (41.1)	31 (42.5)

IX. Provision of Information That Helps Clinicians Identify a Course of Action

- Finally, the impact and clinical consequences of identifying AL patients at risk for VTE are yet to be determined. This is in great part due to the fact that **leukemia patients have usually a high bleeding risk** due to thrombocytopenia, chemotherapy use, and other risk factors, **and . . . primary thromboprophylaxis is usually deemed to be contraindicated**.
- Potential strategies to deal with this may include **considering higher platelet transfusion thresholds with use of prophylactic doses of LMWH** in patients at a high risk of thrombosis.
- Another consideration is that the duration of treatment in leukemia patients is usually long, in particular in patients with acute lymphoblastic leukemia. Thus, using thromboprophylaxis for the duration of treatment is impractical. In this regard, our model showed that the **higher risk was in the first 3 to 6 months [after the diagnosis of acute leukemia]**, and thus extended prophylaxis may not necessarily be needed.
- Additionally, **nonpharmacologic prophylaxis could be considered** as well in selected cases.
- Lastly, a potential alternative that could be considered is **implementing more proactive VTE surveillance programs** or policies, which could rapidly identify patients at risk of impending thrombotic complications with the objective of initiating rapid pharmacological interventions on a more selective basis.

X. Assessment of Whether the Rule Affects Practice

Not Done.

Conclusions

Discussion
In the present study, we developed a prediction model for VTE in patients with AL that is easy to use and includes variables that are reproducible and can be used consistently. . . . Our model is novel and addresses the existing gap from previous scoring systems.

Compare these conclusions with the original objective

[We think that] identifying risk factors for the development of VTE among patients with AL will enable clinicians to stratify their patients according to their VTE risk . . . for tailoring surveillance or prophylaxis strategies. . . .

Issues

More important issues

- How the authors selected the cutoff point for the prediction score.
  - Why did the authors dichotomize the prediction score?
  - How did they pick this particular cutoff value?
- *Is the C-statistic good enough for clinical use?
- The results of the Hosmer-Lemeshow test (used to measure calibration) are not reported.
- *External validity is yet to be determined.
- *Are the opportunities for preventing VTE realistic?

Other issues

- We know what the candidate predictors were, but we don't know how the authors collected data about them.
- The overall cumulative incidences of VTE reported in the text do not equal the 12-month VTE rates in Table 3.
- Is this the right way to report a bootstrap?

What Should We Do with This Manuscript?

- Accept the manuscript
- Reject the manuscript
- Ask the authors to revise the manuscript and submit the revised version for reconsideration
  - What revisions should we ask the authors to make?

Derivation of a Decision Rule for the Use of Radiography in Acute Knee Injuries

[redacted]
Ian G Stiell, MD, FRCPC*[15]
Gary H Greenberg, MD, FRCPC*
George A Wells, PhD*[15]
R Douglas McKnight, MD, FRCPC*
A Adam Cwinn, MD, FRCPC*
Teresa Cacciotti, RN[5]
Ian McDowell, PhD[5]
Norman A Smith, MD, FRCPC*

Study objective: To derive a highly sensitive decision rule for the selective use of radiography in acute knee injuries.

Design: Prospectively administered survey.

Setting: Emergency departments of two university hospitals.

Participants: Convenience sample of 1,047 adults with acute knee injuries.

OCTOBER 1995 26:4 ANNALS OF EMERGENCY MEDICINE

From the Division of Emergency Medicine, Departments of Medicine; and Epidemiology and Community Medicine ~, and the Clinical Epidemiology Unit", University of Ottawa, Ontario, Canada.*

Supported by grant 06992N from the Emergency Health Services Branch of the Ontario Ministry of Health.

Dr Stiell is a career scientist of the Ontario Ministry of Health, Health Research Personnel Development Program, Toronto.

I. Hypothesis Generation

What is the purpose of this study?
Why does the sensitivity have to be 1.0?

The objective of this study was to derive a decision rule for the use of radiography in acute knee injuries. To be clinically useful for emergency physicians, such a rule should have a sensitivity of 1.0 for identifying fractures, and should be reliable and easy to apply.

We previously have shown that experienced physicians have the ability, using clinical judgment, to determine accurately which knee-injury patients have fractures. However, often they are reluctant to use this skill.(40)

II. Gold Standard

The criterion that the decision rule was designed to identify was any fracture of the knee or patella seen on standard plain knee radiography. We also defined a clinically insignificant fracture as any avulsion fragment that was less than 5 mm in breadth and that was not associated with a complete tendon or ligament disruption.

Those patients who did not have radiography in the ED answered a structured telephone questionnaire to determine the possibility of a missed fracture. Patients were classified as having no fracture if they satisfied all five of the explicit criteria listed in Figure 1. Patients who could not fulfil the criteria were recalled for clinical reassessment and radiography.

Figure 1.

Criteria used in structured 14-day follow-up telephone questionnaire of patients who did not undergo radiography.

Patients who could not fulfill all of the following criteria were recalled for reassessment and radiography:

- — Pain is better.
- — Ability to walk is better.
- — Does not require assistance to walk (crutches/cast/splint).
- — Has returned to usual occupational activities (work, housework, or school).
- — Has no plans to see a physician about knee injury.

How many patients were not reached by telephone?

From Results: During the 14 months of the study (September 1992 through October 1993), 1,054 of 1,212 (87%) eligible knee-injury patients were enrolled in the study. Telephone follow--up was achieved in 340 of 347 (98%) patients who did not have ED radiography. None of these patients proved to have a fracture. The seven patients who could not be reached in follow-up to have their fracture status confirmed were excluded from further analysis.

From Discussion: We are fully confident that our explicit structured telephone follow-up questionnaire most likely would have identified any patients harboring a missed fracture. The eight patients who could not be reached in follow-up were excluded from the analysis. We have used this technique successfully to identify missed fractures in previous studies.18·20

III. Choice of Predictor Variables

Patients were assessed for 23 standardized clinical variables, which had been selected by the investigators on the basis of their clinical experience, data from the literature, and the results of a 2-month pilot study.

Table 2 lists the proportions of patients with and without knee fractures who were positive for the clinical variables, including those created by means of a cutoff point or combination. Most associations were statistically significant; X2 values with 1 df, the basis of the recursive partitioning splits, are given for dichotomous variables. Interobserver agreement, however, exceeded .5 for only 18 of the variables.

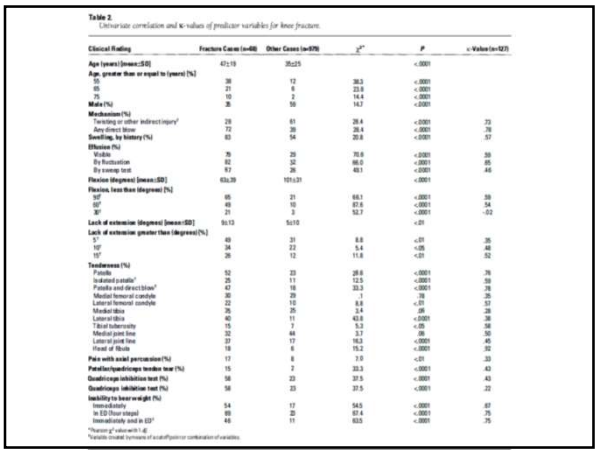


Table 2.
Univariate correlation and κ-values of predictor variables for knee fracture.

Clinical finding	Fracture Cases (n=68)	Other Cases (n=979)	χ^2*	P	κ-Value (n=127)
Age (years) [mean±SD]	47±19	35±15		<.0001	
Age, greater than or equal to (years) [%]					
55	38	12	38.3	<.0001	
65	21	6	23.8	<.0001	
75	10	2	14.4	<.0001	
Male [%]	35	56	14.7	<.0001	
Mechanism [%]					
Twisting or other indirect injury†	28	61	28.4	<.0001	.73
Any direct blow	72	39	28.4	<.0001	.78
Swelling, by history [%]	83	54	20.8	<.0001	.57
Effusion [%]					
Visible	75	29	70.8	<.0001	.59
By fluctuation	82	32	66.0	<.0001	.65
By sweep test	67	26	48.1	<.0001	.46
Flexion (degrees) [mean±SD]	63±39	101±31		<.0001	
Flexion, less than (degrees) [%]					
90°	85	21	68.1	<.0001	.59
60°	49	10	67.6	<.0001	.54
30°	21	3	52.7	<.0001	-.02
Lack of extension (degrees) [mean±SD]	9±13	5±10		<.01	
Lack of extension greater than (degrees) [%]					
5°	49	31	8.8	<.01	.35
10°	34	22	5.4	<.05	.48
15°	26	12	11.8	<.01	.52
Tenderness [%]					
Patella	52	23	38.6	<.0001	.76
Isolated patella†	25	11	12.5	<.0001	.59
Patella and direct blow†	42	16	33.3	<.0001	.76
Medial femoral condyle	20	29	1	.16	.35
Lateral femoral condyle	22	10	8.8	<.01	.57
Medial tibia	25	25	3.4	.06	.28
Lateral tibia	40	11	43.8	<.0001	.38
Tibial tuberosity	15	7	5.3	<.05	.58
Medial joint line	32	44	3.7	.06	.50
Lateral joint line	37	17	16.3	<.0001	.45
Head of fibula	18	6	15.2	<.0001	.92
Pain with axial percussion [%]	17	8	7.0	<.01	.33
Patellar/quadriceps tendon tear [%]	15	7	33.3	<.0001	.43
Quadriceps inhibition test [%]	58	23	37.5	<.0001	.43
Quadriceps inhibition test [%]	58	23	37.5	<.0001	.22
Inability to bear weight [%]					
Immediately	54	17	54.5	<.0001	.67
In ED (hour elapsed	69	20	67.4	<.0001	.75
Immediately and in ED‡	46	11	63.5	<.0001	.75

*Pearson χ^2 value with 1 df.
†Variable created by means of a scatterplot or combination of variables.

IV. Study Sample

RESULTS
During the 14 months of the study (September 1992 through October 1993), 1,054 of 1,212 (87%) eligible knee-injury patients were enrolled in the study. Tele-phone follow--up was achieved in 340 of 347 (98%) patients who did not have ED radiography. None of these patients proved to have a fracture.

V. Data Collection

MATERIALS AND METHODS
This study was conducted in the EDs of two teaching institutions affiliated with the University of Ottawa, the Ottawa Civic and Ottawa General hospitals. We included adult patients who presented with acute blunt injuries of the knee caused by any mechanism of injury The "knee" was considered to include the patella, the head and neck of the fibula, the proximal 8 cm of the tibia, and the distal 8 cm of the femur. We excluded patients who were younger than 18 years, were pregnant, had isolated injuries of the skin without underlying soft-tissue or bone involvement. . . , had an altered level of consciousness, were paraplegic, or had multiple trauma or other fractures.

V. Data Collection 2

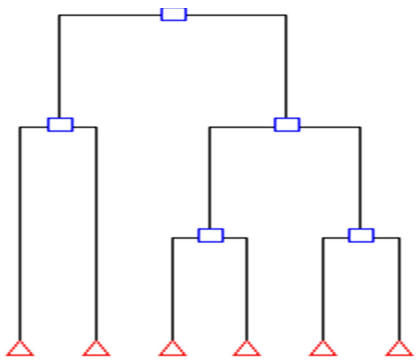
Eligible patients were entered into the study when 1 of 33 designated staff emergency physicians was on duty. . . The physicians were trained by means of a 1-hour lecture and practical demonstration to assess the clinical variables in a standardized fashion. . . . Furthermore, explicit definitions of each variable were . . . on the back of the data collection sheet. Flexion and lack of extension were measured with a goniometer. The findings were recorded . . . before radiography . . . . To determine the interobserver reliability of the physical findings, the patients were examined, where feasible, by a second emergency physician who was blinded to the results of the first assessment.

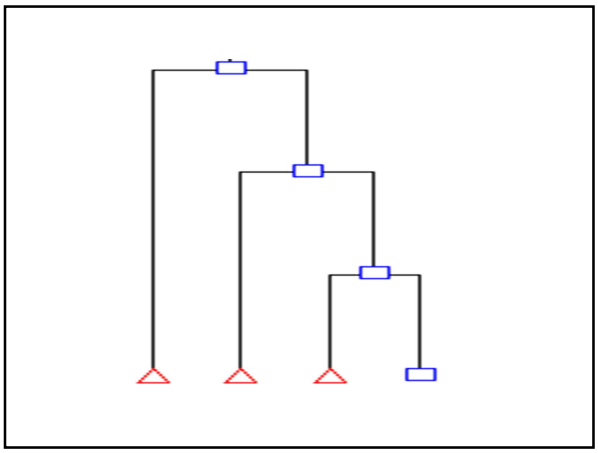
VI. Construction of the Rule

Recursive Partitioning

CART

Classification and Regression Trees





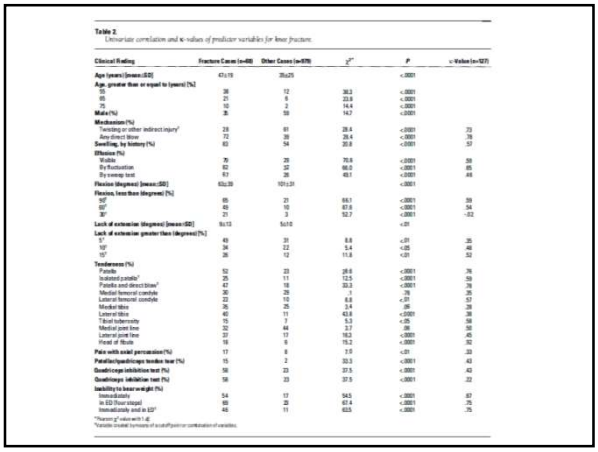


Table 2.
Univariate correlation and κ-values of predictor variables for knee fracture.

Clinical Finding	Fracture Cases (n=69)	Other Cases (n=879)	χ^{2*}	*P*	κ-Value (n=127)
Age (years) [mean±SD]	47±19	35±25		<.0001	
Age, greater than or equal to (years) [%]					
55	38	12	38.3	<.0001	
65	21	6	23.8	<.0001	
75	10	2	14.4	<.0001	
Male (%)	35	59	14.7	<.0001	
Mechanism (%)					
Twisting or other indirect injury†	28	61	28.4	<.0001	.73
Any direct blow	72	39	28.4	<.0001	.78
Swelling, by history (%)	83	54	20.8	<.0001	.57
Effusion (%)					
Visible	70	29	70.6	<.0001	.59
By fluctuation	82	32	66.0	<.0001	.65
By sweep test	67	26	48.1	<.0001	.46
Flexion (degrees) [mean±SD]	63±39	101±31		<.0001	
Flexion, less than (degrees) [%]					
90°	85	21	66.1	<.0001	.59
60°	49	10	67.6	<.0001	.54
30°	21	3	52.7	<.0001	−.02
Lack of extension (degrees) [mean±SD]	9±13	5±10		<.01	
Lack of extension greater than (degrees) [%]					
5°	49	31	8.8	<.01	.35
10°	34	22	5.4	<.05	.48
15°	26	12	11.8	<.01	.52
Tenderness (%)					
Patella	52	23	28.6	<.0001	.76
Isolated patella†	25	11	12.5	<.0001	.59
Patella and direct blow†	47	18	33.3	<.0001	.76
Medial femoral condyle	30	29	1	.76	.25
Lateral femoral condyle	22	10	8.8	<.01	.57
Medial tibia	26	25	2.4	.06	.28
Lateral tibia	40	11	43.8	<.0001	.38
Tibial tuberosity	15	7	5.3	<.05	.58
Medial joint line	32	44	3.7	.09	.50
Lateral joint line	37	17	16.3	<.0001	.45
Head of fibula	18	9	15.2	<.0001	.52
Pain with axial percussion (%)	17	8	7.0	<.01	.33
Patellar/quadriceps tendon tear (%)	15	2	33.3	<.0001	.43
Quadriceps inhibition test (%)	58	23	37.5	<.0001	.43
Quadriceps inhibition test (%)	58	23	37.5	<.0001	.22
Inability to bear weight (%)					
Immediately	54	17	54.5	<.0001	.67
In ED (four steps)	68	25	67.4	<.0001	.75
Immediately and in ED†	46	11	63.5	<.0001	.75

*Pearson χ^2 value with 1 df.
†Variable created by means of a cutoff point or combination of variables.

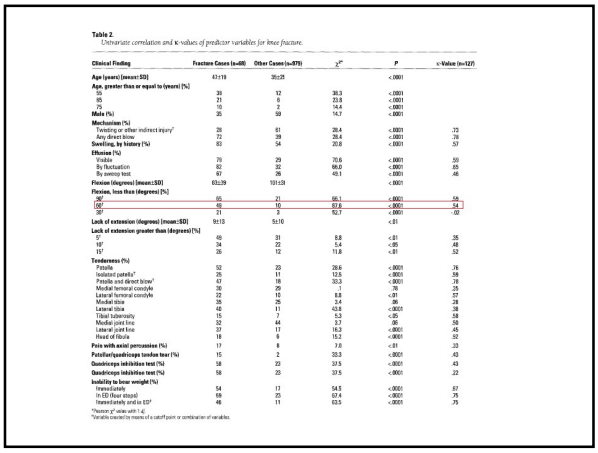
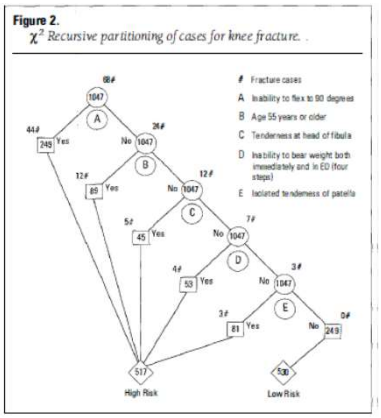


Table 2.
Univariate correlation and κ-values of predictor variables for knee fracture.

Clinical Finding	Fracture Cases (n=69)	Other Cases (n=879)	χ^{2*}	*P*	κ-Value (n=127)
Age (years) [mean±SD]	47±19	35±25		<.0001	
Age, greater than or equal to (years) [%]					
55	38	12	38.3	<.0001	
65	21	6	23.8	<.0001	
75	10	2	14.4	<.0001	
Male (%)	35	59	14.7	<.0001	
Mechanism (%)					
Twisting or other indirect injury†	28	61	28.4	<.0001	.73
Any direct blow	72	39	28.4	<.0001	.78
Swelling, by history (%)	83	54	20.8	<.0001	.57
Effusion (%)					
Visible	70	29	70.6	<.0001	.59
By fluctuation	82	32	66.0	<.0001	.65
By sweep test	67	26	48.1	<.0001	.46
Flexion (degrees) [mean±SD]	63±39	101±31		<.0001	
Flexion, less than (degrees) [%]					
90°	85	21	66.1	<.0001	.59
60°	49	10	67.6	<.0001	.54
30°	21	3	52.7	<.0001	−.02
Lack of extension (degrees) [mean±SD]	9±13	5±10		<.01	
Lack of extension greater than (degrees) [%]					
5°	49	31	8.8	<.01	.35
10°	34	22	5.4	<.05	.48
15°	26	12	11.8	<.01	.52
Tenderness (%)					
Patella	52	23	28.6	<.0001	.76
Isolated patella†	25	11	12.5	<.0001	.59
Patella and direct blow†	47	18	33.3	<.0001	.76
Medial femoral condyle	30	29	1	.76	.25
Lateral femoral condyle	22	10	8.8	<.01	.57
Medial tibia	26	25	2.4	.06	.28
Lateral tibia	40	11	43.8	<.0001	.38
Tibial tuberosity	15	7	5.3	<.05	.58
Medial joint line	32	44	3.7	.09	.50
Lateral joint line	37	17	16.3	<.0001	.45
Head of fibula	18	9	15.2	<.0001	.52
Pain with axial percussion (%)	17	8	7.0	<.01	.33
Patellar/quadriceps tendon tear (%)	15	2	33.3	<.0001	.43
Quadriceps inhibition test (%)	58	23	37.5	<.0001	.43
Quadriceps inhibition test (%)	58	23	37.5	<.0001	.22
Inability to bear weight (%)					
Immediately	54	17	54.5	<.0001	.67
In ED (four steps)	68	25	67.4	<.0001	.75
Immediately and in ED†	46	11	63.5	<.0001	.75

*Pearson χ^2 value with 1 df.
†Variable created by means of a cutoff point or combination of variables.



From the Second Article

A series of 2 recursive partitioning analyses yielded a model that was more specific than the original decision rule. This refined model was identical to the original rule except that "inability to flex to 90°" was replaced by "inability to flex to 60°." Application of this refined model to the current study population would have yielded a sensitivity of 1.0, a specificity of 0.56, and a potential relative reduction in radiography of 36%. Application of the refined model to the 1047 derivation set patients (1992 to 1993), however, revealed that five clinically important fractures would have been missed. The investigators felt that this loss in sensitivity was unacceptable and that the refined model should not be adopted.

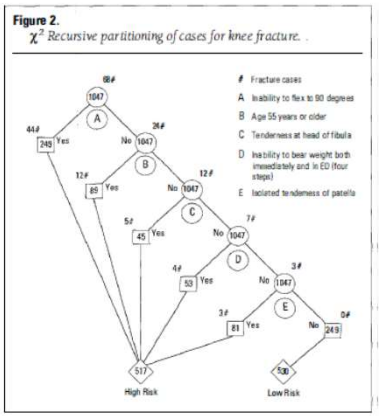


Figure 3.
Decision rule for radiography in acute knee injury.

A knee radiograph is required only for acute knee-injury patients with one or more of these findings related to age, tenderness, or function:
- Age 55 years or older
- Tenderness at head of fibula
- Isolated tenderness of patella
- Inability to flex to 90 degrees
- Inability to bear weight both immediately and in the ED (four steps)

VII. Test Characteristics, Incremental Information, and Cost

If applied to the study population, the decision rule would have had a sensitivity of 1.0 (95% CI, .95 to 1.0) and a specificity of .54 (95% CI, .51 to .57) for identifying fractures of the knee (Table 3). Furthermore, application of the rule would have led to a 28.0% relative reduction in use of radiography from a baseline rate of 68.6% to a potential rate of 49.4%.

VIII. Validation

The major limitations of this study are that the decision rule has not been validated prospectively and has not undergone an implementation trial. No decision rule should be considered for clinical use until it has been validated prospectively.(41) Many guidelines or decision rules do not perform well when tested in a new patient population.(44) We currently are conducting a validation study. . . We then plan to conduct an implementation trial to demonstrate the true effect of the decision rule on clinical practice. Very few decision rules have undergone field trials to test their effectiveness in altering patient care.(41·45)

JAMA
The Journal of the American Medical Association

Volume 275(8) 28 February 1996 pp 611-615

Prospective Validation of a Decision Rule for the Use of Radiography in Acute Knee Injuries
[Original Contributions]

Stiell, Ian G. MD, MSc, FRCPC; Greenberg, Gary H. MD, FRCPC; Wells, George A. PhD; McDowell, Ian PhD; Cwinn, A. Adam MD, FRCPC; Smith, Norman A. MD, FRCPC; Cacciotti, Teresa F. RN; Sivilotti, Marco L. A. MD, MSc, FRCPC

From the Division of Emergency Medicine (Drs Stiell, Greenberg, Cwinn, Smith, and Sivilotti), the Departments of Epidemiology and Community Medicine (Drs Stiell, Wells, and McDowell) and Medicine (Drs Stiell and Wells), and the Clinical Epidemiology Unit (Drs Stiell and Wells and Ms Cacciotti), University of Ottawa (Ontario) Faculty of Medicine. Dr Stiell is an MRC scientist with the Medical Research Council of Canada Research Personnel Program.
Reprint requests to Clinical Epidemiology Unit, Loeb Medical Research Institute, Ottawa Civic Hospital, 1053 Carling Ave, Ottawa, Ontario, Canada K1Y 4E9 (Dr Stiell).

This study was supported by grant 6898SN from the Emergency Health Services Branch of the Ontario Ministry of Health, Toronto.

Table 3.—Classification Performance of the Decision Rule for Identifying Clinically Important Knee Fractures Among the Study Patients*

	Fracture		
	Yes	No	
Decision rule			
Positive	63	522	
Negative	0	511	
Sensitivity (95% CI)			1.0 (0.94-1.0)
Specificity (95% CI)			0.49 (0.46-0.52)
Negative predictive value (95% CI)			1.0 (0.99-1.0)
Positive predictive value (95% CI)			0.11 (0.08-0.13)

*CI indicates confidence interval.

Is this Internal Validation, External Validation, or some combination?

From Study 1

MATERIALS AND METHODS
This study was conducted in the EDs of two teaching institutions affiliated with the University of Ottawa, the Ottawa Civic and Ottawa General hospitals.

From Study 2

METHODS
Study Population
The study was conducted in the emergency departments of two teaching hospitals serving adults affiliated with the University of Ottawa (Ontario) Faculty of Medicine: Ottawa Civic Hospital and Ottawa General Hospital.

Why the change in definition?

Study 1

The criterion that the decision rule was designed to identify was any fracture of the knee or patella seen on standard plain knee radiography. We also defined a clinically insignificant fracture as any avulsion fragment that was less than 5 mm in breadth and that was not associated with a complete tendon or ligament disruption

Study 2

Outcome Measure
The criterion standard that the decision rule was developed to identify was a clinically important fracture of the knee demonstrated on a standard knee radiographic series.

Three of the five clinically unimportant fractures would not have been identified by the rule; none of these cases were treated with a cast.

X. Assessment of Whether the Rule Affects Practice

How many decisions about knee x-rays were influenced by the Ottawa Knee decision rule in the first two studies?

Implementation of the Ottawa Knee Rule for the Use of Radiography in Acute Knee Injuries

Ian G. Stiell, MD, MSc, FRCPC; George A. Wells, PhD; Roger H. Hoag, MD, FRCPC; Marco L. A. Sivilotti, MD, MSc, FRCPC; Teresa F. Cacciotti, RN; P. Richard Verbeek, MD, FRCPC; Keith T. Greenway, MD, FRCPC; Ian McDowell, PhD; A. Adam Cwinn, MD, FRCPC; Gary H. Greenberg, MD, FRCPC; Graham Nichol, MD, FRCPC; John A. Michael, MD, FRCPC

JAMA. 1997;278:2075-2079

From the Division of Emergency Medicine (Drs Stiell, Sivilotti, Cwinn, Greenberg, and Michael), the Department of Medicine (Drs Wells and Nichol), the Department of Epidemiology and Community Medicine (Dr McDowell), and the Clinical Epidemiology Unit (Ms Cacciotti), University of Ottawa, Ottawa, Ontario; the Department of Emergency Medicine (Dr Hoag), Queensway-Carleton Hospital, Nepean, Ontario; the Division of Emergency Medicine (Dr Verbeek), University of Toronto, Toronto, Ontario; the Department of Emergency Medicine (Dr Greenway), Peel Memorial Hospital, Brampton, Ontario.

This study was supported by grant H1056N from the Emergency Health Services Branch of the Ontario Ministry of Health, Toronto. Dr Stiell holds a Scientist Award from the Medical Research Council of Canada.

Table 2.—Referral for Knee Radiography of All Knee Injury Patients Seen at the Intervention and Control Hospitals During the 12-Month Before and After Study Periods*

	Referred for Radiography at Intervention Hospitals, No. (%)			Referred for Radiography at Control Hospitals, No. (%)		
	Before (n=982)	After (n=1063)	RR, % (95% CI)	Before (n=962)	After (n=900)	RR, % (95% CI)*
Overall†	762 (77.6)	607 (57.1‡)	26.4 (21.7-30.8)	740 (76.9)	683 (75.9)	1.3 (–3.8-6.2)
Individual hospitals§						
Ottawa Civic	311 (67.8)	265 (48.7)‖	28.1 (20.0-35.4)	. . .	. . .	. . .
Queensway Carleton	451 (86.2)	342 (65.9)‖	23.6 (18.0-28.8)			
Sunnybrook HSC	. . .	. . .	. . .	286 (76.3)	281 (73.8)	3.3 (–5.0-10.9)
Peel Memorial	. . .	. . .	. . .	454 (77.3)	402 (77.5)	–0.1 (–6.7-6.0)

*RR indicates relative reduction; CI, confidence interval; and ellipses, not applicable.
†P<.001; compares intervention group with control group for change from before period to after period.
‡P<.001; compares before period to after period for intervention group.
§Ottawa Civic and Sunnybrook Health Science Center (HSC) are teaching hospitals. Queensway Carleton and Peel Memorial are community hospitals.
‖P<.001; compares before period to after period at each intervention hospital.

Three items to note:
1. The frequency of referral for x-rays went down in the 2 intervention hospitals (77.6% to 57.1%) and did not change in the control hospitals (76.9% to 75.9%).
2. Ottawa Civic Hospital was in the 2 previous studies. The other 3 hospitals were not.
3. The number of people with knee injuries increased in the 2 intervention hospitals (982 to 1063) but decreased in the 2 control hospitals (962 to 900) (Chi-square 22.958, $p < 0.0001$).

VII. Test Characteristics, Incremental Information, and Cost

- Incremental information
  - Compared to clinical intuition (Not done)

    We previously have shown that experienced physicians have the ability, using clinical judgment, to determine accurately which knee-injury patients have fractures. However, often they are reluctant to use this skill.(40)
  - Compared to other prediction rules (Not done, perhaps because there were no other such rules)
- Cost (Done in subsequent studies)

Topics in Diagnostic Test Accuracy

"The field of test evaluation is plagued with poor design, low sample sizes, poor reporting, and a low volume of research"

Johannes B. Reitsma, et al. *J Clin Epidemiol.* 2009;62:797-806.

Resources for Doing and Reporting Studies of Diagnostic Tests

Studies of Individual Diagnostic Tests

1. Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. *Cochrane Database of Systematic Reviews* 2013, Issue 9. Art. No.: MR000022. DOI: 10.1002/14651858.MR000022.pub3.
2. Sources of bias and variation in studies of diagnostic test accuracy. (*J Clin Epidemiol.* 2013;66:1093-1104)
3. Sources of Variation and Bias in Studies of Diagnostic Accuracy: A Systematic Review. *Ann Intern Med.* 2004;140:189-202.
4. How to assess quality in studies of diagnostic test accuracy (QUADAS-2) (*Ann Intern Med.* 2011;155:529-536)
5. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;6:e012799. doi:10.1136/bmjopen-2016-012799

Systematic Reviews and Meta-Analyses of Diagnostic Tests

1. How to conduct systematic reviews of diagnostic test accuracy (*Ann Intern Med.* 2008;149:889-897) Trevor A. McGrath, Mostafa Alabousi, Becky Skidmore, Daniël A. Korevaar, Patrick M. M. Bossuyt, David Moher, Brett Thombs and Matthew D. F. McInnes. Recommendations for reporting of systematic reviews and meta-analyses of diagnostic test accuracy: a systematic review. *Systematic Reviews*. 2017**6**:194. https://doi.org/10.1186/s13643-017-0590-8
2. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA* 2018;319(4):388–396.

3a. Kim KW, Lee J, Choi SH, Huh J, Park SH. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers--part I. general guidance and tips. *Korean J Radiol.* 2015;16:1175-1187

3b. Systematic Review and Meta-Analysis of Studies Evaluating Diagnostic Test Accuracy: A Practical Review for Clinical Researchers–Part II. Statistical Methods of Meta-Analysis. *Korean J Radiol.* 2015;16(6):1188-1196.

Meta-analysis of Published Studies Evaluating Sensitivity and Specificity When Fecal Occult Blood Testing (FOBT) Is Used to Screen for Colorectal Cancer

	FOBT Sensitivity	FOBT Specificity
Colonoscopy depended on FOBT results (19 studies)	0.70	0.88
Colonoscopy did not depend on FOBT results; everyone had colonoscopy (7 studies) or longitudinal follow up (3 studies)	0.36	0.96

J Gen Intern Med 25(1):1211–21
DOI: 10.1007/s11606-010-1375-0

A meta-analysis of 147 published studies summarized the operating characteristics of the exercise-ECG test for diagnosing coronary artery disease (CAD) as follows:

	CAD	CAD	
Exercise ECG	Present	Absent	
Positive	7,830	2,896	10,726
Negative	3,686	9,662	13,348
	11,516	12,558	24,074

Sensitivity = 7,830 / 11,516 = .68
Specificity = 9,662 / 12,558 = .77

(Journal of the American College of Cardiology. 1997;30:260-311)

One year later, a study in which all patients had both exercise-ECG testing and coronary angiography described the operating characteristics of the exercise-ECG test as follows:

	CAD	CAD	
Exercise ECG	Present	Absent	
Positive	185	60	245
Negative	226	343	569
	411	403	814

Sensitivity = 185 / 411 = .45
Specificity = 343 / 403 = .85

(Ann Intern Med. 1998;128:965-974)

Verification Bias

When the results of a diagnostic test affect whether the gold standard procedure is used to verify the test result, verification bias is introduced. This problem is also called work-up bias.

Verification bias is common because many gold standard procedures, such as biopsy, surgery, and angiography, are invasive, risky, and expensive. Under these conditions, physicians are reluctant to refer patients for the gold standard procedure, and patients are reluctant to undergo the gold standard procedure, unless preliminary diagnostic tests have positive results.

What Verification Bias Does

	CAD+	CAD-
ECG +	↑	↑
ECG -	↓	↓

When verification bias is present, we observe too many TPs and too few TNs relative to FNs and FPs because people with positive ECG results are more likely to have coronary angiography than people with negative ECG results. Therefore, observed sensitivity (TP/(TP+FN)) is too high with too many TPs relative to FNs, and observed specificity (TN/(TN+FP)) is too low with too few TNs relative to FPs.

Original Effort To Adjust for Verification Bias

Biometrics, Vol. 39, No. 1 (Mar., 1983), pp. 207-215

One, you must know the total numbers of people with each type of result for the test in question (in this case, everyone who had an exercise ECG, including those who did not have coronary angiography).

Two, you must assume that the PPV and NPV are the same in the people who had the gold standard procedure and the people who did not have the gold standard procedure.

Hypothetical Example: Assume that in the meta-analysis about exercise-ECG testing 80% of people with a positive result for the exercise-ECG test had coronary angiography and 25% of people with a negative result for the exercise-ECG test results had coronary angiography.

Start with the original table from the meta-analysis study.

	CAD Present	CAD Absent	Total
ECG positive	7,830	2,896	10,726
ECG negative	3,686	9,662	13,348
Total	11,516	12,558	24,074

Inflate row totals to adjust for differential verification. For example, 13,408 = 10,726/0.80 and 53,392 = 24,072/0.20.

	CAD Present	CAD Absent	Total
ECG positive			13,408
ECG negative			53,392
Total			

Use predictive values (PPV=.73, NPV=.72) from the original 2by2 table to calculate the number of true positives (.73 X 13,408 = 9,788) and the number of true negatives (.72 X 53,392 = 38,648) in the new 2by2 table, and use subtraction to get the remaining cell numbers. Calculate sensitivity and specificity.

	CAD Present	CAD Absent	Total
ECG positive	9,788	3,620	13,408
ECG negative	14,744	38,648	53,392
Total	24,532	42,268	

Sensitivity = 9,788 / 24,532 = .40
Specificity = 38,648 /42,268 = .91

Compare Original with Revised Results

Original	CAD Present	CAD Absent	Total
ECG positive	7,830	2,896	10,726
ECG negative	3,686	9,662	13,348
Total	11,516	12,558	24,074

Sensitivity = 7,830 / 11,516 = .68
Specificity = 9,662 / 12,558 = .77

Revised	CAD Present	CAD Absent	Total
ECG positive	9,788	3,620	13,408
ECG negative	14,744	38,648	53,392
Total	24,532	42,268	66,800

Sensitivity = 9,788 / 24,532 = .40
Specificity = 38,648 / 42,268 = .91

Compare All Three

	Sensitivity	Specificity
Original meta-analysis	.68	.77
Optimal study	.45	.85
Meta-analysis adjusted with PPV and NPV	.41	.91

Why are the adjusted values for the meta-analysis different from the values in the optimal study?

Why the Assumption is Incorrect

Consider two people with suspected CAD, both with negative test results. One person is a 50-year-old woman with atypical chest pain and no risk factors for CAD. The other person is a 65-year-old male smoker who has typical angina, diabetes mellitus, and a strong family history of CAD. The woman is less likely to have CAD, and she is less likely to have her negative exercise-ECG result "verified" with angiography. In contrast, the man is more likely to have CAD, and he is more likely to have his negative exercise-ECG result "verified" with angiography. A similar, but perhaps less powerful, effect likely occurs when the test result is positive. Therefore, people with disease are more likely to have their test results verified with the gold standard procedure than people without disease.

One effect of verification bias is that people with positive test results are more likely to be verified, which means the numbers in the upper row of the 2by2 table are increased relative to the cells in the lower row.

	CAD+	CAD-
ECG +	↑	↑
ECG -	↓	↓

Another effect of verification bias is that people who are more likely to have disease are more likely to be verified, which means the numbers in the left-hand column of the 2by2 table are increased relative to the cells in the right-hand column.

	CAD+	CAD-
ECG+	↑	↓
ECG -	↑	↓

	CAD+	CAD-
ECG +	↑	↑
ECG -	↓	↓

	CAD+	CAD-
ECG+	↑	↓
ECG -	↑	↓

The combination of these two separate effects further increases the number of true positives relative to false negatives and further increases the number of true negatives relative to false positives. These additional increases are the reason why the proposed fix overcorrects when it Is applied.

	CAD+	CAD-
ECG +	↑↑	↓↑
ECG -	↓↑	↓↓

The Result of Adjusting for Verification Bias

In most situations, the adjustment for verification bias with this method produces a sensitivity that is lower than the true sensitivity and a specificity that is higher than the true specificity. The adjustment "over corrects" for verification bias.

The Bottom Line

In most situations affected by verification bias, the true values for sensitivity and specificity are between the values reported in the original article and the values that are calculated using this adjustment method. The reported and adjusted values, however, may be useful because they define a range surrounding the true values for sensitivity and specificity.

The Bottom Line 2

- Verification bias is important because it leads to incorrect reports of the operating characteristics for diagnostic tests, usually with falsely elevated sensitivity and falsely lowered specificity, and the differences can be substantial.
- Verification bias is even more important because many, maybe most, decision makers do not recognize that it must be considered when making clinical and policy decisions.

My Gold Standard is "Tarnished" or I Don't Have a Gold Standard

No Gold Standard

Number of Tests: 3 , Number of Cases: 1692
1 = Test Result Negative, 2 = Test Result Positive

Result of Test	1	2	3		
	1	1	1	Frequency	1513
	2	1	1	Frequency	23
	1	2	1	Frequency	59
	2	2	1	Frequency	12
	1	1	2	Frequency	21
	2	1	2	Frequency	19
	1	2	2	Frequency	11
	2	2	2	Frequency	34

Is the prevalence of disease in this group of patients high or low?

Number of Tests: 3 , Number of Cases: 1692
1 = Test Result Negative, 2 = Test Result Positive

Result of Test	1	2	3		
	1	1	1	Frequency	1513
	2	1	1	Frequency	23
	1	2	1	Frequency	59
	2	2	1	Frequency	12
	1	1	2	Frequency	21
	2	1	2	Frequency	19
	1	2	2	Frequency	11
	2	2	2	Frequency	34

Are the sensitivity and specificity of test 1 more like those of test 2 or test 3?

Number of Tests: 3 , Number of Cases: 1692
1 = Test Result Negative, 2 = Test Result Positive

Result of Test	1	2	3		
	1	1	1	Frequency	1513
	2	1	1	Frequency	23
	1	2	1	Frequency	59
	2	2	1	Frequency	12
	1	1	2	Frequency	21
	2	1	2	Frequency	19
	1	2	2	Frequency	11
	2	2	2	Frequency	34

Are the sensitivity and specificity of test 1 more like those of test 2 or test 3?

Number of Tests: 3 , Number of Cases: 1692
1 = Test Result Negative, 2 = Test Result Positive

Result of Test	1	2	3		
	1	1	1	Frequency	1513
	2	1	1	Frequency	23
	1	2	1	Frequency	59**
	2	2	1	Frequency	12*
	1	1	2	Frequency	21*
	2	1	2	Frequency	19**
	1	2	2	Frequency	11
	2	2	2	Frequency	34

*The result of test 1 agrees with the result of test 2 and disagrees with the result of test 3 (12 + 21 =33).

**The result of test 1 agrees with the result of test 3 and disagrees with the result of test 2 (59 + 19 =78).

Sample Output

Theta = prevalence of disease
Alpha = false positive rate (1 - alpha = specificity)
Beta = false negative rate (1 - beta = sensitivity)

Likelihood = -891.1428

Estimated Theta value = .0545 95% C.I.= (.040, .069)

Test	Estimated Beta	S.E.	95% C.I.
1	.2350	.067	(.105, .365)
2	.3565	.067	(.226, .487)
3	.2509	.067	(.119, .383)

Test	Estimated Alpha	S.E.	95% C.I.
1	.0109	.004	(.004, .018)
2	.0354	.005	(.026, .045)
3	.0100	.003	(.003, .017)

Bedside Diagnosis of Influenzavirus Infections in Hospitalized Children

Katherine A. Poehling, MD, MPH*¶; Marie R. Griffin, MD, MPH‡§; Robert S. Dittus, MD, MPH¶¶; Yi-Wei Tang MD, PhD‖; Kathy Holland, BS*; Haijing Li, BS‡; and Kathryn M. Edwards, MD*

Pediatrics. 2002 Jul;110(1 Pt 1):83-8.

ABSTRACT. *Objective.* For preventing nosocomial influenza infections and to facilitate prompt antiviral therapy, an accessible, rapid diagnostic method for influenzavirus is needed. We evaluated the performance of a lateral-flow immunoassay (QuickVue Influenza Test) completed at the bedside of hospitalized children during the influenza season.

Methods. All children who were evaluated at a large teaching hospital during the 1999 to 2000 influenza season were eligible if they were 1) younger than 19 years and hospitalized with respiratory symptoms or 2) younger than 3 years and hospitalized with fever. Each study child had 2 nasal swabs obtained—1 for influenzavirus culture and polymerase chain reaction (PCR) and the other for the QuickVue Influenza Test. The performance of the rapid diagnostic test was compared with the results of culture or PCR for influenza A or B.

Results. Of 303 eligible children, 233 (77%) were enrolled. In this population, 19 children had culture- and/or PCR-confirmed influenza A infection, prevalence of 8%. The QuickVue Influenza Test had a sensitivity of 74%, specificity of 98%, positive predictive value of 74%, and negative predictive value of 98%.

Conclusions. Among children hospitalized with fever/respiratory symptoms during the influenza season, negative bedside QuickVue Influenza Tests indicated very low likelihood of influenza infection, whereas positive tests greatly increased the probability of influenza-associated illness. *Pediatrics* 2002;110:83–88; *influenzavirus, diagnosis, rapid test, polymerase chain reaction, children.*

TABLE 2. Results From Viral Culture, QuickVue Influenza Test (Rapid), and 2 Consecutive PCRs

Culture	Rapid	PCR	Influenza Infection*	Total
+	+	+	+	8
+	−	+	+	2
+	−	−	+	1
−	+	+	+	6
−	+	−	−	5
−	−	+	+	2
−	−	−	−	209
11	19	18	19	233

* Defined as either a positive viral culture or 2 consecutive positive PCRs.

Output

Theta = prevalence of disease
Alpha = false positive rate (1 - alpha = specificity)
Beta = false negative rate (1 - beta = sensitivity)

Estimated Theta value = .0750 = **Prevalence** (.0729 in article)

Test	Estimated Beta	**Sensitivity**		
1	.4281	.5719	Culture	(.5789 in article)
2	.1998	.8002	QuickVue	(.7368 in article)
3	.0000	1.0000	PCR	(.9474 in article)
Test	Estimated Alpha	**Specificity**		
1	.0047	.9953	Culture	(1.0000 in article)
2	.0233	.9767	QuickVue	(.9766 in article)
3	.0024	.9976	PCR	(1.0047 in article)

Resources for Estimating Sensitivity and Specificity When there is No Gold Standard or the Gold Standard is "Tarnished"

Programs written by: S.D. Walter, Ph.D., Professor, McMaster University, Department of Clinical Epidemiology and Biostatistics, 1200 Main Street West, Room HSC 2C16, Hamilton, Ontario L8N 3Z5 Canada. E-Mail WALTER@FHS.MCMASTER.CA

These programs estimate the error rates of diagnostic tests or measurements when there is no gold standard. Maximum likelihood estimation methods are applied to latent class models representing the observed data.

1. LATENT1 (Version 3) - used when all the observations are subject to error, i.e. there are no gold standard measurements. There must be 3 or more observations per patient.
2. LATENT2 - used when there are 2 diagnostic measurements, and there are definitive gold standard assessments available in follow-up for patients with one or two positive results. Patients with both initial tests negative have no further observations made, and so may be true disease cases or true non-cases.
3. LATENT3 - similar to LATENT2, but there are three initial tests. Patients with 3 negative results have no further follow-up; other patients have a gold standard diagnosis available.

Latent Class Analysis in Other Software

- STATA 15
  - https://www.stata.com/new-in-stata/latent-class-analysis/
- Other software
  - Not available in SAS except as a plug-in program
  - Available in R, which is pretty much all plug-ins
  - Free-standing software
- None of this software is designed specifically for diagnostic tests, so all require substantial methodological expertise.

Other Resources for Using Latent Class Methods When the Gold Standard is Tarnished or Absent

- **Methods**
  - van Smeden M, Naaktgeboren CA, Reitsma JB, et al. Latent Class Models in Diagnostic Studies When There is No Reference Standard—A Systematic Review. *American Journal of Epidemiology*. 2014;179(4):423–431.
  - Chikere CMU, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard – An update. *PLOS ONE*. October 11, 2019, https://doi.org/10.1371/journal.pone.0223832.
  - STARD-BLCM: Standards for the Reporting of Diagnostic accuracy studies that use Bayesian Latent Class Models. 2017. http://www.equator-network.org/reporting-guidelines/stard-blcm/
  - Tanya Walsh. Fuzzy gold standards: Approaches to handling an imperfect reference standard. *Journal of Dentistry*. 2018. https://doi.org/10.1016/j.jdent.2018.04.022.
- **Examples**
  - Liu Y, Mwapasa V, Khairallah C, et al. Rapid Diagnostic Test Performance Assessed Using Latent Class Analysis. for the Diagnosis of Plasmodium falciparum Placental Malaria. *American Journal of Tropical Medicine and Hygiene*. 2016. 95(4), pp. 835–839.
  - Wiegand RE, Cooley G, Goodhew B, et al. Latent class modeling to compare testing platforms for detection of antibodies against the *Chlamydia trachomatis* antigen. *Scientific Reports*. 2018. *8:4232*. DOI:10.1038/s41598-018-22708-9.
  - Galappaththi-Arachchige HN, Holmen S, Koukounari A, Kleppa E, et al. Evaluating diagnostic indicators of urogenital *Schistosoma haematobium* infection in young women: A cross sectional study in rural South Africa. *PLoS ONE*. 2018. 13(2): e0191459. https://doi.org/10.1371/journal.pone.0191459.
  - Karaman BF, Acıkalın A, Unal I, et al. Diagnostic values of KOH examination, histological examination, and culture for onychomycosis: a latent class analysis. *International Journal of Dermatology*. 2018. doi: 10.1111/ijd.14255.

If it is possible to estimate sensitivity and specificity when there is no gold standard, isn't there a better way to estimate sensitivity and specificity when verification bias is present?

See LATENT2 and LATENT3 in the set of programs written by S.D. Walter, Ph.D.

John Collins, and Minh Huynh. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Statistics in Medicine*. 2014 October 30; 33(24): 4141–4169. doi:10.1002/sim.6218.

Xu Z, Meijuan Li. Statistical Considerations for Bias and Protocol Deviation in Medical Device Pivotal Clinical Study. *Therapeutic Innovation & Regulatory Science*. 2019, Vol. 53(5) 623-629. DOI:10.1177/2168479018804175

COURSE SUMMARY

The overall goal of this course is for students to learn quantitative tools that can be used to analyze and understand medical decisions.

- Diagnostic tests with dichotomous results
- Diagnostic tests with continuous results
- Prediction rules
- Introduction to cost-effectiveness analysis
- Costing / Analysis of cost / Discounting
- Mathematical modeling with decision trees
- Mathematical modeling with Markov techniques
- Measuring outcomes in "utility" terms
- Confidence intervals / sample size for cost-effectiveness analysis
- Economic assessment and policy analysis