Published in final edited form as:

Eur Urol. 2012 October ; 62(4): 590–596. doi:10.1016/j.eururo.2012.04.022.

How should we evaluate prediction tools? Comparison of three different tools for prediction of seminal vesicle invasion at radical prostatectomy as a test case

Giovanni Lughezzani^{1,2,*}, Kevin C. Zorn^{1,3,*}, Lars Budäus¹, Maxine Sun¹, David I. Lee⁴, Arie L. Shalhav³, Gregory P. Zagaya³, Sergey A. Shikanov³, Ofer N. Gofrit³, Alan E. Thong³, David M. Albala⁵, Leon Sun⁵, Angel Cronin⁶, Andrew J. Vickers⁷, and Pierre I. Karakiewicz¹

¹Cancer Prognostics and Health Outcomes Unit, University of Montreal Health Center, Montreal, QC, Canada

²Vita-Salute San Raffaele University, Milan

³University of Chicago Medical Center, Chicago, Illinois

⁴University of Pennsylvania, Philadelphia, Pennsylvania

⁵Duke University, Durham, North Carolina

⁶Harvard University, Boston, Massachusetts

⁷Memorial Sloan-Kettering Cancer Center, New York, New York

Abstract

Background—Statistical prediction tools are increasingly common in contemporary medicine but there is considerable disagreement about how they should be evaluated. Three tools (Partin tables, the European Society for Urological Oncology (ESUO) criteria and the Gallina nomogram) have been proposed for the prediction of seminal vesicle invasion (SVI) in patients with clinically localized prostate cancer. We aimed to determine which of these tool, if any, should be used clinically.

Methods—The independent validation cohort consisted of 2584 patients treated surgically for clinically localized prostate cancer between 2002 and 2007 at one of four North American tertiary-care referral centers. Traditional (area-under-the-receiver-operating-characteristic-curve (AUC), calibration plots, the Brier score, sensitivity and specificity, positive and negative predictive value) and novel (risk stratification tables, the net reclassification index, decision curve analysis and predictiveness curves) statistical methods quantified the predictive abilities of the three tested models.

Results—Traditional statistical methods (receiver operating characteristic (ROC) plots and Brier scores), as well as two of the novel statistical methods (risk stratification tables and the net reclassification index) could not provide clear distinction between the SVI prediction tools. For example, receiver operating characteristic (ROC) plots and Brier scores seemed biased against the binary decision tool (ESUO criteria) and gave discordant results for the continuous predictions of the Partin tables and the Gallina nomogram. The results of the calibration plots were discordant

Address all correspondence to: Pierre I. Karakiewicz, MD, FRCSC, Cancer Prognostics and Health Outcomes Unit, University of Montreal Health Center (CHUM), 1058, rue St-Denis, Montréal, Québec, Canada, H2X 3J4, Tel: 514-890-8000-35336, Fax: 514-412-7363, pierre.karakiewicz@umontreal.ca.

*Equal study contribution

with those of the ROC plots. Conversely, the decision curve clearly indicated that the Partin tables represent the ideal strategy for stratifying the risk of SVI.

Conclusions—Based on decision curve analysis results, surgeons should consider using the Partin tables to predict SVI. Decision curve analysis provided clinically meaningful comparisons between predictive models; other statistical methods for evaluation of prediction models gave inconsistent results that were difficult to interpret.

Keywords

prostate; prostatic neoplasms; prostatectomy; seminal vesicles; algorithms; statistics

Introduction

Statistical prediction tools are increasingly common in contemporary medicine. As just one example, a recent literature review identified over 100 models in prostate cancer alone [1]. This profusion of models begs the question of model evaluation: how do we know whether a model is a good one? Typical recommendations from the methodologic literature emphasize external validation – that is, testing a model on a data set other than that used to generate the model – and that “both calibration and discrimination should be evaluated” [2]. However, there is often little guidance as to how calibration or discrimination should be assessed or the results evaluated. For example, in the words of one group of well-regarded group of experts, researchers should “pre-specify acceptable performance of a model in terms of calibration and discrimination ...it is, however, unclear how to determine what is acceptable” [2].

We sought to evaluate three published models for seminal vesicle invasion (SVI) in prostate cancer patients, namely the 2007 update of Partin tables [3], the European Society for Urological Oncology (ESUO) criteria[4] and the nomogram developed by Gallina et al.[5]. Complete removal of seminal vesicles is commonly performed during radical prostatectomy for prostate cancer. The tip of the seminal vesicles is close to the arterial supply of the bladder base and to the proximal neurovascular bundles. Some investigators have suggested that sparing the tip of these structures would decrease erectile and urinary dysfunction [6–9]. Three small retrospective studies have demonstrated improved functional outcomes following seminal vesicle-sparing surgery [6–8]. In a fourth study, Albers et al. randomly subjected patients with localized prostate cancer either to a seminal vesicle-sparing approach or to a radical prostatectomy with complete seminal vesicle removal. The authors observed better urinary, but not erectile, outcomes in patients who underwent a seminal vesicle sparing radical prostatectomy [9].

Moreover, SVI is associated with poor prognosis after radical prostatectomy [10–13]. It is reasonable to assume that omission of seminal vesicle removal in patients who have cancer in the seminal vesicles would result in worse cancer control outcomes. As such, seminal vesicle sparing surgery should be restricted to men at low risk of SVI. Furthermore, higher likelihood of SVI may favor a wider resection of the prostate as well as the performance of pelvic lymphadenectomy.

Several published tools have been devised to determine SVI risk. We have previously published a comparison of the Gallina and Partin prediction models [14]. Our results were rather indeterminate, with one model (Gallina nomogram) showing better discrimination and the other (Partin tables) better calibration. Here we include evaluation of a third tool, the ESUO criteria, and use a broad range of evaluation methods. We hoped that our results would shed light not only on the best tool for SVI prediction, but on methods for statistical evaluation of prediction methods.

Materials and Methods

Study population, clinical and pathological assessment

The study population consisted of 2606 consecutive patients treated with radical prostatectomy for clinically localized prostate cancer at one of four North American tertiary care centers between 2002 and 2007. None of these patients were involved in the creation of any of the three prediction tools and as such, this constitutes an entirely independent, external validation. To comply with the inclusion criteria of the Gallina nomogram and of the 2007 Partin tables, 7 patients were excluded for clinical stage T₃–T₄, 11 patients for unknown clinical stage and 4 patients for PSA>45ng/mL, resulting in a study population of 2584 patients. The three prediction tools are based on similar, routinely collected data: clinical stage, biopsy grade and PSA; the Gallina nomogram and the ESUO criteria require, in addition, percentage of positive cores.

The clinical stage was assigned by the attending urologist according to the 1992–2002 American Joint Committee on Cancer (AJCC) TNM guidelines. In all men, pretreatment PSA was measured before digital rectal examination and transrectal ultrasound (TRUS). All patients underwent multi-core (≥ 10) TRUS-guided prostate biopsy. Biopsy Gleason grades were assigned by dedicated genito-urinary pathologists at each institution [15]. All prostate specimens were processed according to the Amin et al. protocol and were graded according to the Gleason system. In all patients complete removal of the seminal vesicles was performed.

Statistical analysis

For each model we calculated area-under-the-receiver-operating-characteristic-curve (AUC), calibration plots, the Brier score (mean square error), sensitivity and specificity, positive and negative predictive value. For the continuous predictors, namely Partin tables and the Gallina nomogram, we used a probability of 2.5% to dichotomize into positive and negative results for calculation of binary statistics such as sensitivity. We considered several novel techniques including risk stratification tables [16], the net reclassification index [17], decision curve analysis [18] and predictiveness curves [19]. Decision curve analysis incorporates the clinical consequences of using a prediction rule by applying a different weight to positive and false-positive results. A false-negative result (not removing cancerous seminal vesicles) has more serious consequences than a false-positive result (removal of seminal vesicles free of cancer). However, the weighting of false-negative and false-positive results can be varied according to patient preferences, or differences in opinion about the risks of the procedure. These preferences represent the threshold probability for action (seminal vesicle removal). For example, an individual who had a threshold probability of 3% would choose complete seminal vesicle removal when risk of SVI is 3% or greater, but seminal vesicle sparing when risk of SVI is less than 3%. For each threshold probability, decision curve analysis quantifies the net benefit of using a SVI predictive model relative to preserving seminal vesicle in all men. The optimal strategy is the one with the highest net benefit across the complete range of reasonable threshold probabilities. Net benefit can be interpreted as the number of true positive instances of SVI treated with seminal vesicle removal at surgery, if no individual without SVI was subjected to seminal vesicle removal.

A second novel method is the predictiveness curve. This plots cumulative proportion of predictions against absolute risk. The proportion of predictions thought to be in an “indeterminate” region – where risk is not obviously too low to warrant seminal vesicle removal or too high to warrant sparing – is indicative of a model’s value. The two other novel methods we considered, risk stratification and net reclassification, have been explicitly

deemed inappropriate for comparison of two models [16, 17] and so were not considered further. All analyses were conducted using Stata 11 (Stata Corp., College Station, Tx).

Results

Table 1 shows the characteristics of the study cohort (n=2854). The cohort is typical of stage-shifted US radical prostatectomy cohort, with most patients with low grade, early stage disease. Seminal vesicle invasion was observed in 109 (4.2%) patients.

Table 2 shows statistics for each model; ROC curves are given in figure 1 with calibration plots for the Partin and Gallina models shown in figure 2. The Gallina nomogram had the highest AUC (0.81 vs 0.79 for Partin and 0.69 for ESUO). The Partin tables had better calibration and very slightly better Brier score than the Gallina nomogram (0.0376 vs. 0.0382). No calibration plot could be plotted for ESUO, and Brier score was far inferior (0.506). Yet ESUO criteria had the highest negative predictive value, an important criterion for a decision not to treat. The Partin tables had the highest specificity and the Gallina nomogram the highest sensitivity, but it is unclear how the differences in specificity (23.2%) should be weighted against differences in sensitivity (3.7%). Figure 3 shows the predictiveness curves. Conclusions were highly sensitive to the definitions used for intermediate risk. If risks of 2–5% are considered intermediate, then the proportion of patients at intermediate risk is 14% for the Partin tables but 45% for the Gallina nomogram; if the range is 1–4%, the proportions are 55% and 43%, respectively.

Figure 4 shows the decision curves. Use of the Partin tables to determine treatment approach has the highest net benefit across the whole range of threshold probabilities, even for very low threshold probabilities likely associated with surgery. In particular, use of the Partin tables has higher net benefit than the current clinical strategy of seminal vesicle removal in all men. Table 3 shows the values for net benefit plotted on in the figures. Although differences between strategies are sometimes small, decision theory would hold that the strategy likely to lead to the best outcome should be chosen, irrespective of the size of the difference.

Table 3 also shows the advantage of using Partin tables to determine treatment instead of the current strategy of treating all men. The table gives *net* reduction in interventions. A difference of 31 for a threshold probability of 2%, can be interpreted as follows: “using the Partin tables to determine seminal vesicle resection is equivalent to a strategy that led to 31.2 fewer patients per 100 undergoing unnecessary seminal vesicle resection, but did not fail to treat any man with affected seminal vesicles”.

Discussion

We have evaluated three prediction tools, namely the 2007 Partin tables, the ESUO criteria and the Gallina nomogram, that have been proposed to inform clinical decisions about the removal of seminal vesicles at radical prostatectomy. We found that the traditional statistical methods were not of value for distinguishing between the three tools. Using sensitivity and specificity required us to dichotomize two continuous predictors (Partin and Gallina models) and it was not entirely clear whether increases in sensitivity were worth corresponding decreases in specificity. AUC and Brier score seemed biased against our binary decision tool (the ESUO criteria) and gave discordant results for which of the two continuous prediction models, namely Partin tables or Gallina nomogram, was optimal. The results of the calibration plot seemed to favor the Partin tables, although no calibration plot was possible for the binary predictor. The predictiveness curves were similarly restricted to comparison of Partin tables and Gallina nomogram, and gave inconsistent results depending on how

intermediate risk was defined. The two other novel evaluation tools, risk stratification tables and the net reclassification index, were also found to be inappropriate for comparison of published models.

In contrast, the decision curve analysis gave an unambiguous result applicable to both the continuous and binary models. With respect to ambiguity, the decision curve result stands by itself; in comparison, there is no need to trade-off sensitivity and specificity or compare calibration and discrimination. This was also true of the Brier score. However, the Brier score is not easily applicable to a binary predictor, such as the ESUO criteria. To further explore Brier scores in the assessment of the ESUO criteria, we set the sensitivity of the ESUO model to 100%, and increased its specificity by randomly reclassifying 20% of the population to ESUO negative SVI status. After this modification, the AUC of the ESUO criteria became almost identical to that of the Partin tables (0.793). The decision curve showed that the newly improved ESUO model was an excellent prediction tool, with the highest net benefit of all models from a threshold probability of 0.1% to 8%. However, the Brier score was still very poor (0.396). Even when we modified the ESUO tool until it was virtually perfect – 100% sensitivity, 90% specificity – its Brier score was still far inferior to that of the Gallina nomogram and Partin tables. This suggests that the Brier score is not applicable to tools that provide binary coded predictions, such as presence or absence of SVI. Conversely, the Brier score is applicable to prediction tools that provide a probability which can be quantified on a continuous scale, such as the Gallina nomogram and Partin tables.

It is highly noteworthy that of all tested methods, only the decision curve analysis provided information allowing to identify the most clinically useful model. Other methods, such as the Brier score or AUC analysis, failed to provide such information. For example, the AUC analysis found the Gallina nomogram superior to Partin tables. Conversely, the Brier score analysis reversed this preference order. However, imagine that new data were published indicating that the benefits to SV preservation were much lower than reported and the risks much greater. In this case, the threshold probability at which a surgeon might consider seminal vesicle resection would be much less than 1%. Statistics such as AUC would be affected leading us to conclude, for example, that the Gallina nomogram had superior discrimination. That neither model should be used to determine surgical approach - instead all men should be treated - would not be apparent. Decision curve analysis provides a clear answer to the question of comparative effectiveness: instead of having to wonder whether discrimination is “high enough” to warrant use of a model, or whether miscalibration is sufficiently severe to prevent its use, the approach with the highest net benefit is chosen.

A final advantage of the decision curve analysis is that it evaluates the consequences of using a prediction tool in clinical terms. Instead of providing abstract and potentially conceptually challenging numerical benchmarks, such as the concordance index – the probability of correct classification for a randomly selected discordant pair – or the Brier score – the mean square error between predictions and outcome - decision curve analysis is based on the numbers of cancers treated vs. unnecessarily performed surgeries. Such concept is clearly reflective of daily clinical decision making.

In conclusion, in the current population, the 2007 Partin tables are better than alternative tools in predicting the presence of SVI. Unless a clinician would consider seminal vesicle resection at a very low risk – less than 1% - using the Partin tables to determine resection would improve clinical results compared to the current strategy of seminal vesicle resection in all men. Moreover, our analyses showed that one methodological approach, namely decision curve analysis, appears best suited to identify the SVI prediction tool that provides the optimal prediction characteristics. Decision curve analysis has the ability to provide

clinically meaningful comparisons between predictive models and can readily determine whether use of a model would lead to better clinical decisions. Other statistical methods for evaluation of prediction models gave inconsistent results that were difficult to interpret.

Acknowledgments

Pierre I. Karakiewicz is partially supported by the University of Montreal Health Center Urology Specialists, Fonds de la Recherche en Santé du Québec, the University of Montreal Department Of Surgery and the University of Montreal Health Center (CHUM) Foundation. Andrew Vickers is supported in part by funds from David H. Koch provided through the Prostate Cancer Foundation, the Sidney Kimmel Center for Prostate and Urologic Cancers and P50-CA92629 SPORE grant from the National Cancer Institute to Dr. P. T. Scardino.

References

1. Shariat SF, Kattan MW, Vickers AJ, et al. Critical review of prostate cancer predictive tools. *Future Oncol.* 2009 Dec; 5(10):1555–1584. [PubMed: 20001796]
2. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. *BMJ.* 2009; 338:b605. [PubMed: 19477892]
3. Makarov DV, Trock BJ, Humphreys EB, et al. Updated nomogram to predict pathologic stage of prostate cancer given prostate-specific antigen level, clinical stage, and biopsy Gleason score (Partin tables) based on cases from 2000 to 2005. *Urology.* 2007 Jun; 69(6):1095–1101. [PubMed: 17572194]
4. Zlotta AR, Roumeguere T, Ravary V, et al. Is seminal vesicle ablation mandatory for all patients undergoing radical prostatectomy? A multivariate analysis on 1283 patients. *Eur Urol.* 2004 Jul; 46(1):42–49. [PubMed: 15183546]
5. Gallina A, Chun FK, Briganti A, et al. Development and split-sample validation of a nomogram predicting the probability of seminal vesicle invasion at radical prostatectomy. *Eur Urol.* 2007 Jul; 52(1):98–105. [PubMed: 17267098]
6. John H, Hauri D. Seminal vesicle-sparing radical prostatectomy: a novel concept to restore early urinary continence. *Urology.* 2000 Jun; 55(6):820–824. [PubMed: 10840084]
7. Sanda M, Dunn R, Wei J, et al. Seminal vesicle sparing technique is associated with improved sexual HRQOL outcome after radical prostatectomy. *J Urol, suppl.* 2002; 167:151. Abstract 606.
8. Bellina M, Mari M, Ambu A, et al. Seminal monolateral nerve-sparing radical prostatectomy in selected patients. *Urol Int.* 2005; 75(2):175–180. [PubMed: 16123574]
9. Albers P, Schafers S, Lohmer H, et al. Seminal vesicle-sparing perineal radical prostatectomy improves early functional results in patients with low-risk prostate cancer. *BJU Int.* 2007 Nov; 100(5):1050–1054. [PubMed: 17760889]
10. Epstein JI, Carmichael M, Walsh PC. Adenocarcinoma of the prostate invading the seminal vesicle: definition and relation of tumor volume, grade and margins of resection to prognosis. *J Urol.* 1993 May; 149(5):1040–1045. [PubMed: 8483205]
11. Jewett HJ, Bridge RW, Gray GF, et al. The palpable nodule of prostatic cancer. Results 15 years after radical excision. *JAMA.* 1968; 203:403. [PubMed: 5694122]
12. Villers AA, McNeal JE, Redwine EA, et al. Pathogenesis and biological significance of seminal vesicle invasion in prostatic adenocarcinoma. *J Urol.* 1990 Jun; 143(6):1183–1187. [PubMed: 2342179]
13. Secin FP, Bianco FJ Jr, Vickers AJ, Reuter V, Wheeler T, Fearn PA, et al. Cancer-specific survival and predictors of prostate-specific antigen recurrence and survival in patients with seminal vesicle invasion after radical prostatectomy. *Cancer.* 2006 Jun 1; 106(11):2369–2375. [PubMed: 16649221]
14. Zorn KC, Capitanio U, Jeldres C, et al. Multi-institutional external validation of seminal vesicle invasion nomograms: head-to-head comparison of Gallina nomogram versus 2007 Partin tables. *Int J Radiat Oncol Biol Phys.* 2009 Apr 1; 73(5):1461–1467. [PubMed: 18938046]
15. Gleason DF. Histologic grading of prostate cancer: a perspective. *Hum Pathol.* 1992 Mar; 23(3):273–279. [PubMed: 1555838]

16. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med.* 2008 Nov 18; 149(10):751–760. [PubMed: 19017593]
17. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med.* 2008 Jan 30; 27(2):157–172. discussion 207–12. [PubMed: 17569110]
18. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006 Nov-Dec;26(6):565–574. [PubMed: 17099194]
19. Huang Y, Sullivan Pepe M, Feng Z. Evaluating the predictiveness of a continuous marker. *Biometrics.* 2007 Dec; 63(4):1181–1188. [PubMed: 17489968]

Figure 1. Receiver-operating-characteristic curves

Gallina nomogram (thick black line); Partin tables (thin black line); ESUO criteria (thin grey line).

Figure 2. Calibration plots for the Gallina nomogram (top) and Partin tables (bottom)
 Ideal calibration (black line); model calibration (dashed line); observed proportion by quartiles, with 95% confidence intervals (forest plot).

Figure 3.
Predictiveness curves for Gallina nomogram (black line) and Partin tables (grey line).

Figure 4. Decision curve analysis

Gallina nomogram (black dashed line), Partin tables (black solid line), and ESUO criteria (grey dashed line) predictive models; the thick grey line indicates benefit of seminal vesicle dissection in all cases, thick black line indicates net benefit of seminal vesicle preservation in all cases.

Table 1

Baseline characteristics of the sample

Variable	Frequency (%) or Median (quartiles)
PSA	5.2 (4.0, 7.0)
Clinical Stage	
T1	2100 (81%)
T2a	381 (15%)
T2B+	103 (4%)
Biopsy Gleason Score	
6	1387 (54%)
7	1010 (39%)
8	187 (7%)
Year of surgery	
2002	61 (2%)
2003	123 (5%)
2004	410 (16%)
2005	459 (18%)
2006	788 (30%)
2007	743 (29%)
Cores: 60%+ positive	244 (9%)
Pathological Stage	
pT2a	374 (14%)
pT2b	477 (18%)
pT2c	1167 (45%)
pT3a	455 (18%)
pT3b	107 (4%)
pT4	107 (4%)
Seminal Vesicle Invasion	109 (4%)

Table 2

Evaluation of the three prediction tools using a variety of traditional statistics.

<i>Model</i>	AUC	Brier Score	Sensitivity	Specificity	PPV	NPV
Gallina nomogram	0.805	0.0382	92.7	33.1	5.7	99.0
Partin tables	0.792	0.0376	89.0	56.3	8.2	99.1
ESUO criteria	0.692	0.5062	90.8	47.6	7.1	99.2

Decision analytic evaluation. The 2007 Partin tables have highest net benefit for all threshold probabilities other than 1.5%, where they are equal to ESUO.

Table 3

Threshold probability	Net benefit				Net reduction in unnecessary seminal vesicle resections per 100 patients		
	Treat all	Gallina nomogram	Partin tables	ESUO criteria	Gallina nomogram	Partin tables	ESUO criteria
0.5%	0.0374	0.0374	0.0376	0.0358	0.00	4.61	-31.46
1.0%	0.0325	0.0325	0.0331	0.0332	0.00	5.65	7.24
1.5%	0.0276	0.0280	0.0306	0.0307	2.44	19.84	20.14
2.0%	0.0226	0.0235	0.0290	0.0281	4.02	31.19	26.59
2.5%	0.0176	0.0227	0.0268	0.0254	19.62	35.84	30.46
3.0%	0.0126	0.0208	0.0237	0.0228	26.79	35.95	33.04
3.5%	0.0074	0.0186	0.0215	0.0201	30.76	38.72	34.88
4.0%	0.0023	0.0159	0.0196	0.0174	32.66	41.68	36.26
4.5%	-0.0030	0.0135	0.0177	0.0146	35.01	43.72	37.34
5.0%	-0.0082	0.0118	0.0156	0.0119	37.96	45.36	38.20