A critical appraisal of logistic regression-based nomograms, artificial neural networks, classification and regression-tree models, look-up tables and risk-group stratification models for prostate cancer


*Cancer Prognostics and Health Outcomes Unit, University of Montreal, Montreal, Quebec, Canada, †Department of Urology, and ¶Martini Clinic – Prostate Cancer Center, University of Hamburg, Germany, ‡Department of Quantitative Health Sciences, Cleveland Clinic Foundation, Cleveland, Ohio, USA

Accepted for publication 31 October 2006
FKHC and PIK contributed equally to the manuscript

OBJECTIVE

To evaluate several methods of predicting prostate cancer-related outcomes, i.e. nomograms, look-up tables, artificial neural networks (ANN), classification and regression tree (CART) analyses and risk-group stratification (RGS) models, all of which represent valid alternatives.

METHODS

We present four direct comparisons, where a nomogram was compared to either an ANN, a look-up table, a CART model or a RGS model. In all comparisons we assessed the predictive accuracy and performance characteristics of both models.

RESULTS

Nomograms have several advantages over ANN, look-up tables, CART and RGS models, the most fundamental being a higher predictive accuracy and better performance characteristics.

CONCLUSION

These results suggest that nomograms are more accurate and have better performance characteristics than their alternatives. However, ANN, look-up tables, CART analyses and RGS models all rely on methodologically sound and valid alternatives, which should not be abandoned.

KEYWORDS

prediction tools, outcome, prostate cancer, nomogram, artificial neural network, classification and regression tree, look-up table, risk group stratification

INTRODUCTION

The field of prognostics has burgeoned in the last decade and clinicians have been provided with numerous tools to assist with evidence-based medical decision-making. Most of these decision aids consist of nomograms, artificial neural networks (ANNs), look-up tables, classification and regression tree analyses (CART) and risk group stratification (RGS) models [1–21]. They address numerous prostate cancer outcomes, which range from predicting biopsy outcome in men considered at risk of prostate cancer [1–9], to predicting the likelihood of Gleason upgrading between biopsy and radical prostatectomy (RP) pathology [10,11], to predicting side-specific extracapsular extension (SS-ECE) [12–14] at RP, to assessing the risk of biochemical recurrence [20–22], and to death from hormone-refractory cancer [15]. For some outcomes more than one model might be available, which makes model selection difficult. Model selection criteria might be proposed:

(i) The level of complexity represents an important consideration. Excessively complex models are clearly impractical in busy clinical practice. Similarly, models that require computational infrastructure might pose problems with their applicability. For example, ANNs can accurately predict several outcomes of interest [5–9], but the use of ANNs might be restricted due to lack of access to ANN code or lack of computer infrastructure. Look-up tables [10,16], such as the Partin Tables [16], RGS models [20,21], decision-trees based on CART models [12,17,18] or nomograms [1–3,11,13–15,19] represent user-friendly, paper-based alternatives, which bypass these problems.

(ii) Accuracy represents the most important consideration [23–30]. Current statistical methods offer the possibility of assessing a model's predictive accuracy. Usually, it is quantified using receiver operating characteristic area under the curve, and is expressed as a percentage. Values are 50–100%, where 50% is equivalent to the flip of a coin and 100% represents a perfect prediction. No model is perfect and generally accepted accuracy ranges are 70–80% [1–9,13–15,19]. Accuracy should be confirmed in an external cohort. Alternatively, statistical methods such as bootstrapping can be used to internally validate the model [31,32].

(iii) Performance characteristics represent another important consideration. Accuracy
FIG. 1. A comparison between a nomogram and look-up table to predict Gleason upgrading at RP with their corresponding nonparametric loess smoothing calibration plots. (A) the nomogram predicting Gleason sum upgrading between biopsy and RP pathology. BX Gleason Pri, primary biopsy Gleason score; BX Gleason Sec, secondary biopsy Gleason score; P (Upgrade), probability of biopsy Gleason sum upgrading. Instructions: To obtain the nomogram predicted probability of biopsy upgrading, locate patient values at each axis. Draw a vertical line to the ‘Point’ axis to determine how many points are attributed for each variable value. Sum the points for all variables. Locate the sum on the ‘Total Points’ line to be able to assess the individual probability of biopsy Gleason sum upgrading on the ‘P(Upgrade)’ line. (B) The calibration plot of a nomogram for predicting biopsy Gleason sum upgrading, where the x-axis represents the predicted probability and the y-axis represents the observed fraction with evidence of upgrading between the biopsy and final pathology. Perfect predictions correspond to the 45° line. Points estimated below the 45° line represent over-prediction, whereas those above the 45° line represent under-prediction. A nonparametric, smoothed curve indicates the relationship between predicted probability and observed frequency of biopsy Gleason sum upgrading. Vertical lines indicate the frequency distribution of predicted probabilities. (C) The look-up table predicting Gleason sum upgrading between biopsy and RP specimen. (D) The calibration plot of a look-up table for predicting biopsy Gleason sum upgrading, where the x-axis represents the predicted probability and the y-axis represents the observed fraction with evidence of upgrading between biopsy and RP pathology. The details of the over-prediction and under-prediction are as listed in (D). AUC, area under the receiver operating characteristic curve.

<table>
<thead>
<tr>
<th>PSA (ng/ml)</th>
<th>AJCC Stage 1992</th>
<th>Prostate gland volume</th>
<th>0.01</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>T1c, T2a</td>
<td>(0.19–0.35)</td>
<td>0.27</td>
<td>0.12</td>
<td>0.12</td>
<td>0.25</td>
<td>0.37</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>T2b, T2c</td>
<td>(0.05–0.25)</td>
<td>0.37</td>
<td>0.19</td>
<td>0.19</td>
<td>0.32</td>
<td>0.50</td>
<td>0.57</td>
</tr>
<tr>
<td>&gt;10–20</td>
<td>T1c, T2a</td>
<td>(0.37–0.63)</td>
<td>0.47</td>
<td>0.25</td>
<td>0.25</td>
<td>0.38</td>
<td>0.50</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>T2b, T2c</td>
<td>(0.08–0.32)</td>
<td>0.37</td>
<td>0.12</td>
<td>0.12</td>
<td>0.38</td>
<td>0.50</td>
<td>0.57</td>
</tr>
<tr>
<td>&gt;20</td>
<td>T1c, T2a</td>
<td>(0.38–0.81)</td>
<td>0.61</td>
<td>0.38</td>
<td>0.38</td>
<td>0.66</td>
<td>0.72</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>T2b, T2c</td>
<td>(0.25–0.75)</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
</tr>
</tbody>
</table>

The availability of several high-quality predictive models should encourage the clinician to adopt these tools into everyday clinical practice. Arguments favouring such behaviour include standardization of care and of decision-making. Moreover, it has been shown that nomograms predict more accurately than clinicians [19]. Most decision tools are based on thousands of observations, and it is virtually impossible to achieve that level of clinical exposure and expertise on an individual level. Moreover, most clinicians do not have the capacity to systematically record or remember the risk characteristics of several thousands of patients. Also, unlike computers, clinicians are incapable of systematically and cumulatively processing the recorded risk characteristics and outcomes of historic cases, and to derive an estimated probability of outcome for a new case at hand. These considerations might motivate clinicians to adopt the use of decision-tools.

Besides the methodological and practical considerations, patient perspective is also relevant when the use of most unbiased decision tools is considered. Patients are becoming increasingly aware of the existence of predictive tools. This trend is likely to increase in future; patients are also increasingly demanding to actively participate in decision-making.

Despite this advantage, decision tools are not meant to replace clinical judgement. Their
The Mantel-Haenszel test was used to test the look-up table predictive accuracy was 52.3%. To address this issue, we provide four representative examples of decision-aid comparisons. Those consist of comparisons between nomogram and look-up table, nomogram and CART model, nomogram and ANN, and nomogram and RGS model.

RESULTS

DIRECT COMPARISON OF A LOOK-UP TABLE AND A NOMOGRAM TO PREDICT GLEASON SUM UPGRADE BETWEEN BIOPSY AND RP PATHOLOGY

Previous studies [35–38] indicate that up to 43% of men with low grade prostate cancer at biopsy will be finally diagnosed with high-grade cancer at RP. The pathological Gleason score is a better predictor of biochemical recurrence than the biopsy Gleason score [39]. A high RP Gleason grade is associated with a higher rate of biochemical recurrence and worse cancer-specific survival [40–42]. Thus, Gleason sum upgrading from biopsy to final pathology might affect the treatment options [16,17,43]. To substantiate our hypothesis, we compared a nomogram and a previously published look-up table [11].

The total PSA level, clinical stage, primary and secondary biopsy Gleason scores were used as predictors in a logistic regression-based nomogram which addressed the probability of Gleason sum upgrading between biopsy and RP pathology in 2982 men (Fig. 1A). Two hundred bootstrap re-samples were used for internal validation of the accuracy estimates and to reduce overfit bias, and this gave a predictive accuracy of 80.4%. D’Amico et al. [10] previously reported a model for predicting pathological Gleason sum upgrading in the form of a look-up table (Fig. 1C) based on total PSA level, clinical stage and prostate gland volume, which was compared to the above nomogram. Within this direct comparison we applied the look-up table to a dataset of 2982 patients, according to the authors’ original specifications. The look-up table predictive accuracy was 52.3%. The Mantel-Haenszel test was used to test the statistical significance between the predictive accuracy of the nomogram and of the look-up table, and confirmed significance (80.4% vs 52.3%, P < 0.001). Finally, we explored the performance characteristics of the nomogram (Fig. 1B) and of the look-up table (Fig. 1D), to assess the rate of agreement between the predicted probability and the observed proportion of Gleason sum upgrading, across the entire range of predicted probabilities. The calibration plots of the nomogram and the look-up table are shown in Fig. 1B. The x-axes, respectively, represent the predicted probability; the y-axes represent the observed rate of Gleason upgrading, and the 45° line represents ideal predictions. The nomogram

<table>
<thead>
<tr>
<th>Points</th>
<th>Preoperative PSA</th>
<th>Clinical stage</th>
<th>Bx. Gleason sum</th>
<th>% + cores</th>
<th>% cancer</th>
<th>Total Points</th>
<th>Prob. of + ECE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 80</td>
<td>0, 70</td>
<td>T1a, T1b</td>
<td>4, 5, 6</td>
<td>0, 25</td>
<td>0, 3</td>
<td>0, 10</td>
<td>0.02</td>
</tr>
<tr>
<td>10, 15</td>
<td>10, 13</td>
<td>T1c, T2a</td>
<td>6, 7, 8</td>
<td>0, 45</td>
<td>0, 5</td>
<td>0, 25</td>
<td>0.10</td>
</tr>
<tr>
<td>20, 25</td>
<td>20, 17</td>
<td>T2b, T2c</td>
<td>7, 8, 9, 10</td>
<td>0, 65</td>
<td>0, 6</td>
<td>0, 50</td>
<td>0.20</td>
</tr>
<tr>
<td>30, 35</td>
<td>30, 23</td>
<td>T3a, T3b</td>
<td>8, 9, 10, 11</td>
<td>0, 80</td>
<td>0, 7</td>
<td>0, 75</td>
<td>0.30</td>
</tr>
</tbody>
</table>

FIG. 2. A comparison between a nomogram and CART model to predict SS-ECE with their corresponding nonparametric loess smoothing calibration plots. (A) The nomogram predicting SS-ECE at RP; BX Gleason sum, biopsy Gleason sum; % + cores, percentage of positive biopsy cores; % cancer, percentage of cancer in all cores; Prob. of + ECE, probability of ECE at RP. The instructions are as in Fig. 1. (B) The calibration plot of a nomogram for predicting SS-ECE at RP, where the x-axis represents the predicted probability and the y-axis represents the observed fraction with evidence of SS-ECE. Other details for predictions are as detailed in Fig. 1. (C) The CART model predicting SS-ECE. (D) The calibration plot of the CART model; details as in B.
FIG. 3. The comparison between a nomogram and an ANN to predict prostate cancer on initial biopsy with their corresponding nonparametric loess smoothing calibration plots. (A) The nomogram; DRE (0 not suspicious; 1, suspicious); Perc.fPSA, percentage of free PSA; BX Gleason sum; biopsy Gleason sum; % + cores, the percentage of positive biopsy cores; % cancer, percentage of cancer in all cores; P [prostate cancer on needle biopsy], the probability of prostate cancer on initial biopsy. The instructions on use are as for Fig. 1. (B) The calibration plot of the nomogram; details as in Fig. 1. (C) A schematic display of an ANN predicting prostate cancer on initial biopsy. (D) the calibration plot of the ANN; details as in Fig. 1.

The comparison between a nomogram and an ANN to predict prostate cancer on initial biopsy with their corresponding nonparametric loess smoothing calibration plots. (A) The nomogram; DRE (0 not suspicious; 1, suspicious); Perc.fPSA, percentage of free PSA; BX Gleason sum; biopsy Gleason sum; % + cores, the percentage of positive biopsy cores; % cancer, percentage of cancer in all cores; P [prostate cancer on needle biopsy], the probability of prostate cancer on initial biopsy. The instructions on use are as for Fig. 1. (B) The calibration plot of the nomogram; details as in Fig. 1. (C) A schematic display of an ANN predicting prostate cancer on initial biopsy. (D) the calibration plot of the ANN; details as in Fig. 1.

A NOMOGRAM TO PREDICT THE PROBABILITY OF SS-ECE

A NOMOGRAM TO PREDICT PROSTATE CANCER ON INITIAL BIOPSY

The concept of anatomical retropubic RP improved the potency and continence rates [44,45]. Despite its quality-of-life benefits, preservation of neurovascular bundles (NVBs) carries the risk of compromising cancer control and might result in a positive surgical margin. The risk is particularly high in the presence of ECE, which frequently occurs postero-laterally, where the NVBs are located [46]. Two models were recently published, a CART analysis [12] and a nomogram [13]; both predict SS-ECE. We tested both models in an external cohort consisting of 1118 men or 2236 prostate lobes, to compare their accuracy and performance characteristics.

The nomogram [14] (Fig. 2A) is based on continuously coded pre-treatment total PSA level, clinical stage and biopsy Gleason sum, the ipsilateral percentage of positive biopsy cores and the ipsilateral percentage of cancer in all cores. Two hundred bootstrap re-samples were used for internal validation of the accuracy estimates and to reduce overfit bias. The nomogram predictive accuracy was 84.0%. The CART [12] model (Fig. 2C) is based on pre-treatment total PSA level, ipsilateral number of positive biopsy cores and ipsilateral number of biopsy cores with Gleason 4/5. The model was applied to the same population as the nomogram. The CART model predictive accuracy was 70.0%. The comparison of the predictive accuracy estimates, using the Mantel-Haenszel test, had statistically significantly ($P<0.001$) higher accuracy for the nomogram. Finally, we explored the performance characteristics of the nomogram (Fig. 2B) and of the CART model (Fig. 2D), to assess the rate of agreement between predicted probability and the observed proportion of SS-ECE across the predicted probability range. The calibration plots of the nomogram and the CART model are shown in Fig. 2B,D. The nomogram calibration plot gave virtually ideal predictions, as the rate of predicted SS-ECE closely paralleled the observed rate of Gleason upgrading and virtually corresponded to the 45° line (Fig. 2B). Conversely, the CART predictions, which are represented by the logistic calibration curve, had appreciable differences from ideal predictions and worse performance than the nomogram predictions (Fig. 2D). Taken together, these findings show that the nomogram is statistically significantly more accurate than the CART model and performs better throughout the range of predicted probabilities.

DIRECT COMPARISON OF A CART MODEL AND A NOMOGRAM TO PREDICT THE PROBABILITY OF SS-ECE

DIRECT COMPARISON OF AN ANN AND A NOMOGRAM TO PREDICT PROSTATE CANCER ON INITIAL BIOPSY

There are several models predicting the probability of prostate cancer at initial needle biopsy. We identified one nomogram [3] and one ANN model [6] which were subjected to strict tests of accuracy and performance characteristics [4]. ANNs are computational methods that use multifactorial analysis; they contain layers of richly interconnected computing nodes, for which weights are adjusted when data are presented to the network during a ‘training’ process. Successful training can result in ANNs that predict output values or recognize patterns in multifactorial data [47].

The initial biopsy nomogram is based on four input variables, i.e. age, DRE findings, serum total PSA level and percentage free PSA,
and its predictive accuracy was originally estimated at 78.0% (Fig. 3A) [3]. The ANN [6] also includes prostate volume as a risk variable and its predictive accuracy was estimated at 84.0% (Fig. 3C). In several contemporary analyses, the prostate volume represents an important predictor of cancer risk on needle biopsy [48–50]. Thus, its inclusion should bias the ability of the ANN to predict more accurately than the nomogram, where this variable is not considered. Moreover, unlike the ANN, the nomogram variables are not allowed to interact with one another, which should further undermine the predictive ability of the nomogram. Both models were tested in a cohort of 3980 patients subjected to at least an 8-core initial biopsy [4]. Despite these ’a priori’ disadvantages, the results indicated that the nomogram (70.6%) was 3.4% (Mantel Haenszel test, P < 0.001) more accurate than the ANN (67.0%) [29,30]. Both models predicted less accurately than in the original studies [3,6]. The decrease in predictive accuracy relative to original data was probably related to the development of both tools on populations subjected virtually exclusively to sextant biopsies, while their direct comparison was in a cohort exposed to an extended–biopsy scheme. Finally, we explored the performance characteristics of the nomogram and of the ANN. The calibration plots of the nomogram and the ANN are shown in Fig. 3B,D. The nomogram calibration plot gave virtually ideal predictions, as the rate of predicted cancer on initial biopsy closely paralleled the observed rate of cancer on initial biopsy and virtually corresponded to the 45° line (Fig. 3B). Conversely, the ANN had important departures from ideal predictions (Fig. 3D), which were manifested by underestimation throughout the range of predicted probabilities. The most important departures were recorded for predicted probabilities of 10–70%, where most of the ANN predictions were situated. Taken together, this example of a direct comparison between a nomogram and an ANN shows that the nomogram is statistically significantly more accurate and is associated with better performance characteristics.

**DIRECT COMPARISON OF A RGS MODEL AND A NOMOGRAM TO PREDICT THE 5-YEAR BIOCHEMICAL RECURRENCE-FREE (BCRF) SURVIVAL AFTER RP**

Of all patients who have RP for prostate cancer, ~30% will have BCR in the form of increasing serum PSA levels [51]. As BCR proceeds there can be overt metastases by 8 years and BCR is often used as a surrogate marker for disease relapse [52]. Several prediction models were developed to predict BCR [20,22]. We compared a RGS model with a nomogram to compare the ability to predict 5-year BCR survival using preoperative total PSA level, clinical stage and biopsy Gleason score.

The above cancer characteristics were used as predictors in a logistic regression-based nomogram, which addressed the probability of BCR at 5 years after RP (Fig. 4A). We used a separate cohort of 1960 men to develop the nomogram; a cohort of 929 patients was used for external validation. The predictive accuracy was 71.7% using the nomogram.

D’Amico et al. [20] previously reported a model for predicting the risk of 5-year BCR after RP in form of a RGS model. Three risk groups were defined, including low risk (≤ 85% 5-year PSA failure-free rate: 1992 TNM clinical category T1c and T2a, PSA level ≤10 ng/mL and biopsy Gleason score ≤6), intermediate (≤60% 5-year PSA failure-free survival rate; clinical category T2b or PSA level >10 and ≤20 ng/mL, or biopsy Gleason score 7), and high (≤40% 5-year PSA failure-free survival rate, category T2c disease, or PSA level >20 ng/mL or biopsy Gleason score ≥8; Fig. 4C). In the context of this direct comparison we applied the risk groups to the external validation dataset of 929 patients, using the authors’ original specifications. The RGS model had a predictive accuracy of 65.5%. The Mantel–Haenszel test showed that the difference in predictive accuracy between the nomogram (71.7%) and the RGS model (65.5%) was highly significant (6.2%, P < 0.01). Finally, we explored the

![Graph](image_url)
performance characteristics of the nomogram (Fig. 4B) and the RGS model (Fig. 4D), to assess the rate of agreement between the predicted probability and the observed proportion of 5-year BCRF survival, across the entire range of predicted probabilities. The calibration plots of the nomogram and of the risk group stratification model are shown in Fig. 4B,D. The nomogram calibration plot had less pronounced departures from the ideal prediction than the RGS model. Taken together, these findings show that the nomogram is statistically significantly more accurate than the RGS model. In the data used for the comparison both models showed departures from the ideal predictions, but they were more pronounced for the RGS model. In addition, to the better performance, the nomogram predicted a wide range of probabilities, of 0.1–90%, whereas the RGS model provided only predictions of 40–86%. It might be postulated that a wider range is biologically more plausible than a narrow range limited to three levels. This is shown by better predictive accuracy and better calibration.

CONCLUSION

We provide theoretical guidelines for selecting a medical decision aid. Moreover, we also present practical examples of direct comparisons between various decision aids. These showed that nomograms are more accurate than ANNs, look-up tables, CART and RGS models.

ACKNOWLEDGEMENTS

Pierre I Karakiewicz is partially supported by the Fonds de la Recherche en Santé du Québec, the CHUM Foundation, the Department of Surgery and Les Urologues Associés du CHUM.

CONFLICT OF INTEREST

None declared.

REFERENCES

22 Kattan MW, Eastham JA, Stapleton...


40 Stamey TA, McNeal JE, Yenmo CM et al. Gleason sums of 7 lose valuable prognostic information in comparison to estimates of % grade 4 and 5 cancer. J Urol 1997; 157: 204, Abstract

41 Humphrey PA, Frazier HA, Vollmer RT, Paulson DF. Stratification of pathologic features in radical prostatectomy specimens that are predictive of elevated initial postoperative serum prostate-specific antigen levels. Cancer 1993; 71: 1821–7


43 Ackerman DA, Barry JM, Wicklund RA, Olson N, Lowe BA. Analysis of risk factors associated with prostate cancer extension to the surgical margin and pelvic node metastasis at radical prostatectomy. J Urol 1993; 150: 1845–50


45 Burkhard FC, Kessler TM, Fleischmann A, Thalmann GN, Schumacher M, Studer UE. Nerve sparing open radical retropubic prostatectomy – does it have an impact on urinary continence? J Urol 2006; 176: 189–95


49 Karakiewicz PI, Baiznet M, Aprikian AG et al. Outcome of sextant biopsy according to gland volume. Urology 1997; 49: 55–9


52 Pound CR, Partin AW, Eisenberger MA, Chan DW, Pearson JD, Walsh PC. Natural history of progression after PSA elevation following radical prostatectomy. JAMA 1999; 281: 1991–7

Correspondence: Pierre I. Karakiewicz, Cancer Prognostics and Health Outcomes Unit, University of Montreal Health Center (CHUM), 1058, rue St-Denis, Montréal, Quebec, Canada, H2X 3J4. e-mail: pierre.karakiewicz@umontreal.ca

Abbreviations: ANN, artificial neural network; CART, classification and regression-tree analysis; RGS, risk group stratification; RP, radical prostatectomy; SS-ECE, side-specific extracapsular extension; BCRF, biochemical recurrence-free; NVB, neurovascular bundle.