Journal of Biomedical Informatics 56 (2015) 87–93

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Improving patient prostate cancer risk assessment: Moving from static, globally-applied to dynamic, practice-specific risk calculators

Andreas N. Strobl [a,b,*], Andrew J. Vickers [c], Ben Van Calster [d], Ewout Steyerberg [e], Robin J. Leach [f,g], Ian M. Thompson [g], Donna P. Ankerst [a,b,g,h]

[a] TU München, Department of Mathematics, Munich, Germany
[b] HelmholtzZentrum München, Institute of Computational Biology, Munich, Germany
[c] Memorial Sloan-Kettering Cancer Center, Department of Epidemiology and Biostatistics, New York City, NY, USA
[d] KU Leuven, Department of Development and Regeneration, Leuven, Belgium
[e] Erasmus MC, Department of Public Health, Rotterdam, The Netherlands
[f] University of Texas Health Science Center at San Antonio, Department of Cellular and Structural Biology, San Antonio, TX, USA
[g] University of Texas Health Science Center at San Antonio, Department of Urology, San Antonio, TX, USA
[h] University of Texas Health Science Center at San Antonio, Department of Epidemiology and Biostatistics, San Antonio, TX, USA

ARTICLE INFO

Article history:
Received 18 December 2014
Revised 14 March 2015
Accepted 4 May 2015
Available online 16 May 2015

Keywords:
Prediction
Discrimination
Calibration
Prostate cancer
Logistic regression
Revision

ABSTRACT

Clinical risk calculators are now widely available but have generally been implemented in a static and one-size-fits-all fashion. The objective of this study was to challenge these notions and show via a case study concerning risk-based screening for prostate cancer how calculators can be dynamically and locally tailored to improve on-site patient accuracy. Yearly data from five international prostate biopsy cohorts (3 in the US, 1 in Austria, 1 in England) were used to compare 6 methods for annual risk prediction: static use of the online US-developed Prostate Cancer Prevention Trial Risk Calculator (PCPTRC); recalibration of the PCPTRC; revision of the PCPTRC; building a new model each year using logistic regression, Bayesian prior-to-posterior updating, or random forests. All methods performed similarly with respect to discrimination, except for random forests, which were worse. All methods except for random forests greatly improved calibration over the static PCPTRC in all cohorts except for Austria, where the PCPTRC had the best calibration followed closely by recalibration. The case study shows that a simple annual recalibration of a general online risk tool for prostate cancer can improve its accuracy with respect to the local patient practice at hand.

1. Introduction

Clinical risk prediction tools are now widely available on the internet and provide a valuable decision-aid to doctors and patients regarding treatment choices. There are currently hundreds of clinical risk prediction tools available online, with objectives ranging from the prediction of onset of disease for use in screening to prognosis of outcomes following treatment for disease [1–3]. Interestingly, despite the recent interest in personalized approaches to medicine, the big data daily flowing into clinical practices, and changes in patient populations and clinical practice over time, these risk calculators have generally remained static and applied in a one-size-fits-all fashion. For instance, 2013 US national guidelines for the prevention of cardiovascular diseases prescribed statins for persons with elevated risk based on a global score that was developed using a pooled cohort of patients monitored from the late 1980s to the early 2000s [4]. Subsequent validations on five external cohorts showed that the recommended risk score would greatly overestimate actual risk on contemporary populations, with up to 40–50% of the millions classified as high-risk in fact over-prescribed [5]. The widespread availability of electronic medical data raises the possibility that such models could instead evolve over time, automatically changing in tandem with evolving

Abbreviations: AUC, area under the ROC curve; BIC, Bayesian information criterion; DRE, digital rectal exam; EMR, electronic medical record; HLS, Hosmer–Lemeshow test statistic; MCMC, Markov Chain Monte Carlo; PBCG, prostate biopsy collaborative group; PCPT, prostate cancer prevention trial; PCPTRC, prostate cancer prevention trial risk calculator; PSA, prostate specific antigen; ROC, receiver operating characteristic; SABOR, San Antonio center of biomarkers of risk for prostate cancer.

* Corresponding author at: Department of Mathematics, Technische Universität München, M12, Boltzmannstr. 3, D-85747 Garching near Munich, Germany. Tel.: +49 89 289 18390; fax: +49 89 289 18435.
E-mail address: a.strobl@tum.de (A.N. Strobl).

http://dx.doi.org/10.1016/j.jbi.2015.05.001
1532-0464/

global clinical practice patterns [6]. Within individual hospitals, the ability to capitalize electronic medical record (EMR) data would additionally permit tailoring of a global risk tool to the hospital-specific patient population at hand, for example, allowing a different dynamic evolution of predictions for high-risk clinically referred versus healthy screening institutions.

As the case study to be investigated in this article, the Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) is a static risk tool that predicts the likelihood of detecting prostate cancer if a prostate biopsy were to be performed. It uses as inputs the commonly collected clinical risk factors: prostate-specific antigen (PSA), digital rectal exam (DRE), age, race, family history of prostate cancer, and prior biopsy history [3]. The model it is based on was developed using prostate biopsy data from participants on the placebo arm of a very unique prostate cancer prevention trial, the PCPT [7]. The PCPT provided the only patient population ever to be free of selection bias because at the end of seven years on the study all participants were requested to undergo prostate biopsy even if they lacked a clinical indication for biopsy (n = 5519) [8]. The posting of the calculator online in 2006 facilitated subsequent external validation on a range of cohorts that differed both in terms of patient composition and date of collection [9–21]. The latter was important since a shift in prostate biopsy practice occurred after the PCPT was completed: the number of sampled tissue cores on biopsy increased from 6 cores (3 on each side) in the PCPT to the now contemporary practice of 12 cores (6 on each side). It has been documented that a greater number of biopsy cores retrieved at biopsy increases the chance of detection of prostate cancer [22].

Statistical approaches to updating an existing risk prediction tool have been proposed, ranging from simple adjustment of the intercept of a model to re-estimation of multiple coefficients in the original model [23]. One-time updating approaches have been implemented in a variety of clinical settings, resulting in improved diagnostic or prognostic performance [24–29]. The need for continual temporal recalibration of a risk tool has been emphasized [30,31], along with the concept of transfer learning from similar hospitals when sample sizes at individual institutions are low [32].

In an era where patient data are housed electronically, risk prediction tools could and should be automatically updated with local data as soon as such data arrive. The objective of this study was to challenge the ubiquitous notion of static universal risk prediction and show via a case study how prediction can easily be adapted to the patient data on-site, and thus improve the accuracy of prediction for local patients.

2. Methods

2.1. Participants and biopsy results

Five international cohorts from the Prostate Biopsy Collaborative Group (PBCG) were used to compare various methods for developing an institution-specific risk calculator. These have been previously described [21]. Three screening cohorts, the San Antonio Center of Biomarkers of Risk for Prostate Cancer study (SABOR), Texas, U.S., ProtecT, UK, and Tyrol, Austria followed primarily a 10-core biopsy scheme. Two clinical cohorts from the U.S., Cleveland Clinic, Ohio and the Durham VA, North Carolina, comprised patients referred for clinical symptoms. Those three cohorts used mixed biopsy schemes, but primarily 10- to 14-cores. Not all cohorts had all of the PCPTRC risk factors available; only those risk factors that were missing in less than 15% of the cases were used in the analysis. Biopsy records with associated PSA values higher than 50 ng/ml or with unknown Gleason grade were excluded. If cohorts had only few biopsies in the beginning and ending years, those years were aggregated into the first and last year. The number of biopsies per year in the resulting data set ranged from 73 (Durham) to 1106 (ProtecT).

2.2. PCPTRC

A modification of version 2.0 of the PCPTRC was used for the methods that tailored an existing risk tool [33]. While PCPTRC 2.0 provides separate estimates of the risks of low- versus high-grade prostate cancer, for this study a logistic regression of any prostate cancer was performed using the same dataset and the same covariates as the PCPTRC model: PSA, age, DRE, first-degree family history of prostate cancer, race (African American versus not) and history of a prior biopsy. When a risk factor was missing in more than 15% of biopsies in a cohort, it was not used in the analysis. This was the case for three of the binary covariates: African American race, prior biopsy and family history. Eight separate logistic regressions were run for each possible combination of missing values from these three variables and the corresponding model was used for the cohort. The PCPTRC logistic regression models are given in Table 1 of the Supplementary Appendix.

2.3. Validation sets and metrics

The different statistical methods for annually updating a risk tool were compared using each consecutive year, starting with year 2, as the validation set, and all past years as a training set. In this manner the training set grew cumulatively in size with each year and the validation set changed each year. To compare methods in absence of a fluctuating validation set, the process was repeated using a fixed validation set consisting of the biopsies in the last three years of each cohort. The methods were compared in terms of discrimination and calibration. Discrimination was measured using the area-underneath-the-receiver-operating-characteristic-curve (AUC), which equals the probability that for a randomly chosen cancer case/control pair, the case has a higher predicted risk of cancer. AUCs vary from 50% (chance discrimination) to 100% (perfect discrimination), with higher values indicating better discrimination. Ninety-five percent confidence intervals (95% CI) for AUCs were calculated using non-parametric U-statistics as commonly implemented in statistical packages. Calibration was measured via the Hosmer–Lemeshow statistic (HLS), which provides a single summary of the commonly used calibration plots. For each method of estimating risk, patients in the validation set were grouped into ten decile groups according to estimated risk: patients with the lowest 10th percentile of risks, risks in the 10th to 20th percentile and so on up to patients with the highest 10th percentile of risks. The observed rate of prostate cancer in each of the decile groups was computed (O_g) and compared to the mean of the n_g estimated risks in each decile group (E_g). The HLS equals the sum $\sum_{g=1}^{10} \frac{n_g(O_g-E_g)^2}{E_g(1-E_g)}$, with larger values indicating poorer fit; 95% CIs for the HLS were generated from 200 bootstrapped samples stratified by outcome.

2.4. Statistical methods

Details of the individual methods follow.

2.4.1. PCPTRC

This method performed no model building or augmentation and thus tests the value of a static model. For each individual in the training set the PCPTRC score was computed, allowing for missing values for some of the variables; see Supplementary Appendix, Table 1.