Journal of Data Science 21 (2), 177–192
April 2023
DOI: 10.6339/22-JDS1078
Data Science in Action

Active Data Science for Improving Clinical Risk Prediction

Donna P. Ankerst[1,2,*] and Matthias Neumair[1]

[1]*Depts. of Mathematics and Life Science Systems, Technical University of Munich, Germany*
[2]*Munich Data Science Institute, Technical University of Munich, Germany*

Abstract

Clinical risk prediction models are commonly developed in a post-hoc and passive fashion, capitalizing on convenient data from completed clinical trials or retrospective cohorts. Impacts of the models often end at their publication rather than with the patients. The field of clinical risk prediction is rapidly improving in a progressively more transparent data science era. Based on collective experience over the past decade by the Prostate Biopsy Collaborative Group (PBCG), this paper proposes the following four data science-driven strategies for improving clinical risk prediction to the benefit of clinical practice and research. The first proposed strategy is to actively design prospective data collection, monitoring, analysis and validation of risk tools following the same standards as for clinical trials in order to elevate the quality of training data. The second suggestion is to make risk tools and model formulas available online. User-friendly risk tools will bring quantitative information to patients and their clinicians for improved knowledge-based decision-making. As past experience testifies, online tools expedite independent validation, providing helpful information as to whether the tools are generalizable to new populations. The third proposal is to dynamically update and localize risk tools to adapt to changing demographic and clinical landscapes. The fourth strategy is to accommodate systematic missing data patterns across cohorts in order to maximize the statistical power in model training, as well as to accommodate missing information on the end-user side too, in order to maximize utility for the public.

Keywords *logistic regression; missing data; prostate cancer; risk calculator*

1 Introduction

The Prostate Biopsy Collaborative Group (PBCG) was formed in order to mobilize the collection of contemporary clinical data for the improvement of prostate cancer etiologic research and prediction (Vickers et al., 2010). Two major online prostate cancer risk tools were available at the formation of the PBCG, which had been built from large prostate cancer screening trials performed in North America and Europe (Roobol et al., 2012; Thompson et al., 2006). However, clinical practice and technology had changed since the time of those trials, potentially invalidating or at least diminishing the use of the models built on them for patients receiving contemporary care. The large expense, duration and conflicting findings meant that similar trials were unlikely to be performed in the future. Clinical trials or other well-resourced studies supply high-quality data at a level hard to reach in daily clinical practice. The PBCG initially collected retrospective clinical data, and then switched to prospective data collection on the

*Corresponding author. Email: ankerst@tum.de.

Received July 11, 2022; Accepted November 8, 2022

Figure 1: World map of the eleven PBCG participating institutions collecting prostate biopsy data.

standard prostate cancer risk factors and prostate biopsy outcomes from heterogeneous clinical practices across North America and Europe (Figure 1, Vickers et al. (2010)). The PBCG continues prospective collection and expansion to new partnering institutions to this day.

More than a decade after the PBCG commenced, reflection shows the practical lessons for more actively participating in the design, analysis and communication of clinical risk tools in order to maximize their use by the public. Specifically, this report advocates four principles for constructing clinical risk tools: 1) actively design prospective data collection, 2) make risk tools and corresponding code available online, 3) dynamically update risk tools and tailor to individual clinical centers, and 4) accommodate missing data both for training risk models and for the end users. The logistics behind putting these principles into action by the PBCG will be discussed in turn in the following sections.

2 Actively Design Prospective Data Collection

The PBCG started as a collection of retrospective prostate biopsy outcome and risk factor data sets from several institutions in order to understand the variability in risk factor outcome associations across institutions (Vickers et al., 2010). Biopsies in the data sets had been performed from the late 1990's to mid-2000's and only the six standard risk factors ubiquitously recommended for assessment for prostate biopsy referral were requested. These included prostate-specific antigen (PSA), digital rectal examination (DRE), race, age, first-degree prostate cancer family history and whether or not a prior prostate biopsy had been performed that was negative for prostate cancer; patients with prior positive prostate biopsies or other types of prostate cancer diagnoses were excluded from all analyses. Missing data from the prospective PBCG data collection from 2006 to 2019 are given in the heatmap shown in Figure 2. The graph shows percent missing data according to risk factor on the y-axis and cohort on the x-axis. Both axes have been sorted so that risk factor – cohort combinations with the highest amount of missing data appear in the lower right of the graph. This graph is the starting point for cohort-specific identification of missing data issues, to be followed by discussion with individual data-providers.

From the graph one can quickly see that all cohorts complied with provision of PSA and age, but problematic variable collection began with African ancestry and cohort 11 had the most

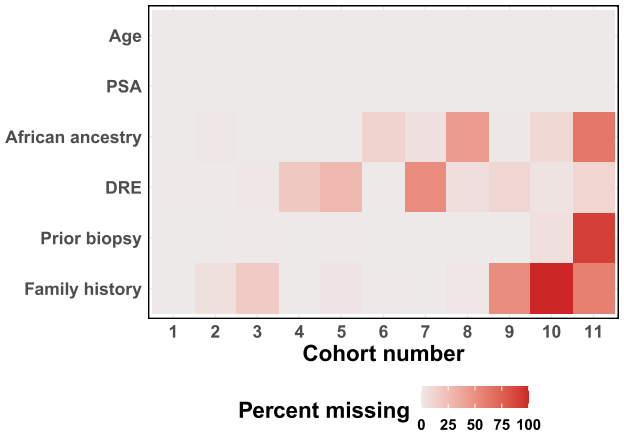

Figure 2: Overview of missing data for 18243 prostate biopsies collected from 11 PBCG cohorts between 2006 and 2019 with number of prostate biopsies performed in each cohort ranging from 243 to 5540, with median 1498.

missing data. The graph from the retrospective PBCG data was much worse and not reproduced here because the study has been closed. Some risk factors were intermittently missing for some patients but not others, and others were systematically missing in that some cohorts did not collect them at all, for example, race is often not collected in Europe due to data laws. The retrospective data presented an additional barrier in that risk factor coding sometimes varied across institutions, with some fields entered as text. Many risk factors were out of a valid range, meaning an error must have occurred. So although data were not missing, they were also not usable.

Requested data from an institution arrives pooled over a time period, but a helpful quality control measure visualizes summary statistics according to year collected. For the retrospective PBCG data collected up until 2008, the PBCG inspected number of biopsies performed, the outcome percent of biopsies positive for cancer, and prevalence of risk factors versus year the biopsy was performed for each institution, thus revealing anomalies within certain institutions and years that should be excluded (Strobl et al., 2015). For example, for one institution, during the earlier years 1992 and 1993, the provided data indicated that 100% of the biopsies performed were positive for prostate cancer. This cannot realistically occur in any clinical practice, and provided empirical evidence of mistakes during local data extraction. Further investigation in collaboration with the institution revealed that only biopsies positive for prostate cancer and not negative biopsy cases were stored in the database during earlier years. The data for all years had been extracted simultaneously via a single query, enabling this error to escape the attention of the local database programmer. All data from the problematic early years for the institution in question were subsequently deleted and examination of data by year of collection became mandated in the prospective PBCG.

Finally, as an additional quality control measure, the PBCG showed institution data contributors how risk factor outcome associations for their cohort compared to the overall average (Ankerst et al., 2018). These plots were most easily visualized for univariate rather than multivariate analyses, one risk factor at a time, and by dichotomizing the risk factors into low- versus high-risk values; an example is shown in Figure 3. This type of exploratory analysis can be helpful for institutions to quality check their own procedures by identifying outliers in either the

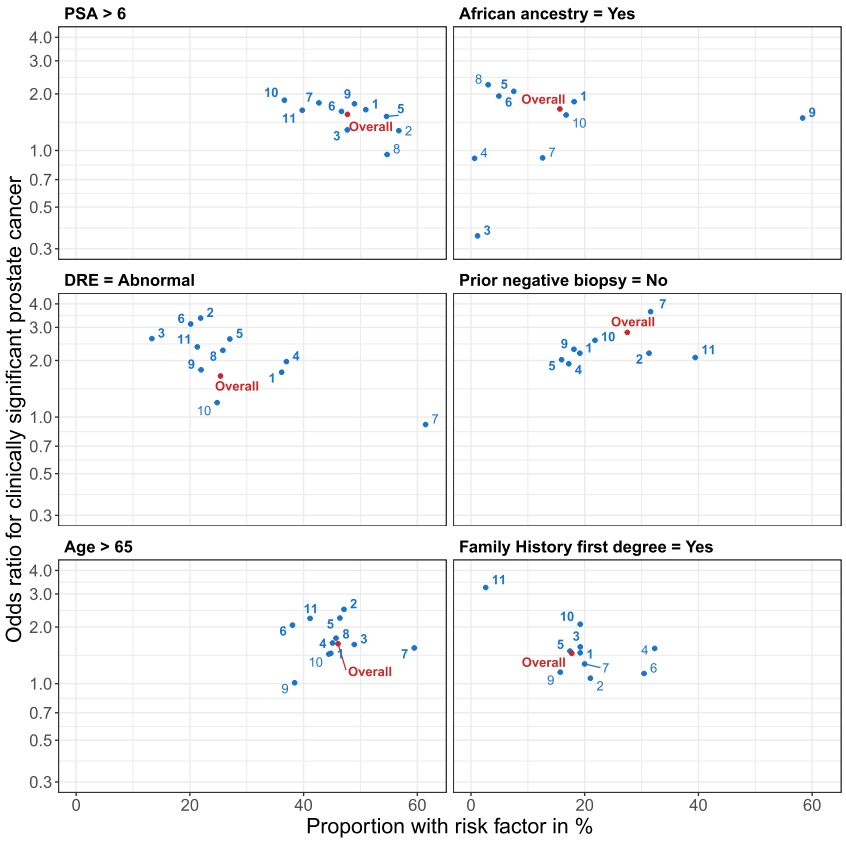

Figure 3: Odds ratios versus prevalence of risk factors evaluated on the same PBCG data used in Figure 2. Individual cohorts are indicated by number (not necessarily matching those of Figure 2), the estimate for the pooled data by "Overall". Bold indicates significance at the 0.05 level.

prevalence of a risk factor or the association with the outcome. For example, one of the participating cohorts was an outlier for both the risk factor and odds ratio association with prostate cancer. It had an unusually high proportion of abnormal DRE's with the lowest association with prostate cancer. This led the contributing investigator to query the clinical practice, finding that too many of the DRE's, which are subject assessments of whether lumps are present, were assigned as suspicious. The visualization thus led to a change in practice at the institution.

In response to shortcomings of the retrospective data and improvements worldwide in prostate biopsy techniques, including recommended increases in the number of core samples, the PBCG received a grant from 2013 to 2018 to prospectively collect contemporary data on par with routine best practice for clinical and prevention trials. Partner institutions were identified who would provide prospective data over a five-year period following local institutional review board approval. Participating institutions included the University of Texas Health Science Center at San Antonio, Memorial Sloan Kettering Cancer Center (MSKCC), Mayo Clinic, University of Zurich, United States Department of Veterans Affairs (VA) Durham and Puerto Rico affiliates, San Raffaele Hospital in Italy, Sunnybrook Health Systems in Canada, University of California San Francisco, and Hamburg University Hospital in Germany. Statistical analyses were performed on anonymized data following human research ethics approval by the Technical University of Munich medical ethics board in Germany as well as at MSKCC.

As a first step to prospective data collection, a set of common data elements (CDEs) were determined as the set of risk factors and outcomes needed to form an updated accurate risk prediction tool for prostate cancer on biopsy utilizing the latest technologies. Decisions on the CDEs involved meetings of multiple stakeholders, including participating statisticians and epidemiologists who would be analyzing the data, and clinicians who perform the biopsies. Committee determination of CDEs is standard practice among consortia aiming to collect standardized data across multiple electronic health record systems or institutional platforms for pooled analysis (Grinspan et al., 2021; Patel et al., 2006; Westra et al., 2020; Austin et al., 2020).

To improve data completeness and quality, CDEs were converted to standardized electronic case report forms via the localized Microsoft Access/Excel database software system. On site health specialists entered data via user-friendly interface pages that included immediate error notifications for out-of-range entries as well as prompts for empty fields (Figure 4). Notifications had to be resolved before automatic transportation to the linked Excel spreadsheet. A locked version of the cumulative spreadsheet containing only anonymized data was sent to the PBCG centralized data center for secure storage at MSKCC at approximately monthly intervals, where it was immediately reviewed by statisticians, who reported the data summaries back to the local institutions along with any identified issues with the data. The timely review ensured that central statisticians could work with the local data suppliers to adjust any errors while it was still fresh on their minds. As further aims of the grant were to instill best data practices by incorporating CDEs into the institutional electronic health record (EHR) systems, especially for sites with high minority representation, select sites received on-site training by PBCG investigators.

The user-friendly Microsoft Access/Excel framework is popular among tissue banks and other registries due to its ubiquitous presence in laboratory and clinical practice; it has been used for example by the Cooperative Prostate Cancer Tissue Resource (Patel et al., 2006). In addition to user-friendly data entry systems, timely communication between central statisticians and local data collectors, verbally, in-person and electronically, treating all individuals as equal partners, is key motivation for obtaining high-quality data.

The PBCG completed analyses on the prospective data collected as part of its funded grant from 2018 to 2022. During this period, the prostate biopsy procedure experienced a change to magnetic resonance image (MRI)-guided biopsies as opposed to the standard 10 to 12 core procedure. This led to the formation of the international PBCG focused on collecting only MRI-biopsies and extending to sites in Asia and Australia, which is currently ongoing.

In 2015, with the rise of clinical risk model development and validation publications of varying quality, a group of interdisciplinary stakeholders convened to develop the Transparent Reporting of multivariable prediction models for Individual Prognosis Or Diagnosis (TRIPOD) statement (Collins et al., 2015). The TRIPOD statement comprises a checklist of 22 items to include in publications on risk models, grouped by requested descriptions in all sections of the paper, from the title/abstract, introduction, methods, results, discussion to the supplementary information. It has its own website and has been reproduced in many publications, but is also included in the Supplementary Appendix to this paper for easy reference. It is not uncommon for journals to request this checklist upon submission of an article on risk model development. Included in the methods specifications are requirements for clear definitions of all outcomes and risk factors, as well as transparent reporting of missing data as espoused here. The TRIPOD statement was not available until the end of the PBCG period and was written as a checklist for publications. However, given the PBCG experience, it is recommended to address the checkpoints in the TRIPOD statement as part of the prospective design of data collection.

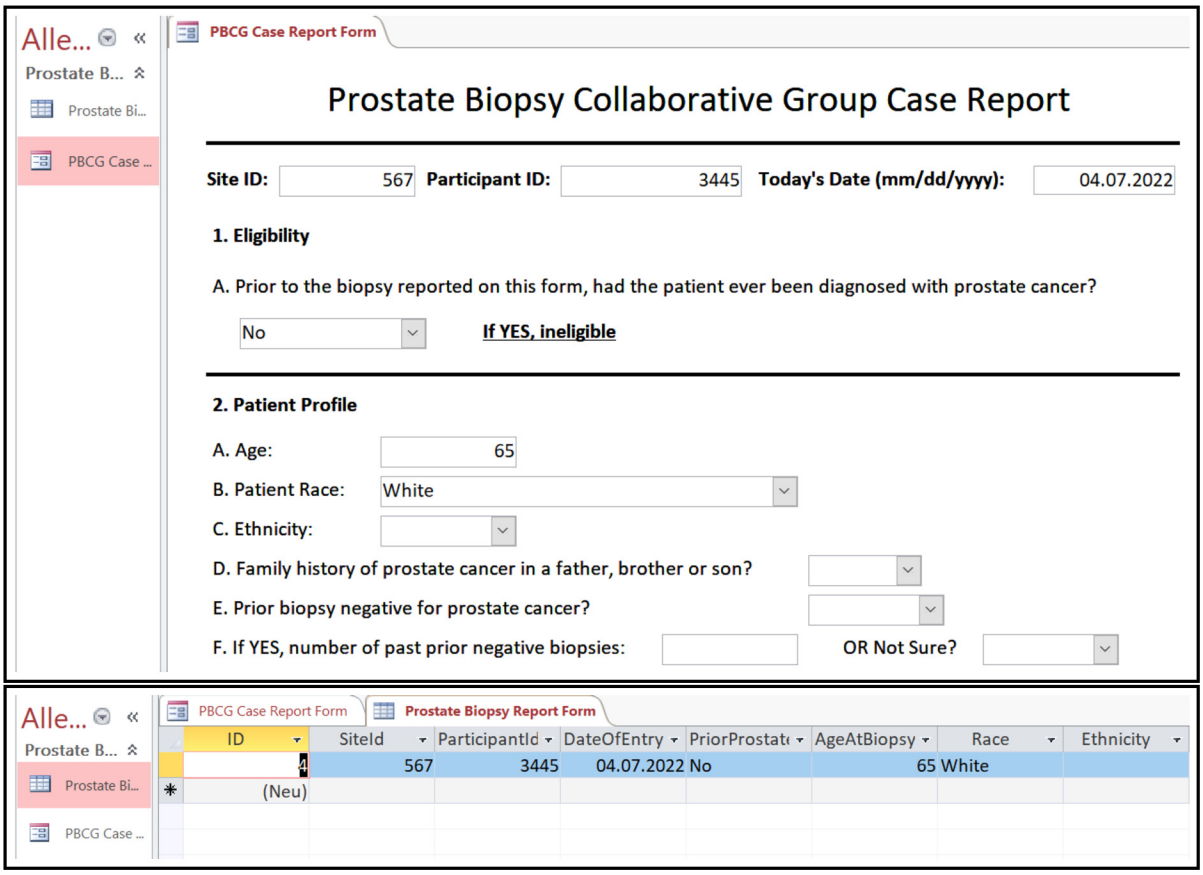

Figure 4: Access data entry form for clinicians with dropdown menus for categorical variables and automated range checks. Entries are automatically transferred to an embedded Excel file, which can also be manually edited.

An additional TRIPOD statement has been developed for the validation of risk tools and is available on the TRIPOD website. Artificial intelligence (AI) has greatly enhanced the capabilities for building online risk tools capitalizing on big data, but such highly-parameterized models are accompanied by an increased risk of overfitting and bias. This has led to current protocols for developing TRIPOD guidelines specific to AI models (Collins et al., 2021).

3 Make Risk Tools Available Online

Once the standardized prospective data have been accrued, quality-checked and corrected, development of a risk prediction tool can proceed. The choice of statistical or machine learning model to use for developing a risk prediction tool should be driven by the specific application at hand, and a good strategy is to research and compare the leading state-of-the-art recommendations. The unique situation of the PBCG was that it comprised outcomes from multiple heterogeneous institutions from all over the world. The initial PBCG risk model used the six established risk factors for prostate cancer. With the small number of predictors, machine learning methods, including random forests, K-nearest neighbors, bagging options, and artificial neural networks,

did not outperform standard logistic regression (Tolksdorf, 2019). Logistic regression has the advantage of interpretable odds ratios and analytical risk prediction formulas for transparent documentation to the public.

After settling on logistic regression, the question remained of how to adjust for the cohort effects. In order to make this modeling decision, five leading suggestions in the literature were identified and compared (Tolksdorf et al., 2019). Three of the methods pooled data from all institutions together, either ignoring the effect of cohort (method 1) or adjusting for it using a normal distributed random effect, the latter using median (method 2) versus mean (method 3) prediction for validation. Pooling or combining data across cohorts requires individual-level data from the cohorts analyzed in a centralized fashion. This was possible for the PBCG because the grant had obtained medical ethics approval from all participating centers allowing de-identified data to be sent across institutional and country borders. The process of executing funding, data curating and statistical analysis for the grant took several years. To circumvent data confidentiality and ethics issues in transporting data, the last two meta-analysis methods considered would analyze data at the local site and only transport estimated regression coefficients and their standard errors for central analysis. Both a fixed effects meta-analysis (method 4) and random-effects meta-analysis (method 5) were considered. All five methods were easily implementable in R using standard packages.

Utilizing data from multiple cohorts allowed an extensive leave-one-cohort–out (LOCO) cross-validation. External validation was also performed on one cohort that had included retrospective cases not following the prospective protocol and hence not included in model development. Validation was assessed according to discrimination, measured by the area underneath the receiver operating characteristic curve (AUC) as well as accuracy according to the negative Hosmer Lemeshow Statistic (HLS), high values of both indicate better validation performance. All five logistic regression methods performed similarly in validation, justifying that any of the methods could be chosen and that predictions were robust to choice of method. Since the current PBCG had expended the effort to collect and centralize the data, method 1 of just pooling the data was chosen. After the grant funding expired data would no longer be able to be transported. Therefore, for future PBCG risk tools, the meta-analyses methods would be preferable.

For the online risk tool, clinicians preferred a further refinement to the logistic regression model that reacted to the evolving prostate cancer clinical landscape by distinguishing low-grade prostate cancer, which may only require active surveillance or watchful waiting based on PSA, family history of prostate cancer and age, versus high-grade prostate cancer that requires more aggressive treatment. Therefore, the PBCG standard risk factor model was created on the pooled PBCG data using multinomial logistic regression for the three outcomes from prostate biopsy: no-, low- and high-grade prostate cancer, based on the six standard risk factors, including those used for monitoring individuals for decision to proceed to biopsy (Ankerst et al., 2018). The model was programmed with an internet interface using the R Shiny application and transported to the Cleveland Clinic Risk Library website at riskcalc.org (Figure 5, R Core Team (2021)). The input page included range checks prohibiting out-of-bound risk factor values. Following consultation with clinicians, patients, and other stakeholders, the output page for risks was designed in easy-to-understand terms as shown in Figure 5.

In addition to immediately assisting patients and clinicians, making the PBCG risk tool available online led to a series of published external validation studies on populations divergent from those on which the tool was developed, thus providing rigorous independent assessment of the generalizability of the risk tool (Engel et al., 2022; Patel et al., 2022; Wei et al., 2021; Yıldızhan et al., 2022; Doan et al., 2021; Amaya-Fragoso and García-Pérez, 2021; Mortezavi

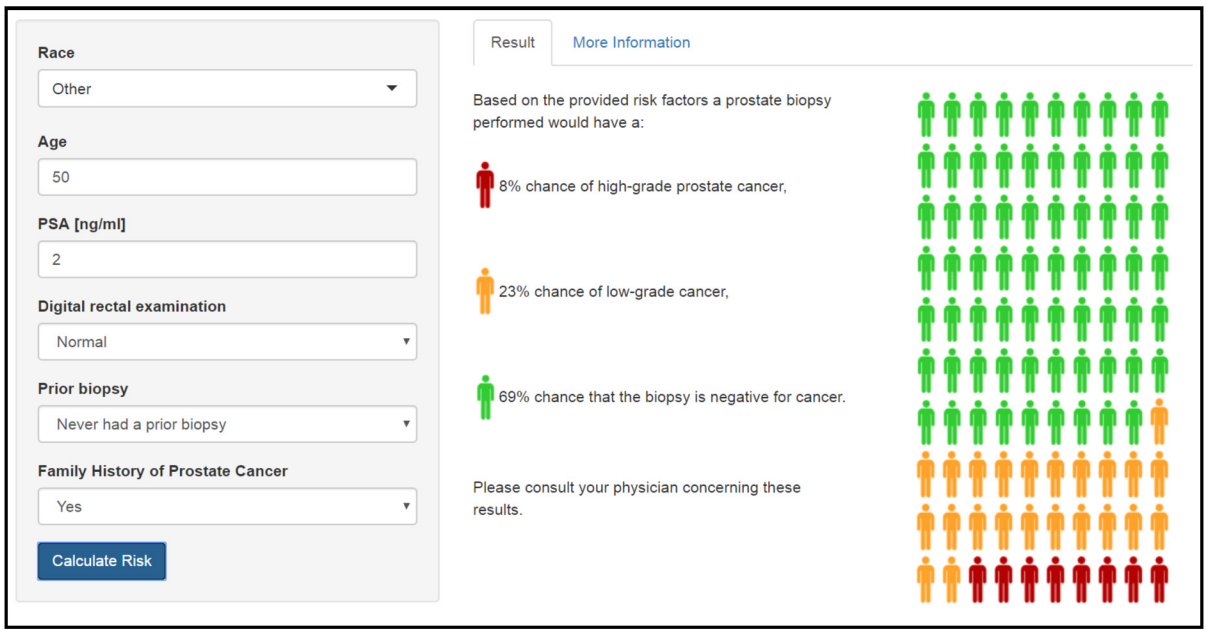

Figure 5: PBCG risk calculator input and output from riskcalc.org/PBCG.

et al., 2021; Carbunaru et al., 2019). Putting the methodology and code in the public domain encouraged other institutions to develop their own risk tools tailored to their populations as competitors to the PBCG risk model (Jalali et al., 2020), as well as to determine whether the PBCG risk tool was still valid for use in patients receiving newer state-of-the-art MRI biopsy techniques (Patel et al., 2022). The advent of the R Shiny package in the ubiquitously used freeware R statistical software system removed the technical difficulties for posting risk tools online, making it feasible for institutions, including those with limited resources, to post their own risk tools. Ji and Kattan (2018) provide a tutorial for this purpose.

A cautionary note about internet risk tools is that there is no referee process: anyone can post a risk tool. Online risk tools should have peer-reviewed references available supporting their validity, ideally with large representative sample sizes and adequate statistical modeling techniques. While posting online facilitates external validation, ideally some validation has occurred and been reported as proof of principle prior to going public. Information should be provided on the website as to the applicable population, and disclaimers provided to warn against unintended use.

4 Dynamically Update Risk Tools

The rapid rise of online risk tools improved clinical practice and accessibility of information to patients. However, clinical practice, medical technologies and patient demographics change over time, which can mean that older clinical risk tools are no longer accurate on contemporary populations. One example of the negative consequences of using an outdated tool was brought to the public fore in 2013, when the American College of Cardiology and American Heart Association released new recommendations for statin use that would be assigned according to risks from a commonly accepted risk tool for cardiovascular events (Cook and Ridker, 2014). Validation on

more contemporary cohorts to those on which the tool was developed showed that cardiovascular risk would be greatly overestimated on current intended populations, and hence lead to over-prescription of statins. Suggested rationale for the over-estimated risks included changes in clinical landscape and population behaviors, such as reduced smoking, increased exercise, and improved diet. The widely publicized high-stakes incident brought to the forefront the need to update clinical risk tools, leading other agencies to question, validate and update their risk-based recommendations in response (Hickey et al., 2013).

Updating a risk tool requires new individual-level data on both outcomes and risk factors. When such data are available there are several modeling options, which are also applicable for tailoring a risk tool to a local population. A previous PBCG study utilized annual data from five international PBCG cohorts spanning 1994 to 2008 and compared six methods for annual cohort-specific updates to a leading risk tool at the time, the Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) (Strobl et al., 2015). Beginning with the second year of data available for each cohort, data from all previous years was used to construct a prostate cancer risk prediction tool that was to be evaluated at the current year as a test set. For the PBCG data, the set of available risk factors was constant over the years, but in case where new predictors become available, some of the methods investigated could be extended. Validation on the current year was assessed according to the AUC and negative HLS as outlined earlier for the other PBCG analyses. Because all cumulative prior data was used for training the annual models, the training data increased in size with each additional year.

The first of the six modeling methods, serving as a baseline, was static use of the online PCPTRC, meaning no cohort-specific data were used. If this approach were to perform at least as well as all others, it would imply the simple low cost result that no update of risk tools is needed for this disease. The second approach was the clean slate method that just fit a new logistic regression model to the training data, thus ignoring the PCPTRC model altogether. The third approach was recalibration of the PCPTRC, one of the simplest methods for adapting a logistic regression risk model to a local population (Strobl et al., 2015). This method requires that the estimated regression coefficients, including the intercept, from the existing risk tool are available, called β_{PCPTRC}, and that the same risk factors X are collected in the training cohort along with their binary outcome of prostate cancer, Y. Then a logistic regression is performed in the training cohort with Y as the outcome and $X'\beta_{PCPTRC}$ as the single covariate, yielding the model $log(P(Y = 1)/P(Y = 0)) = \alpha_0 + \alpha_1 X'\beta_{PCPTRC}$. The intercept α_0 and slope α_1 are estimated from the training data; if α_0 is estimated to be 0 and α_1 to be 1, then the prediction collapses to that based on the static PCPTRC model. The fourth method, referred to as revision, extended the recalibration method by allowing all other covariates to enter the logistic regression individually in addition to the single covariate $X'\beta_{PCPTRC}$. The fifth method followed Bayesian principles and specified a Normal prior distribution for the log odds ratios and their variances from the PCPTRC model. It then computed the posterior distribution of all parameters using the training data to form the likelihood. The sixth method investigated was a machine-learning random forests approach based on the training data, with additional analyses considering $X'\beta_{PCPTRC}$ as a separate covariate as well.

All methods showed no improvement to static use of the PCPTRC in terms of the AUC, but a marked improvement in accuracy as measured by the negative HLS, except for random forests, which had significantly worse accuracy than the static PCPTRC. Recalibration was the easiest to implement and performed equally well to the other updating methods, and hence was recommended in practice (Strobl et al., 2015).

The dynamic analyses described in this section referred to single updates after decades or

annual updates because of the manual data collection, processing and analysis involved. Updating a risk tool through a manual process such as this can only feasibly be performed as necessary, for example, following a major change in medical diagnostics, such as the introduction of prostate MRI-biopsy. The rapid growth of AI methods and massive databanks, which continually and automatically are updated with information, such as through the EHR and daily COVID registries, has led to automated methodologies that can simultaneously receive new data and update risk tools in real time without human involvement. These tools essentially program the manual modeling process. Some applications include kidney transplant survival prognostic models (Raynaud et al., 2021) and real-time monitoring of the adult dengue (Tan et al., 2020).

5 Accommodate Missing Data at Both the Training and End-User Stage

The promise of big data as afforded by EHR, mobile health devices, and large continuous registries yields access to an increasing number of risk factors that should improve the precision of clinical practice. Along with this increase in risk factor availability comes more missing data issues, especially when integrating across different platforms, many of which may not collect the rarer or expensive risk factors. Traditional simple complete case approaches of eliminating entire records with at least one missing risk factor may result in bias. This has paved the way for increased research in methods for accommodating missing data in health research.

When the PBCG sought to develop an extended risk prediction model based on 12 risk factors, some less commonly measured, from the standard six risk factors, it faced multiple types of missing data. In addition to risk factor data missing intermittently within cohorts as in Figure 2, it was also systematically missing by some cohorts that did not measure some risk factors at all. To accommodate intermittent and systematically missing data when pooling data across heterogeneous cohorts, the PBCG identified five approaches using logistic regression (Neumair et al., 2022). For the PBCG, two covariates were available for all individuals and would be required by the end-user, PSA and age, which would automatically be included in the risk model. The remaining ten risk factors would be optional to the user. Figure 6 shows a snapshot of the entry page for the online extended PBCG risk tool at riskcalc.org.

All five methods considered for incorporating missing data were based on logistic regression. The first was termed all available cases. It had previously been proposed in the literature as the 2^k model (Hoogland et al., 2020) and Multiple Models for Missing values at Time Of Prediction (MMTOP) (Ma et al., 2020). The end-user specifies which risk factors they have available, then all the PBCG data with complete cases on that set of risk factors is used to develop a risk model and prediction. Since there were ten optional missing covariates in the PBCG, there were $2^{10} = 1024$ possible models for any single user depending on their available risk factors. Figure 7 shows examples of the variation in risks that can occur for a single user depending on which predictors the user has available. In general, the more risk factors available the more refined and accurate an estimate of risk should be. However, this is counteracted by the reduced training sample size for fitting the model used to develop the prediction, since the number of individuals with a specified set of risk factors decreases as the size of the required set increases.

The second method, referred to as the cohort ensembles method, was similar to meta-analysis, with separate risk models fit to the different cohorts depending on which risk factors the cohort had available. Only cohort-specific models with the same subsets of risk factors specified by the end-user provided predictions that were averaged according to various weighting

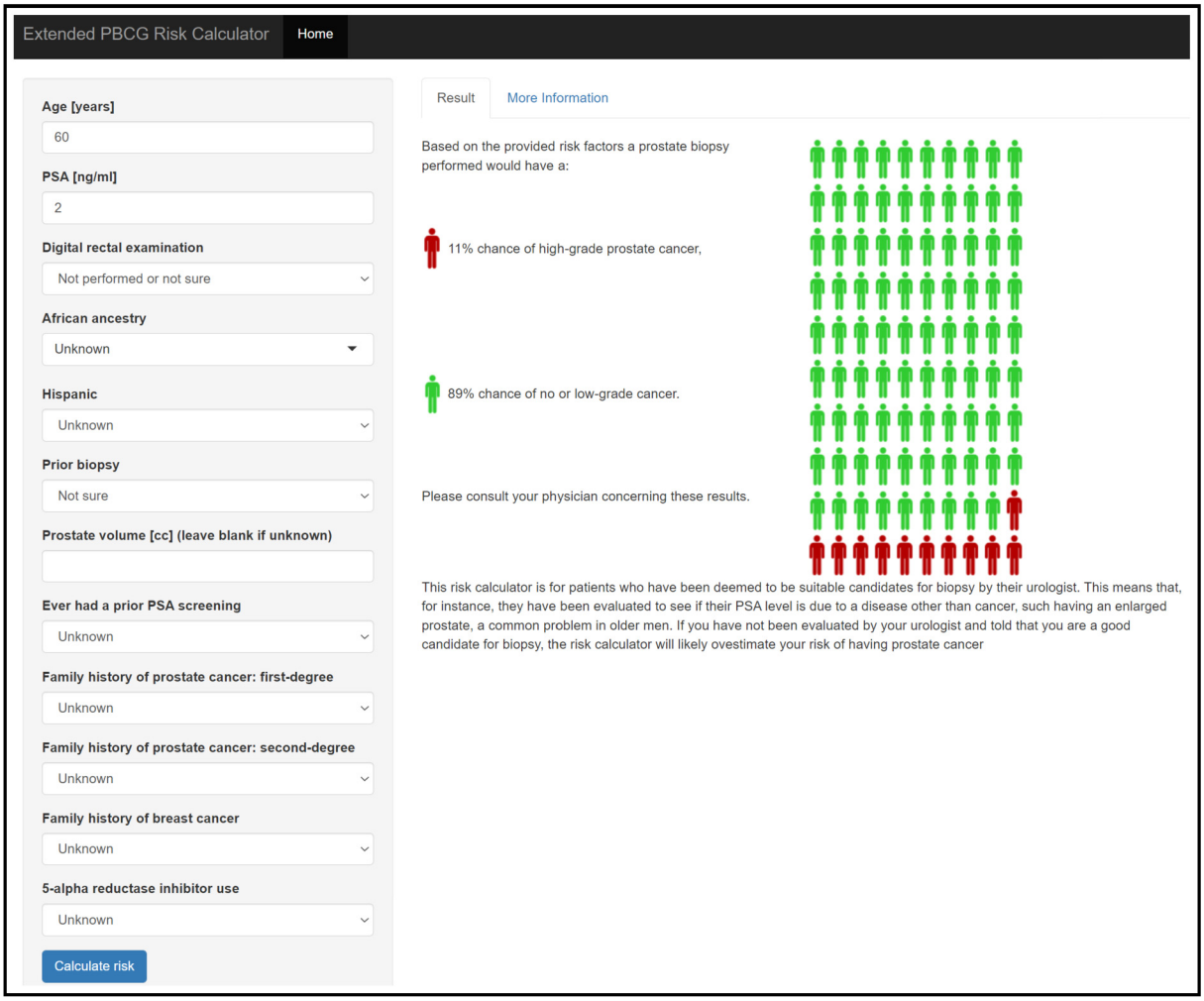

Figure 6: Extended risk factor calculator at riskcalc.org/ExtendedPBCG, with PSA and age required, but all other ten risk factors optional.

schemes. The third method termed categorization and the fourth method termed missing indicators considered missing-ness as a feature of the data. They added a missing factor level to categorical variables or a separate indicator variable for missing versus not for each continuous variable. The categorization method additionally categorized all continuous variables. And finally, the fifth method used multiple imputation. The LOCO and external validation approaches were again used with discrimination and calibration as validation metrics, yielding the available cases method as optimal (Neumair et al., 2022). The available cases method was thus used behind the risk tool interface, with R code containing all 1024 model coefficients included as supplementary information (Neumair et al., 2022).

With big data comes big missing data, and compared to smooth statistical models for small to moderate data, such as logistic regression, more highly parameterized and sensitive machine learning methods may be more susceptible to how the missing data are handled. A recent comprehensive survey reviewed contemporary approaches to using machine learning methods for imputing missing data values, finding that methods such as K-nearest neighbors had reduced

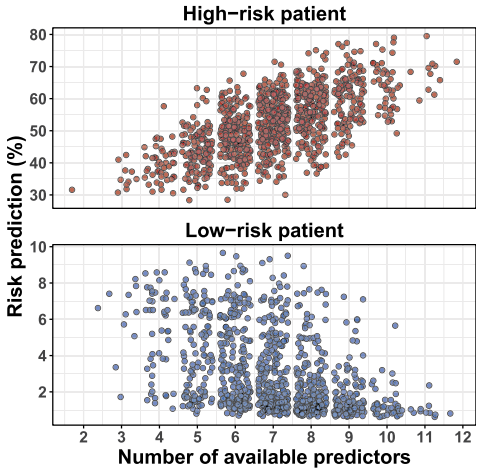

Figure 7: Change in prostate cancer risk according to how many risk factors a patient has available based on the 1024 potential models. The low risk patient is a 60 year-old non-Hispanic, non-African with no family history of cancer of any kind, a PSA of 1 ng/mL, normal DRE, prior negative biopsy, prostate volume of 44 cm^3, prior 5-alpha-reductase inhibitor treatment, and a history of prior PSA screening. The high risk patient is a 75 year-old with African and Hispanic ancestry, a PSA of 4 ng/mL, abnormal DRE, a history of prostate cancer, a prostate volume of 44 cm^3, no 5-alpha-reductase inhibitor treatment, and no prior PSA screening.

root-mean-square-error for recreating the actual missed value compared to random forests (Emmanuel et al., 2021). As the PBCG experience throughout this review has shown, comparison of multiple missing data methods for a given data set using cross- or external-validation provides a robust investigation that the most appropriate method is used for the specific application at hand.

6 Conclusions

Four active data science strategies have been proposed for improving the field of clinical risk prediction. These golden standard proposals are time- and cost-intensive, and hence may not be realistic for often resource-strapped medical informatics and statistical centers. While the strategies may not be implementable in full we offer reasonable alternatives here.

Concerning the first strategy to actively design prospective data collection, the first step is to realize that both clinicians and data scientists spend a disproportionate amount of resource-allocated time cleaning retrospective data collected using systems designed by non-experts that have not been changed or challenged for decades. For example, some data centers interviewed for the PBCG were still collecting the leading biomarker for prostate cancer PSA as part of the optional catch-all notes text field in the patient chart, leading to hours of detective conversations trying to reconstruct and thus salvage data. Some of the effort allocated to data cleaning should be redirected towards smart design of data input, by re-designing the clinician entry forms along the lines proposed by the PBCG. So as not to be overwhelmed, the process should be performed going forward one department or grant-funded project at a time. It should be written in as part of the statistical section of every grant and clinicians should be shown the value of changing entry forms. Clinicians understand the value of preventative medicine for avoiding disease over

treating incurable disease. The analogy holds for data: a simple change of form can prevent thousands of lost or unusable data records, greatly increasing the power of their data.

Concerning the second strategy of making risk tools available online, this does not apply to every risk tool, but rather for large-cohort based tools that have been validated and published. The advent of the R Shiny software, with its early banner of leaving the IT-guy out of it, presented a dramatic shift, bringing internet interfaces to the average researcher adept in R, thus cutting a lot of red tape. The primary usage of R Shiny is for creating internal easy-input interfaces, and not just for prediction tools, but for any analyses of interest for internal data mining by stakeholders. It presents a method for making data visualization and exploration transparent, thus saving valuable time in clinician-statistician meetings. This process is more efficient than back and forth email requests and worth the minor start-up costs.

Concerning the third strategy of dynamically updating risk tools, this need only be performed over intervals where substantial data have been generated. If risk tool development programs are written following sound data collection processes then they can be re-applied efficiently, creating a pipeline of updates. The same philosophy applies as for the other strategies. There is no need to re-invent the wheel for every risk tool, rather invest in foundational approaches that dramatically shave off time for future reproductions.

Finally, for the fourth strategy of accommodating missing data at both the training and end-user stage, as the TRIPOD statement indicates, it is becoming increasingly impossible to avoid addressing missing data for grants or publishable research and by now there are numerous R packages and tutorials available for implementing the methods. When developing a clinical risk prediction tool, keep in mind the end-user, as the impact of the risk tool is measured most by the help it can give patients and clinical care-takers. Making a usable and user-friendly tool requires minimal additional effort. Provide options for missing risk factors as lack of a single risk factor may turn away the end-user from the tool. Provide information on the risk factors to better help the end-user with questions concerning the input and provide range checks to automatically inform the end-user of the mistake. In addition to the primary goal of helping clinical practice, provide underlying formulas or R code to facilitate peer-reviewed external validation of the risk tool by independent centers. This ultimately provides the population with the necessary confidence to use the risk tool in practice, or develop their own tailored tools to maximally benefit their patient population.

The modeling strategies were illustrated using logistic regression for binary outcomes as utilized in the PBCG, but extend directly to commonly used approaches employing risk models for time to events or development of outcomes within specified time-periods, such as the five-year risk of developing prostate cancer (Pfeiffer et al., 2022). Additional missing data issues that arise in these models, including informative censoring and competing events, have been well-researched leading to practical solutions (Coemans et al., 2022). More research is needed for data science strategies dealing with high-dimensional predictors, such as genomic and time-series clinical measurements, as risk models built on increasingly precise comprehensive data have the most potential to improve clinical practice.

Supplementary Material

R code for producing figures is provided along with the TRIPOD checklist for prediction model development.

Funding

Funding for the PBCG was provided by the US National Institutes of Health R01 grant CA179115.

References

Amaya-Fragoso E, García-Pérez CM (2021). Improving prostate biopsy decision making in Mexican patients: still a major public health concern. *Urologic Oncology*, 39(12): 831.e11–831.e18.

Ankerst DP, Straubinger J, Selig K, Guerrios L, de Hoedt A, Hernandez J, et al. (2018). A contemporary prostate biopsy risk calculator based on multiple heterogeneous cohorts. *European Urology*, 74(2): 197–203.

Austin EJ, Lee JR, Ko CW, Kilgore MR, Parker EU, Bergstedt B, et al. (2020). Improving the impact of clinical documentation through patient-driven co-design: experiences with cancer pathology reports. *Healthcare Informatics*, 27(3). https://doi.org/10.1136/bmjhci-2020-100197.

Carbunaru S, Nettey OS, Gogana P, Helenowski IB, Jovanovic B, Ruden M, et al. (2019). A comparative effectiveness analysis of the PBCG vs. PCPT risks calculators in a multi-ethnic cohort. *BMC Urology*, 19(1): 121.

Coemans M, Verbeke G, Döhler B, Süsal C, Naesens M (2022). Bias by censoring for competing events in survival analysis. *BMJ Clinical Research*, 378: e071349.

Collins GS, Dhiman P, Navarro CL, Ma J, Hooft L, Reitsma JB, et al. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*, 11(7): e048008.

Collins GS, Reitsma JB, Altman DG, Moons KGM (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BJS*, 102(3): 148–158.

Cook NR, Ridker PM (2014). Further insight into the cardiovascular risk calculator: the roles of statins, revascularizations, and underascertainment in the Women's Health Study. *JAMA Internal Medicine*, 174(12): 1964–1971.

Doan P, Graham P, Lahoud J, Remmers S, Roobol MJ, Kim L, et al. (2021). A comparison of prostate cancer prediction models in men undergoing both magnetic resonance imaging and transperineal biopsy: are the models still relevant? *BJU International*, 128(Suppl 3): 36–44.

Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O (2021). A survey on missing data in machine learning. *Big Data*, 8(1): 140.

Engel JC, Palsdottir T, Ankerst D, Remmers S, Mortezavi A, Chellappa V, et al. (2022). External validation of the prostate biopsy collaborative group risk calculator and the rotterdam prostate cancer risk calculator in a Swedish population-based screening cohort. *European Urology Open Science*, 41: 1–7.

Grinspan ZM, Patel AD, Shellhaas RA, Berg AT, Axeen ET, Bolton J, et al. (2021). Design and implementation of electronic health record common data elements for pediatric epilepsy: foundations for a learning health care system. *Epilepsia*, 62(1): 198–216.

Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, et al. (2013). Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. *European Journal of Cardio-Thoracic Surgery*, 43(6): 1146–1152.

Hoogland J, van Barreveld M, Debray TPA, Reitsma JB, Verstraelen TE, Dijkgraaf MGW, et al. (2020). Handling missing predictor values when validating and applying a prediction model to new patients. *Statistics in Medicine*, 39(25): 3591–3607.

Jalali A, Foley RW, Maweni RM, Murphy K, Lundon DJ, Lynch T, et al. (2020). A risk calculator to inform the need for a prostate biopsy: a rapid access clinic cohort. *BMC Medical Informatics and Decision Making*, 20(1): 148.

Ji X, Kattan MW (2018). Tutorial: development of an online risk calculator platform. *Annals of Translational Medicine*, 6(3): 46.

Ma S, Schreiner PJ, Seaquist ER, Ugurbil M, Zmora R, Chow LS (2020). Multiple predictively equivalent risk models for handling missing data at time of prediction: with an application in severe hypoglycemia risk prediction for type 2 diabetes. *Journal of Biomedical Informatics*, 103: 103379.

Mortezavi A, Palsdottir T, Eklund M, Chellappa V, Murugan SK, Saba K, et al. (2021). Head-to-head comparison of conventional, and image- and biomarker-based prostate cancer risk calculators. *European Urology Focus*, 7(3): 546–553.

Neumair M, Kattan MW, Freedland SJ, Haese A, Guerrios-Rivera L, de Hoedt AM, et al. (2022). Accommodating heterogeneous missing data patterns for prostate cancer risk prediction. *BMC Medical Research Methodology*, 22(1): 200.

Patel AA, Gilbertson JR, Parwani AV, Dhir R, Datta MW, Gupta R, et al. (2006). An informatics model for tissue banks–lessons learned from the cooperative prostate cancer tissue resource. *BMC Cancer*, 6: 120.

Patel HD, Koehne EL, Shea SM, Fang AM, Gerena M, Gorbonos A, et al.(2022). A prostate biopsy risk calculator based on MRI: development and comparison of PLUM to the PBCG. *BJU International*. https://doi.org/10.1186/1471-2407-6-120

Pfeiffer RM, Chen Y, Gail MH, Ankerst DP (2022). Accommodating population differences when validating risk prediction models. *Statistics in Medicine*. https://doi.org/10.1002/sim.9447.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Raynaud M, Aubert O, Divard G, Reese PP, Kamar N, Yoo D, et al. (2021). Dynamic prediction of renal survival among deeply phenotyped kidney transplant recipients using artificial intelligence: an observational, international, multicohort study. *The Lancet Digital Health*, 3(12): e795–e805.

Roobol MJ, Schröder FH, Hugosson J, Jones JS, Kattan MW, Klein EA, et al. (2012). Importance of prostate volume in the European Randomised Study of Screening for Prostate Cancer (ERSPC) risk calculators: results from the prostate biopsy collaborative group. *The World Journal of Urology*, 30(2): 149–155.

Strobl AN, Vickers AJ, van Calster B, Steyerberg E, Leach RJ, Thompson IM, et al. (2015). Improving patient prostate cancer risk assessment: moving from static, globally-applied to dynamic, practice-specific risk calculators. *The Journal of Biomedical Informatics*, 56: 87–93.

Tan KW, Tan B, Thein TL, Leo YS, Lye DC, Dickens BL, et al. (2020). Dynamic Dengue haemorrhagic fever calculators as clinical decision support tools in adult Dengue. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 114(1): 7–15.

Thompson IM, Ankerst DP, Chi C, Goodman PJ, Tangen CM, Lucia MS, et al. (2006). Assessing prostate cancer risk: results from the prostate cancer prevention trial. *Journal of the National Cancer Institute*, 98(8): 529–534.

Tolksdorf J, Kattan MW, Boorjian SA, Freedland SJ, Saba K, Poyet C, et al. (2019). Multi-

cohort modeling strategies for scalable globally accessible prostate cancer risk tools. *BMC Medical Research Methodology*, 19(1): 191.

Tolksdorf JE (2019). Data scientific approaches to contemporary clinical risk tool construction. Universitätsbibliothek der TU München, München.

Vickers AJ, Cronin AM, Roobol MJ, Hugosson J, Jones JS, Kattan MW, et al. (2010). The relationship between prostate-specific antigen and prostate cancer risk: the prostate biopsy collaborative group. *Clinical Cancer Research*, 16(17): 4374–4381.

Wei G, Kelly BD, Timm B, Perera M, Lundon DJ, Jack G, et al. (2021). Clash of the calculators: external validation of prostate cancer risk calculators in men undergoing mpMRI and transperineal biopsy. *BJUI Compass*, 2(3): 194–201.

Westra BL, Lytle KS, Whittenburg L, Adams M, Ali S, Furukawa M, et al. (2020). A refined methodology for validation of information models derived from flowsheet data and applied to a genitourinary case. *Journal of the American Medical Informatics Association*, 27(11): 1732–1740.

Yıldızhan M, Balcı M, Eroğlu U, Asil E, Coser S, Özercan AY, et al. (2022). An analysis of three different prostate cancer risk calculators applied prior to prostate biopsy: a Turkish cohort validation study. *Andrologia*, 54(2): e14329.