Journal of Clinical Epidemiology 133 (2021) 53−60

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Sample sizes of prediction model studies in prostate cancer were rarely justified and often insufficient

Shane D. Collins[a,b], Niels Peek[b], Richard D. Riley[c], Glen P. Martin[b,*]

[a]*Research Department of Oncology, Cancer Institute, Faculty of Medical Sciences, School of Life & Medical Sciences, University College London, London, UK*

[b]*Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK*

[c]*Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Staffordshire, UK*

Accepted 8 December 2020; Published online 28 December 2020

Abstract

Objective: Developing clinical prediction models (CPMs) on data of sufficient sample size is critical to help minimize overfitting. Using prostate cancer as a clinical exemplar, we aimed to investigate to what extent existing CPMs adhere to recent formal sample size criteria, or historic rules of thumb of events per predictor parameter (EPP) $\geq$ 10.

Study Design and Setting: A systematic review to identify CPMs related to prostate cancer, which provided enough information to calculate minimum sample size. We compared the reported sample size of each CPM against the traditional 10 EPP rule of thumb and formal sample size criteria.

Results: About 211 CPMs were included. Three of the studies justified the sample size used, mostly using EPP rules of thumb. Overall, 69% of the CPMs were derived on sample sizes that surpassed the traditional EPP $\geq$ 10 rule of thumb, but only 48% surpassed recent formal sample size criteria. For most CPMs, the required sample size based on formal criteria was higher than the sample sizes to surpass 10 EPP.

Conclusion: Few of the CPMs included in this study justified their sample size, with most justifications being based on EPP. This study shows that, in real-world data sets, adhering to the classic EPP rules of thumb is insufficient to adhere to recent formal sample size criteria.

Keywords: Prediction models; Prostate cancer; Sample size; Development; Validation

1. Introduction

Clinical prediction models (CPMs) are statistical models or algorithms that can estimate the risk of existing disease (diagnostic) or the probability of future outcomes (prognostic) for an individual [1,2]. Estimations of risk are conditional on the values of multiple predictors that are observable at the time one wishes to make a prediction from the model. Classically, these models are based on multivariate modeling methods such as logistic regression for binary outcomes or survival models for time-to-event outcomes.

Although there is a plethora of CPMs developed across medical domains, very few are implemented clinically, despite their many practical uses [1,3,4]. Commonly, this lack of uptake is attributed to reduced predictive performance when CPMs are validated in independent cohorts (e.g., external validation) [5−7]. Indeed, CPMs can suffer from lack of generalizability, meaning that repeated de novo development of CPMs is not uncommon [8]. The ways in which CPMs are developed could also contribute to the lack of uptake in practice.

Specifically, it is now recognized that the sample size used to develop a CPM is crucial to help achieve robust predictive performance [9,10]. Small sample sizes may result in extreme estimates of predictor effects (i.e., overfitting), subsequently resulting in poor predictive performance when applied to new patients. Although penalization and shrinkage methods (such as LASSO or ridge regression) are available to help with overfitting, these are not a solution to small sample size [11−13].

Historically, studies that develop CPMs have often justified their sample size based on events per predictor parameter (EPP)—the ratio of the number of outcome events,

Funding: None.
Conflict of Interests: None.
* Corresponding author. Tel.: +44 (0) 161 27 50179.
E-mail address: glen.martin@manchester.ac.uk (G.P. Martin).

https://doi.org/10.1016/j.jclinepi.2020.12.011

What is new?

Key findings

- This systematic review highlighted over 200 prediction model published related to prostate cancer, with very few of the included studies justifying their choice of sample size; any justification were usually being based on Events per Predictor Parameter (EPP).

What this adds to what was known?

- The classic use of the EPP $\geq$ 10 rule of thumb is not necessarily enough to guide sample size for prediction model development based on formal criteria.
- Our study highlights the extent to which this situation has previously been an issue, and so serves as a benchmark for comparison in future reviews of studies in the coming years

What is the implication and what should be done now?

- There is large scope to improve the justification of sample sizes used in prediction model studies and single threshold values for EPP are insufficient to do this accurately.
- Regardless of whether a sample size calculation has been used, our recommendation is that the justification for sample size consideration should be included in all prediction model studies going forward.

relative to the number of candidate predictor parameters—with an EPP of $\geq$ 10 often taken as a rule of thumb [14–16]. However, this blanket rule of thumb is too simplistic [10,17,18]. As such, Riley et al. [11,19,20] recently published a series of sample size formula to calculate the minimum required sample size for binary, time to event and continuous prediction models. Hereto, these sample size criteria will be referred to as "Riley et al." with the references being as follows: [11,19,20]. These criteria help to ensure that the prediction model is robustly developed [11]. Indeed, compared with previous guidance around sample size requirements for prediction models, the Riley et al. criteria are tailored to the model (and clinical context) in question. For example, they are context-specific in terms of outcome incidence and model fit. As such, in this study, we take the Riley et al. criteria as the gold standard for sample size calculation.

However, it is unclear to what extent previously developed CPMs adhere to minimum required sample sizes as calculated by the Riley et al. criteria. Therefore, the aim of this study was to use prostate cancer as a clinical exemplar in which to retrospectively assess whether published multivariable CPMs adhere to the minimum sample size criteria as outlined by Riley et al. [19,20] and the level of agreement between these criteria and the EPP $\geq$ 10 rule of thumb. We chose to focus on CPMs within prostate cancer because this is a common context in which CPMs are developed in both a diagnostic and prognostic prediction setting. This is largely due to their many practical uses, including predicting disease onset [21]; risk stratification [22]; predicting risk of upgrading during active surveillance [23]; predicting risk of recurrence [24]; predicting risk of treatment toxicity [25]; and predicting survival [26].

2. Methods

We conducted and reported this systematic review as per the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [27].

2.1. Search strategy for the identification of studies

We undertook a literature search in Pubmed and Embase using a published search filter specific to finding studies related to CPMs [28]; this existing search filter was combined with terms specific to prostate cancer (see supplementary methods). We also searched the reference list of any relevant systematic reviews that we discovered in our database search to identify additional CPM development studies for inclusion. The last search was conducted on the June 30, 2019.

2.2. Eligibility criteria

We included any papers that developed a multivariable model/score/tool/algorithm (hereto termed model) for predicting the individual risk of an outcome of interest in the context of prostate cancer. Because we were interested in models that output the risk of the outcome of interest, we only included prediction models that were based on logistic or cox regression for binary and time-to-event outcomes, respectively [19]. To be included, the papers must have reported sufficient information to allow us to retrospectively calculate the Riley et al. minimum required sample size; any study that did not report sufficient information was excluded. We also excluded any papers that externally validated an existing model (without developing a new model) and those that aimed to update an existing risk model with a new predictor, although such papers were used to identify the paper that developed the original CPM. Furthermore, papers that developed CPMs for multiple anatomical sites were excluded, as were those that were developed within a competing risks/multistate modeling framework (because multiple outcome regression is currently not covered by the Riley et al. [19,20] criteria). Finally, we excluded any papers that were only available as an abstract (e.g., conference

proceedings). We limited to papers published in English or those available with an English translation. The study selection process was documented using the PRISMA flow diagram [29], including reasons for exclusion.

2.3. Screening process

Initially, the titles and abstracts of identified papers were screened by the lead author (S.D.C.), cross-referencing against the inclusion and exclusion criteria. Papers satisfying the inclusion and exclusion criteria at this stage were then full-text-screened (which further excluded papers as required). Any uncertainty in whether to include/exclude a particular study was resolved through discussion and consensus with a second reviewer (G.P.M.).

2.4. Data extraction

Primary information that we extracted from identified papers included the following: the sample size used (for model development), the number of outcome events, the predictor parameters considered, and the reported C-statistic (to retrospectively calculate R^2 as outlined by Riley et al. [11,19,20]). In addition, for time-to-event models, we extracted the mean follow-up and length of follow-up reported in the papers. Extraction of all these values enabled us to retrospectively calculate the Riley et al. minimum sample size criteria [19,20]. In addition, the EPP was retrospectively calculated using the reported number of events and number of predictors. For calculation of the Riley et al. criteria and EPP, we considered the total degrees of freedom for all variables (of all candidate predictors), where this was possible to determine; if the number of candidate predictors was not reported, then the number of parameters in the final model was used for the calculations. Finally, we noted if the study was reported in accordance with the TRIPOD guidelines (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) [30], and whether the study conducted any form of internal and/or external validation alongside the model development.

2.5. Statistical analysis

Results were summarized using descriptive statistics. We used logistic regression to examine if the odds of a paper surpassing the EPP $\geq$ 10 rule of thumb or the Riley et al. sample size criteria changed with time. We used the chi-squared test to examine if the number of papers surpassing the EPP $\geq$ 10 rule of thumb and/or Riley et al. sample size criteria differed by study type (development/internal/external validation) or clinical task (i.e., intended prediction aim). All analyses were performed using R version 3.6.2 [31], with the package ‘‘pmsampsize’’ [32] used to calculate the Riley et al. required sample size.

3. Results

The initial search identified 5,026 papers, with an additional 20 identified through citation searching of systematic reviews (Supplementary Table 1); of these, 2,628 papers remained following removal of duplicates. After the initial title/abstract screening, 305 papers underwent full-text screening, which identified 139 papers for inclusion. These papers resulted in 211 CPMs (because some papers developed more than one CPM). Fig. 1 shows the PRISMA flow diagram, and Supplementary Table 2 gives the full list of included papers.

3.1. Study characteristics

The included papers were published between 1994 and 2019. The intended use of the CPM and the modeling methods varied across the included studies (Table 1). Of the 211 models included in this review, 124 (59%) focused on diagnosis of prostate cancer, 9 (4%) on predicting side effects, 44 (21%) on risk of progression/recurrence, and 34 (16%) on survival/mortality predictions. Overall, 143 CPMs were developed using logistic regression for binary outcomes, and 68 using a Cox proportional hazards model for time-to-event outcomes.

As shown in Table 1, 43 (20%) of the models detailed development only, 116 (55%) also incorporated internal validation (i.e., adjusted performance for in-sample optimism) and 52 (25%) included development and external validation (i.e., validation in an independent data set).

3.2. Adherence to sample size requirements: overall

From the 139 included studies, 34 (24%) acknowledged limitations of the sample size used to develop their proposed model(s), but only 3 studies outlined how they calculated their minimum required sample size. Of these three studies, the first used EPP $\geq$ 10 [33], the second used EPP $\geq$ 20 [34], and the third based their sample size calculation on achieving ‘‘92% power [for] a noninferior sensitivity and superior specificity’’ to a previously developed model [35].

Supplementary Table 2 depicts which of the included papers satisfied the traditional EPP $\geq$ 10 rule of thumb and which satisfied the Riley et al. sample size criteria. Overall, 102 of the included CPMs (48%) were developed on sample sizes that surpassed the Riley et al. criteria, and 145 (69%) of the included models exceeded the traditional 10 EPP rule of thumb. Less than half of the models (47%) satisfied both EPP $\geq$ 10 and the Riley et al. criteria (Fig. 2), and 64 models failed to meet both criteria.

Across the CPMs that satisfied the Riley et al. sample size criteria, there was large variability in their EPP (Supplementary Fig. 1). In most CPMs, the Riley et al. criteria resulted in a higher required sample size than that based on EPP $\geq$ 10 (Fig. 3). About 38 (18%) of the

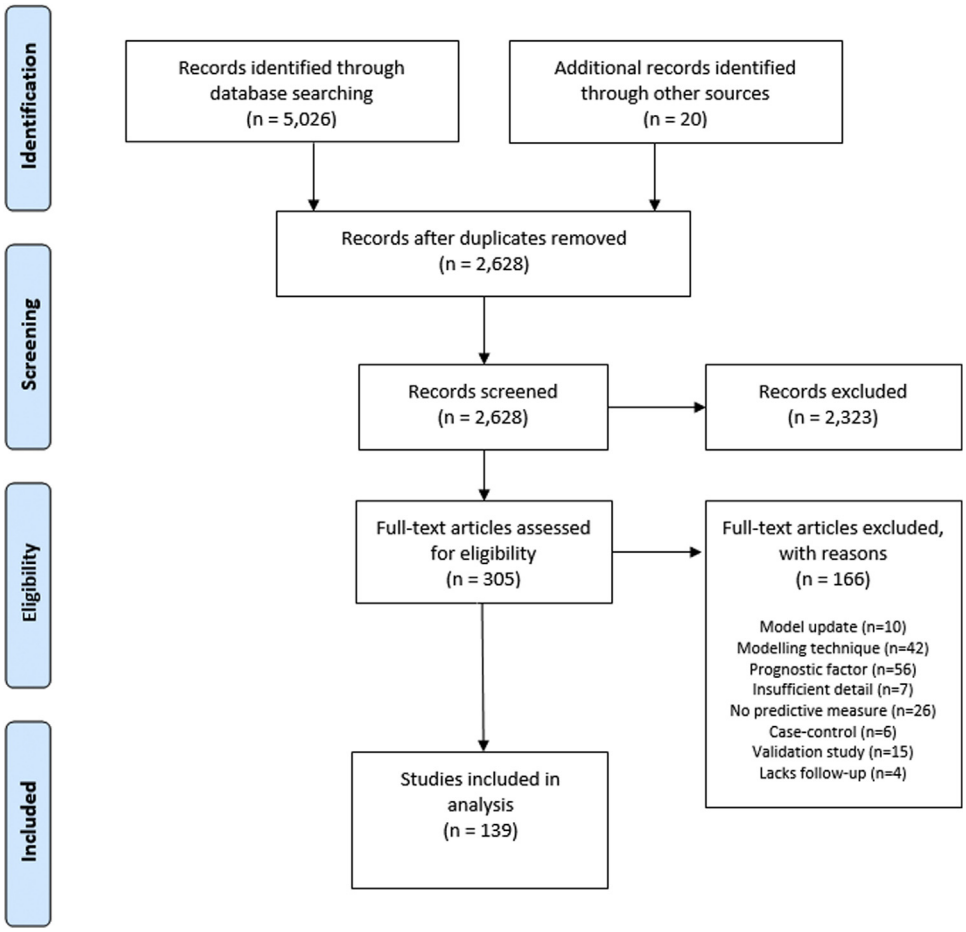

Fig. 1. Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) flow diagram.

included CPMs required a lower minimum sample size to satisfy the Riley et al. criteria compared with the sample size that would be required to surpass EPP $\geq$ 10. Here, the calculated sample size from the Riley et al. formula was driven by criteria 1 (small optimism in predictor effect estimates) in 141 CPMs, criteria 2 (small difference in apparent and adjusted model fit) in 5 CPMs and criteria 3 (precise estimation of overall risk) in 65 CPMs [11,19,20].

There was no evidence that the proportion of papers surpassing the Riley et al. sample size criteria changed through time ($P = 0.323$, Fig. 4). We found that study type (development only/+internal/+external validation) was significantly associated with pass rate ($P < 0.001$). Clinical task was also associated with pass rate ($P = 0.005$), suggesting differing pass rates between diagnostic models, risk of progression/recurrence models, side effects models, and survival models.

3.3. Adherence to sample size requirements: binary outcomes

Most of the included CPMs were developed for binary outcomes (143 models), of which, 71% satisfied the $\geq$ 10 EPP rule of thumb and 51% satisfied the Riley et al. sample size criteria (Supplementary Table 2). Only 24 (17%) models had a lower required minimum sample size to meet the Riley et al. criteria compared with the sample size that would be needed to meet the $\geq$ 10 EPP rule of thumb.

3.4. Adherence to sample size requirements: time-to-event outcomes

Of the 68 included CPMs that were developed for time-to-event outcomes, 65% satisfied the $\geq$ 10 EPP rule of thumb, and 43% satisfied the Riley et al. criteria (Supplementary Table 2). Only 14 (21%) of the time-to-event CPMs had a lower required sample size according to the Riley et al. criteria compared with the sample size that would be required to surpass the $\geq$ 10 EPP rule of thumb, again showing that the Riley et al. criteria are more strict.

3.5. Quality assessment

Of all included studies, 46 were published after the publication of the TRIPOD guidelines [30], with only 4 (9%)

Table 1. Distribution of modeling techniques across prediction aim

Prediction aim	Binary	Time to event	Total
All			
Development	26	17	43
+ Internal validation	85	31	116
+ External validation	32	20	52
Total	143	68	211
Diagnosis			
Development	20	1	21
+ Internal validation	74	1	75
+ External validation	27	1	28
Total	121	3	124
Side effects			
Development	3	-	3
+ Internal validation	2	1	3
+ External validation	3	-	3
Total	8	1	9
Progression/recurrence			
Development	3	8	11
+ Internal validation	7	16	23
+ External validation	2	8	10
Total	12	32	44
Survival/mortality			
Development	-	8	8
+ Internal validation	2	13	15
+ External validation	-	11	11
Total	2	32	34

Note that 139 studies met the inclusion criteria, including 211 models.

Internal validation was defined as any appropriate method that was used to adjust predictive performance for in-sample optimism, such as bootstrap resampling. External validation was defined as any paper that included an independent data set from a distinct population, such as geographical validation.

of these explicitly adhering to them [36–39]. These 4 studies produced 6 CPMs, of which only 1 satisfied both the traditional rule of thumb and Riley et al. criteria [37], and all six were based on cox proportional hazards models.

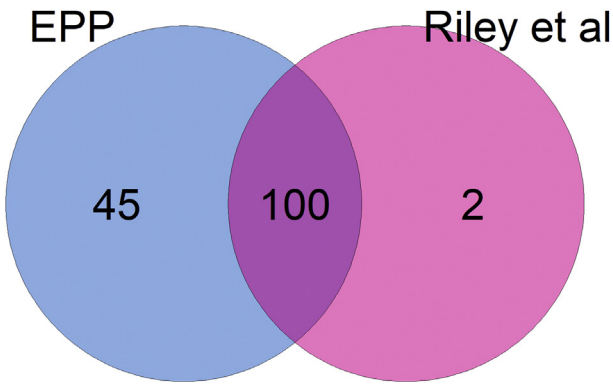

Fig. 2. A Venn diagram of the included models which surpass and fail to meet events per predictor parameter (EPP) ≥ 10 and the Riley et al. criteria. A total of 64 models failed to meet both criteria.

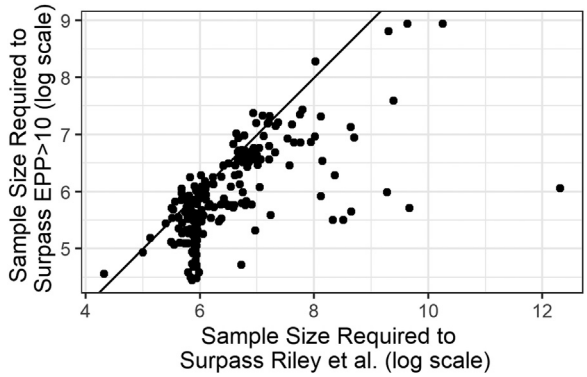

Fig. 3. A scatter plot of the sample size that would be needed to satisfy the events per predictor parameter (EPP) ≥ 10 rule of thumb against the required sample size based on the Riley et al. criteria. Both axes are on the log-scale to aid visual appearance of the plot.

4. Discussion

This systematic review adds important implications to the literature. First, it shows that sample sizes are rarely justified in this field. Second, it highlights that the classic use of EPP≥10 is not necessarily sufficient to guide sample size for prediction model development based on formal criteria. Thirdly, regardless of whether a sample size calculation has been used, our recommendation is that a justification for sample size should be included in all prediction model studies going forward (as a minimum). Finally, our study highlights the extent to which this situation has previously been an issue, and so serves as a benchmark for comparison in future reviews of studies in the coming years (that should aim to examine if improvements have been made). To be clear, the intention of this paper is not to point blame at previous studies in terms of sample size and justification of sample size, but rather highlight to readers that this topic is an important and outstanding issue in the reporting of prediction model studies.

This study found over 200 published CPMs for risk prediction in prostate cancer, but only three justified the sample size used, most of which were based on EPP rules of thumb. We were not able to determine the precise reasons why studies might not justify sample size. One potential reason is that some studies included in this review were published before the TRIPOD guidelines (which includes an item to report how the sample size was arrived at). In addition, we postulate that this also reflects the historic lack of formal guidelines around required sample size in CPM development studies, and the historic blanket use of EPP rules of thumb [14–16]. Indeed, several systematic reviews have observed that prediction model studies—both development and validation—frequently provide no rationale for the sample size used, or discuss the potential for overfitting [3,40,41]. The recently published Riley et al. criteria [19,20] provide the required mechanism to allow sample size justification. It is also possible that if researchers have access to a data set of fixed sample size, then sample size

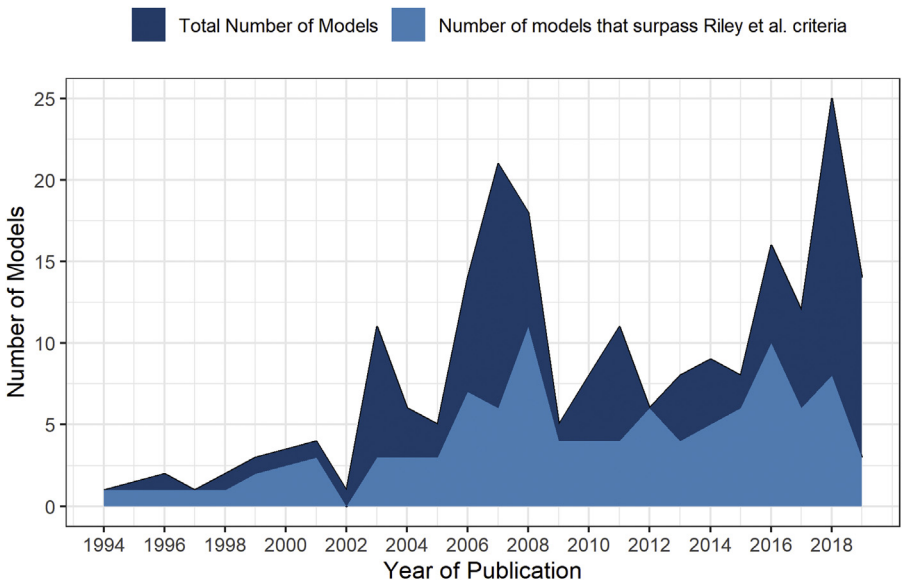

Fig. 4. The number of clinical prediction models (CPMs) related to prostate cancer that have been published, and the number of those that satisfy the Riley et al. sample size criteria.

justification might not be given. Here, the Riley et al. criteria should be used to determine the maximum number of candidate predictor variables for the fixed sample size.

We acknowledge that all of the CPMs considered in this review were published before the Riley et al. sample size criteria [19,20] (which were published in 2019); consequently, one cannot expect historically derived CPMs to adhere to these criteria (and, again, we do not intend to point blame). Nonetheless, our findings highlight the importance of carefully justifying the required sample size in all future CPM development studies. The TRIPOD guidelines include the requirement to report how the study sample size was obtained [30]; our findings suggest that these guidelines should be extended to include a requirement for papers to formally justify the sample size (e.g., using the Riley et al. criteria). To use the Riley et al. criteria in practice, one needs to specify the anticipated model fit (i.e., R^2) in advance of data collection/model fitting [11,19,20]. In this study, we based the sample size calculation on the performance of the fitted model; thus, it could be that—in the future—modelers adhere to the Riley et al. guidance but are too optimistic in the value of R^2 that they choose. In such a case, it would retrospectively appear that the study failed to meet the criteria. Therefore, we suggest that all steps of the Riley et al. calculation should be reported in development papers.

Importantly, we found that only 48% of included CPMs surpassed the Riley et al. sample size criteria [19,20], whereas 68% had an EPP $\geq$ 10. It is now widely accepted that an EPP $\geq$10 is overly simplistic [42–45]. Indeed, this study found that there was large spread in the EPP for studies that satisfied the Riley et al. criteria (Supplementary Fig. 1), demonstrating that a single threshold value for EPP is insufficient; this supports existing research in this area [10,17,18]. In addition, in most cases, the Riley et al. sample size formula [11,19,20] resulted in higher sample sizes than the sample sizes that would be needed to meet an EPP $\geq$ 10, with the former usually being based on minimizing overfitting (i.e., criteria 1 of Riley et al. [11,19,20]). In other words, if a particular study met the 10 EPP rule of thumb, then this does not guarantee that it is likely to meet the Riley criteria, which is often more stringent. This finding reinforces that EPP is not suitable to guide sample size for CPM development.

In addition, we found that studies including internal/external validation alongside model development had higher odds of surpassing the Riley et al. criteria, compared with studies that only included CPM derivation. This finding suggests that sample size improvements are linked to having improved methodology in general within CPM studies. Furthermore, it is currently unknown if adhering to formal sample size criteria leads to greater generalizability of predictive performance [1,3,4]. Lack of generalizability of CPMs might lead to individual institutes/centers/research groups developing de novo models on local data, and this could further compound the issues with low sample size. Indeed, lack of sufficient sample size in local settings increases the risk of overfitting. Further research is required to explore both of these points in more detail.

Similarly, further research is needed around sample size requirements for modeling methods such as ordinal, multinomial, competing risks/multistate models and machine learning methods. Indeed, while the sample size criteria outlined by Riley et al. [19,20] provides guidance for logistic, cox, and linear regression models, this study identified other modeling techniques being used to develop CPMs in prostate cancer, which we excluded due to the lack of

sample size formula. For example, owing to the rise in popularity of machine learning techniques, guidance around sample size has become more pertinent [46], especially because they are often ‘‘data hungry’’ [47].

Several limitations should be considered when interpreting the results of this study. First, the search was completed until June 2019. Prompted by a reviewer comment, we examined a random sample of 10 articles published between June 2019 and November 2020 (time of the first reviewer comments) and we found very similar conclusions; for example, 53% surpassed the Riley sample size criteria, which is similar to the 48% in the main paper. Second, this analysis was retrospective, meaning that most of the included studies were published before the publication of the Riley et al. sample size criteria. Future work should explore if adherence increases in the future. Third, 37 studies were excluded due to insufficient information being reported to calculate the Riley et al. sample size criteria. Such exclusion potentially introduces bias into our findings if such papers differ from those included in terms of their reported sample size relative to the required sample size. Fourth, this review focused on prediction models in the prostate cancer domain; such a domain-specific focus was required because of the large number of CPMs that are published across medical domains. As such, one could argue that the findings might not generalize to other clinical areas.

In conclusion, historically few CPM development studies have justified their choice of sample size, with any justification usually being based on EPP. This systematic review has highlighted that the classic EPP $\geq$ 10 rule of thumb is not necessarily sufficient to guide sample size for CPM development, even though it is the most used metric. The findings also show that there is a need to drastically improve the justification of sample sizes used in prediction model studies; justification for how the sample size was arrived at should be the bare minimum regardless of how data were collected (e.g., by conducting a new cohort study, or by using an existing data set already available). The Riley et al. criteria now provide the required means of doing this in the future. Although a historic lack of justification is perhaps unsurprising (given that the Riley et al. criteria have only recently been published), future work should monitor whether this situation improves in the future.

Acknowledgments

Authors’ contributions: Authors G.P.M. and N.P. were responsible for the conceptualization of the research. Author S.D.C. was responsible for undertaking the systematic review, collecting all data therefrom and performing the statistical analysis. G.P.M., N.P., and R.D.R. supervised the project. S.D.C. and G.P.M. wrote the original draft of the paper, whereas all authors (S.D.C., N.P., R.D.R., and G.P.M.) revised and edited the paper critically for scientific content.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.12.011.

References

[1] Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10(2):e1001381.
[2] Steyerberg. Clinical Prediction Models. New York: Springer; 2009.
[3] Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. BMC Med 2010; 8:21.
[4] Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 2006;144:201−9.
[5] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338: b605.
[6] Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.
[7] Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353:i3140.
[8] Martin GP, Mamas MA, Peek N, Buchan I, Sperrin M. Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models. BMC Med Res Methodol 2017;17:1.
[9] Jinks RC, Royston P, Parmar MK. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. BMC Med Res Methodol 2015;15:82.
[10] van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res 2018. 962280218784726.
[11] Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441.
[12] Van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. Stat Neerlandica 2001;55:17−34.
[13] Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study. Stat Methods Med Res 2020. 096228022092141.
[14] Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. J Clin Epidemiol 1995;48: 1495−501.
[15] Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995;48:1503−10.
[16] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373−9.
[17] Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol 2016;76:175−82.
[18] van Smeden M, de Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol 2016;16:163.

[19] Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38:1276−96.
[20] Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part I - continuous outcomes. Stat Med 2019;38:1262−75.
[21] Poyet C, Wettstein MS, Lundon DJ, Bhindi B, Kulkarni GS, Saba K, et al. External evaluation of a Novel prostate cancer risk calculator (ProstateCheck) based on data from the Swiss arm of the ERSPC. J Urol 2016;196:1402−7.
[22] Lowrance WT, Scardino PT. Predictive models for newly diagnosed prostate cancer patients. Rev Urol 2009;11(3):117−26.
[23] Nieboer D, Tomer A, Rizopoulos D, Roobol MJ, Steyerberg EW. Active surveillance: a review of risk-based, dynamic monitoring. Transl Androl Urol 2018;7(1):106−15.
[24] Borque-Fernando Á, Rubio-Briones J, Esteban LM, Collado-Serra A, Pallás-Costa Y, López-González P, et al. The management of active surveillance in prostate cancer: validation of the Canary Prostate Active Surveillance Study risk calculator with the Spanish Urological Association Registry. Oncotarget 2017;8(65):108451−62.
[25] D'Avino V, Palma G, Liuzzi R, Conson M, Doria F, Salvatore M, et al. Prediction of gastrointestinal toxicity after external beam radiotherapy for localized prostate cancer. Radiat Oncol 2015;10:80.
[26] Moreira DM, Howard LE, Sourbeer KN, Amarasekara HS, Chow LC, Cockrell DC, et al. Predicting time from Metastasis to overall survival in castration-resistant prostate cancer: results from SEARCH. Clin Genitourin Cancer 2017;15(1):60−66.e2.
[27] Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097.
[28] Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One 2012;7:e32844.
[29] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 2009;62:e1−34.
[30] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1−73.
[31] R Core Team. R: A Langauge and Envrionment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019. https://www.R-project.org/.
[32] Riley RD. Pmsampsize: calculates the minimum sample size required for developing a multivariable prediction model. R package version 1.0.3. 2019. https://CRAN.R-project.org/package=pmsampsize.
[33] Shukla-Dave A, Hricak H, Kattan MW, Pucar D, Kuroiwa K, Chen HN, et al. The utility of magnetic resonance imaging and spectroscopy for predicting insignificant prostate cancer: an initial analysis. BJU Int 2007;99:786−93.
[34] Wang JY, Zhu Y, Wang CF, Zhang SL, Dai B, Ye DW. A nomogram to predict Gleason sum upgrading of clinically diagnosed localized prostate cancer among Chinese patients. Chin J Cancer 2014;33(5):241−8.
[35] Gronberg H, Adolfsson J, Aly M, Nordstrom T, Wiklund P, Brandberg Y, et al. Prostate cancer screening in men aged 50-69 years (STHLM3): a prospective population-based diagnostic study. Lancet Oncol 2015;16(16):1667−76.
[36] Dalela D, Santiago-Jimenez M, Yousefi K, Karnes RJ, Ross AE, Den RB, et al. Genomic classifier augments the role of pathological features in identifying optimal candidates for adjuvant radiation therapy in patients with prostate cancer: development and internal validation of a multivariable prognostic model. J Clin Oncol 2017;35:1982−90.
[37] Gnanapragasam VJ, Lophatananon A, Wright KA, Muir KR, Gavin A, Greenberg DC. Improving clinical risk stratification at diagnosis in primary prostate cancer: a prognostic modelling study. PLoS Med 2016;13(8):e1002063.
[38] Peters M, Kanthabalan A, Shah TT, McCartan N, Moore CM, Arya M, et al. Development and internal validation of prediction models for biochemical failure and composite failure after focal salvage high intensity focused ultrasound for local radiorecurrent prostate cancer: presentation of risk scores for individual patient prognoses. Urol Oncol 2018;36(1):13.e1.
[39] Peters M, van der Voort van Zyp JR, Moerland MA, Hoekstra CJ, van de Pol S, Westendorp H, et al. Multivariable model development and internal validation for prostate cancer specific survival and overall survival after whole-gland salvage Iodine-125 prostate brachytherapy. Radiother Oncol 2016;119(1):104−10.
[40] Audige L, Bhandari M, Kellam J. How reliable are reliability studies of fracture classifications? A systematic review of their methodologies. Acta Orthop Scand 2004;75(2):184−94.
[41] Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. Plos Med 2012;9(5):1−12.
[42] Ambler G, Brady AR, Royston P. Simplifying a prognostic model: a simulation study based on clinical data. Stat Med 2002;21:3803−22.
[43] Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making 2001;21:45−56.
[44] Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. Stat Med 2016;35:1159−77.
[45] Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth's logistic regression with rare events: accurate effect estimates and predictions? Stat Med 2017;36:2302−17.
[46] Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Med Inform Decis Mak 2012;12:8.
[47] van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014;14:137.