BMJ Open

Submissions from the SPRINT Data Analysis Challenge on clinical risk prediction: a cross-sectional evaluation

Cynthia A Jackevicius,[1,2,3,4,5] JaeJin An,[1] Dennis T Ko,[2,3,6] Joseph S Ross,[7,8,9] Suveen Angraal,[9] Joshua D Wallach,[10,11] Maria Koh,[2] Jeeeun Song,[1] Harlan M Krumholz[8,9,12]

To cite: Jackevicius CA, An JJ, Ko DT, *et al*. Submissions from the SPRINT Data Analysis Challenge on clinical risk prediction: a cross-sectional evaluation. *BMJ Open* 2019;**9**:e025936. doi:10.1136/bmjopen-2018-025936

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-025936).

Received 9 August 2018
Revised 13 December 2018
Accepted 4 February 2019

For numbered affiliations see end of article.

Correspondence to
Dr Cynthia A Jackevicius; cjackevicius@westernu.edu

ABSTRACT

Objectives To collate and systematically characterise the methods, results and clinical performance of the clinical risk prediction submissions to the Systolic Blood Pressure Intervention Trial (SPRINT) Data Analysis Challenge.

Design Cross-sectional evaluation.

Data sources SPRINT Challenge online submission website.

Study selection Submissions to the SPRINT Challenge for clinical prediction tools or clinical risk scores.

Data extraction In duplicate by three independent reviewers.

Results Of 143 submissions, 29 met our inclusion criteria. Of these, 23/29 (79%) reported prediction models for an efficacy outcome (20/23 [87%] of these used the SPRINT study primary composite outcome, 14/29 [48%] used a safety outcome, and 4/29 [14%] examined a combined safety/efficacy outcome). Age and cardiovascular disease history were the most common variables retained in 80% (12/15) of the efficacy and 60% (6/10) of the safety models. However, no two submissions included an identical list of variables intending to predict the same outcomes. Model performance measures, most commonly, the C-statistic, were reported in 57% (13/23) of efficacy and 64% (9/14) of safety model submissions. Only 2/29 (7%) models reported external validation. Nine of 29 (31%) submissions developed and provided evaluable risk prediction tools. Using two hypothetical vignettes, 67% (6/9) of the tools provided expected recommendations for a low-risk patient, while 44% (4/9) did for a high-risk patient. Only 2/29 (7%) of the clinical risk prediction submissions have been published to date.

Conclusions Despite use of the same data source, a diversity of approaches, methods and results was produced by the 29 SPRINT Challenge competition submissions for clinical risk prediction. Of the nine evaluable risk prediction tools, clinical performance was suboptimal. By collating an overview of the range of approaches taken, researchers may further optimise the development of risk prediction tools in SPRINT-eligible populations, and our findings may inform the conduct of future similar open science projects.

Strengths and limitations of this study

- ► Unique systematic examination of clinical risk prediction submissions to the SPRINT Data Analysis Challenge.
- ► Data extraction in duplicate by independent reviewers.
- ► Examination of study methods and clinical applicability of clinical prediction tools.

INTRODUCTION

The Systolic Blood Pressure Intervention Trial (SPRINT) Data Analysis Challenge, hosted by the *New England Journal of Medicine*, set out to explore the potential benefits of sharing data and results of analyses from clinical trials, in the spirit of encouraging open science.[1] This initiative made available published data from the SPRINT Trial, a multinational, randomised, controlled, open-label trial that was terminated early after a median of 3.3 years of follow-up on showing intensive blood pressure therapy improved clinical outcomes more than standard blood pressure therapy in 9361 patients with hypertension without prior stroke or diabetes.[2] Health professionals, researchers and scientists from all over the world were invited to analyse the SPRINT Trial data set in order to identify novel scientific or clinical findings that may advance our understanding of human health.

The value of open science continues to be a subject of ongoing debate.[3 4] Given that the SPRINT Challenge was a highly publicised competition, with a goal of promoting open science efforts for the SPRINT Trial, there may be value in examining what was initially generated and subsequently published from this competition in order to understand the impact of data sharing.[3–9] The next step is to evaluate what the effort of the SPRINT Challenge produced. Therefore, our objective was to conduct a systematic evaluation that collates, and systematically characterises the methods and results of the submissions. We focused on submissions related to clinical risk

prediction, one of the most popular submission types in the competition. While we hypothesised that divergent results for this common objective of clinical risk prediction may represent differences in quality of the methods used, it may also simply reflect a difference in the approaches used. We also sought to test the clinical relevance of any differences in the risk prediction models. Characterising and disseminating the range of approaches and the findings that resulted from crowdsourcing on this topic using a systematic cross-sectional approach may stimulate conversations about what could be done next, which may subsequently prompt these same authors or others to take further initiative in this area of scientific discovery. Furthermore, our findings may help inform the conduct of future similar open science projects.

METHODS

Study eligibility and selection

We used the SPRINT Challenge website as the data source for this study (https://challenge.nejm.org/pages/home). Submissions to the SPRINT Challenge with an objective to develop a clinical prediction tool or clinical risk score were included in our study. Submissions to the SPRINT Challenge with the objective to simply identify risk factors without an objective to develop a tool or score, or submissions without an objective to create a prediction or risk score were excluded. In addition, we excluded submissions focused on surrogate outcomes, such as blood pressure, but included submissions focused on clinical outcomes.

The title, study objective and abstract of each submission were screened in duplicate by two investigators (JJA, JS) independently to determine whether the submissions met the inclusion and exclusion criteria. Discrepancies between the investigators were reviewed by a third investigator (CAJ) with further discussion resolved by consensus as needed.

Data abstraction

Data were extracted based on a standardised data extraction form and common data variable dictionary which were consistent with the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklist.[10] Data were abstracted in duplicate by three independent reviewers (JJA, JDW and SA). Reviewers were first trained on a common set of three submissions, then iteratively on a second set of two submissions, until an agreement rate for abstraction of 89% was reached. After each iteration, a meeting was held to discuss the interpretation of the items where differences existed. Revisions to the data abstraction dictionary were made at each iteration to ensure a common understanding of data abstraction. Reviewers were not blinded to author names for each submission.

Subsequent to reaching good agreement during the training phase, each investigator (JJA, JDW, SA) received two-thirds of the abstracts so that each submission was abstracted in duplicate. We extracted information on the typical steps that are used when developing a clinical risk score, including the statistical modelling approach, inclusion of variables in the model, how risk and benefit were quantified (absolute risk, absolute risk reduction, etc), methods to assess prediction model performance, and internal and external validation testing approaches.[10 11] Completed abstractions were compared and disagreements were reviewed by a fourth study investigator (CAJ), and differences were resolved through discussion and by consensus.

Hypothetical case vignettes

Four vignettes of patients with hypertension representing typical scenarios of patients at high risk and low risk of adverse clinical outcomes as well as high risk and low risk of adverse therapy effects were created by one clinician investigator (DTK) and reviewed by a second clinician investigator (CAJ). The purpose of the cases was to determine how the tools predicted the recommendation for intensive blood pressure therapy management in order to test the clinical relevance of any differences in the risk prediction models. The cases were then reviewed by two other clinician investigators (HMK, JSR) who manage patients with hypertension to determine, based on their clinical knowledge and expertise, whether they would recommend intensive blood pressure lowering therapy for each of the hypothetical patient cases, and then to rank the patient cases from highest to lowest likelihood to recommend intensive blood pressure management therapy. Among those four cases, the two cases (see box 1) with consistent recommendations from the clinicians (one case to recommend, the other case to not recommend intensive blood pressure control) were then applied to those submissions that provided usable risk scores or prediction tools to determine their clinical recommendation for intensive blood pressure therapy. The purpose of selecting only two cases was to test whether the prediction tools would differentiate high benefit and low benefit patient cases and consistently provide a treatment recommendation aligned with that of the clinicians. The well-performing predictive models were defined as the tools which provided consistent recommendations with the clinicians for both patient cases. Data on application of the cases to the risk scores/tools were applied and extracted by three investigators (JJA, SA, MK), with discrepancies resolved through discussion and consensus with a fourth investigator (CAJ). The investigators applying the risk scores/tools to the cases also provided their opinion on usability of the

Box 1 Two hypothetical patient case vignettes

- A 55-year-old white man with a history of smoking, and prior myocardial infarction, blood pressure 140/90, on aspirin, statin, and β blocker and ACE inhibitor for his prior myocardial infarction (MI), creatinine 1.1.
- A 60-year-old white woman, non-smoker, normal lipids, on one blood pressure medication, systolic blood pressure 130/90, creatinine 1.01.

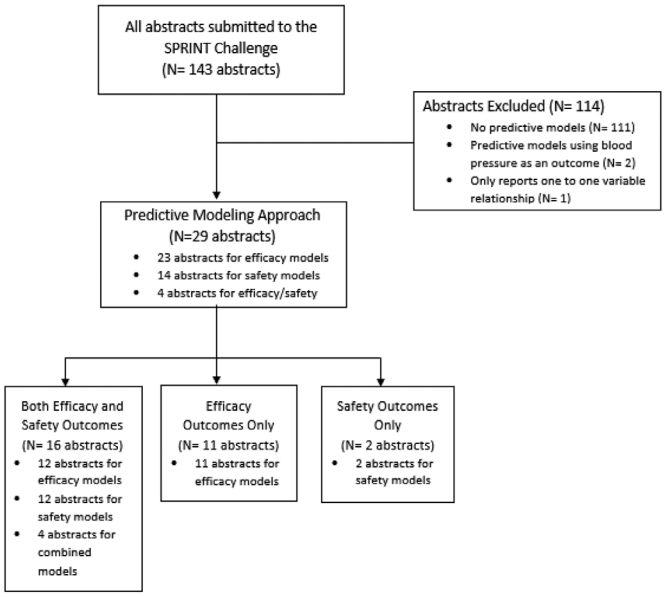

Figure 1 This figure illustrates the selection process of the submissions included in the systematic evaluation and the reasons for exclusion. SPRINT, Systolic Blood Pressure Intervention Trial.

risk scores/tools by completing a survey that included the time required to calculated a score/use the tool, ease of inputting the patient case information into the risk score/tool, understandability of the risk score/tool output and their subjective recommendation on the utility of the risk score/tool for healthcare providers making decisions about managing patients with hypertension. The usability scores were averaged among the three investigators.

Data synthesis and statistical analysis

Data extracted were synthesised quantitatively using descriptive statistics, including mean, median, SD, interquartile intervals (IQIs) or proportions, as appropriate for the data. Risk estimates and recommendations from the tools/scores based on the case scenarios were also summarised descriptively. The proportion of agreement on whether intensive blood pressure lowering was recommended between the tools for each case was determined. Analyses were conducted using SAS V.9.2 (Cary, North Carolina, USA).

Patient involvement

No patients were involved in setting the research question or the outcome measures, nor were they involved in developing plans for recruitment, design or implementation of the study. No patients were asked to advise on interpretation or writing up of results. There are no plans to disseminate the results of the research to study participants or the relevant patient community, aside from publishing the study results.

RESULTS

Out of a total of 143 SPRINT Challenge submissions, 29 submissions met our inclusion/exclusion criteria and were included for analysis (online supplementary appendix I). The most common reason for exclusion was that the submission contained no prediction models (97%; 111 of 114 exclusions) (figure 1). The majority (90%; 26 of 29) of the submissions used the overall SPRINT cohort rather than a subgroup of patients for building prediction models (table 1). Out of the 29 submissions, 10 developed a single prediction model and 12 developed two prediction models, although a maximum of 30 different prediction models were created in one submission. Most submissions (26/29, 89%) considered an efficacy outcome, while 16 of 29 submissions (55%) used both efficacy and safety outcomes in their prediction modelling. The most frequent statistical approach was a traditional multivariable Cox proportional hazards model alone (11/29, 38%), followed by both machine learning and Cox proportional hazards approach combined (9/29, 31%). The most novel approach to create the prediction model was to use machine learning, either with or without a Cox model included. Machine learning techniques were diverse, including supported vector machines, random forest methods, along with use of boosting procedures. Approximately a third (10/29, 35%) of submissions considered absolute net benefit in their risk prediction. Seven of 29 submissions (24%) developed a web-based risk prediction tool and 8 of 29 submissions (28%) developed a clinical score.

A total of 23 distinct abstracts reported prediction models for the efficacy outcome, 14 abstracts presented a model for the safety outcome and four abstracts made predictions for the combined outcome (both efficacy and safety). The vast majority of the efficacy models (20/23, 87%) used the SPRINT primary composite outcome of myocardial infarction, acute coronary syndrome not resulting in myocardial infarction, stroke, acute decompensated heart failure or death from cardiovascular causes as their efficacy outcome, however, safety outcome definitions varied widely. The most frequent safety outcomes used in the model were hypotension, syncope, electrolyte abnormality, acute kidney injury or acute renal failure (9/14, 64%) followed by injurious fall or bradycardia (6/14, 43%).

A median (IQI) of 21 (18–27) candidate variables were used to construct the 23 efficacy models, with 15 models reporting a median of 7 (5–9) variables in the final efficacy prediction models. A median of 20 (18–27) candidate variables were tested in the safety models, with a median of 10 (5–11) variables retained in the 14 final safety models that specified the number of predictors. The highest number of candidate variables and predictors were used in the combined efficacy/safety models, although there were only four models in this category (table 2).

The most common predictor included in the submissions for both efficacy and safety models was age, followed by clinical history of cardiovascular diseases (CVDs) for the efficacy models and race for the safety models (figure 2). Many of these common predictors for efficacy

Table 1 Characteristics of prediction models

Characteristic	N	%
Study population (n=29)	29	
Overall cohort	26	90
Others (patients without CKD, patients without primary end point, unclear)	3	10
Outcomes of prediction models (n=29)		
Both efficacy and safety outcomes	16	55
Efficacy models (a)	12	41
Safety models (b)	12	41
Efficacy and safety combined models	4	14
Efficacy outcome only (c)	11	37
Safety outcome only (d)	2	7
Efficacy outcome model (a), (c) (n=23)		
SPRINT primary composite outcome*	21	91
Safety outcome model (b), (d) (n=14)		
Composite outcome	8	57
Single outcome for each prediction model	6	43
Safety outcome frequencies used in the model		
Hypotension	9	64
Syncope	9	64
Electrolyte abnormality	9	64
Acute kidney injury or acute renal failure	9	64
Bradycardia	6	43
Injurious fall	6	43
Model approach (n=29)		
Multivariable Cox proportional hazards model only	11	38
Multivariable Cox proportional hazards and machine learning†	9	31
Machine learning only†	5	17
Others	4	14
Absolute net benefit calculated (n=29)	10	34
Risk prediction tools (n=29)		
Risk prediction tools developed	7	24
Risk prediction tools provided	2	7
Clinical scores developed (n=29)		
Efficacy clinical scores	4	14
Safety clinical scores	2	7
Efficacy/safety combined clinical scores	2	7
Risk prediction tools/clinical scores provided in a usable format (n=29)	9	31
Web-based risk calculators	2	7
Risk equation	1	3
Clinical scores	3	10
Risk stratification algorithms	3	10

*Myocardial infarction, acute coronary syndrome, stroke, heart failure or death from cardiovascular causes.
†Machine learning techniques include least absolute shrinkage and selection operator (LASSO), generalised, unbiased, interaction detection and estimation (GUIDE) regression tree, weighted k-nearest neighbour model, support vector machines, supervised learning, elastic net regularisation, elastic net binary linear classifier, recursive partition model, random forest, random survival forest, causal forest, boosted classification trees, supervised learning classification and regression trees (CART).
CKD, chronic kidney disease; SPRINT, Systolic Blood Pressure Intervention Trial.

and safety models overlapped. Other frequently identified predictors from the efficacy models were serum urine creatinine ratio, smoking, estimated glomerular filtration rate, sex, race, systolic blood pressure, total cholesterol, high-density lipoprotein and the number of antihypertensive agents. All these predictors were also the most common predictors for the safety models. The frequency of individual predictors included in the final models is shown in figure 2.

Approximately 60% of the abstracts reported prediction model performance measures for the efficacy and safety models, while only one of four of the combined efficacy/safety models did so (table 3). The most frequent performance measure for the 23 efficacy models was the C-statistic; six abstracts (26%) reported C-statistics from the model development phase and seven abstracts (39%) from the internal validation phase. The median (IQI) C-statistic from internal validation was 0.69 (0.64–0.71). Internal validation for the efficacy models was reported in 13 of the abstracts (57%), most frequently using a bootstrapping method (7 abstracts). Only two efficacy model submissions reported external validation of their tools. The performance of the safety models was similar to those of the efficacy models, with a median (IQI) C-statistic from internal validation of 0.68 (0.66–0.72). Five submissions with C-statistics from internal validations were identified with the same purpose, the same data and the same outcomes, but with different methods to build the predictive models. Two submissions using machine learning techniques (elastic net regularisation or least absolute shrinkage and selection operator) reported C-statistics ranging from 0.69 to 0.73, and three submissions using traditional methods (Cox proportional hazards model, or Fine and Gray Cox proportional hazards model) reported C-statistics ranging from 0.64 to 0.69.

Although seven submissions developed web-based risk prediction tools and eight developed clinical scores, only nine of these submissions were available in a usable format in order to apply to the patient cases. These included three clinical scores, three risk-stratification algorithms, two web-based calculators and one risk assessment equation.

Case vignettes

Case 1 represented a patient with high risk of CVD who would be expected to be recommended for intensive blood pressure lowering therapy. After applying the developed tools, the estimated absolute risk of the CVD composite outcome from intensive therapy ranged from 0.05% to 13.1%. Only two of the nine tools explicitly predicted intensive therapy recommendation considering both benefit and risk, while two other prediction tools categorised the patient as having high CVD risk or low harm which may be interpreted as an intensive therapy recommendation, resulting in 44% of the tools providing a recommendation to treat as expected for a high-risk patient. Another three tools categorised the patient into either a low benefit or no significant benefit

Table 2 Variables used in the prediction models

	Efficacy model (abstract, n=23)	Safety model (abstract, n=14)	Efficacy/safety combined models (abstract, n=4)
Candidate variables			
Numbers (%) specified in the abstract	11 (48%)	6 (43%)	2 (50%)
Median number of candidate variables (IQI, range)	21 (IQI: 18–27, range: 9–30)	20 (IQI: 17–26, range: 12–30)	24 (IQI: 22–26, range: 20–28)
All baseline variables/candidate variables	5 (22%)	5 (36%)	1 (25%)
All baseline+blood pressure trajectory	2 (9%)	–	–
Unclear/not available/other	5 (22%)	3 (21%)	1 (25%)
Final variables			
Clearly presented	15 (65%)	10 (71%)	2 (50%)
Median number of final variables (IQI, range)	7 (IQI: 5–9, range: 3–22)	7 (IQI: 5–11, range: 3–22)	12.5 (IQI: 9–16, range: 3–22)
Unclear/not specified	7 (30%)	4 (29%)	2 (50%)
All baseline variables	1 (4%)	–	–

One abstract may report both efficacy and safety models separately, and this abstract is counted twice, as an efficacy model abstract and a safety model abstract.
One abstract may build and report multiple efficacy models, but they are counted as one abstract here.
Note, this table shows the number of abstracts reporting an efficacy, a safety or a combined prediction model.
IQI, interquartile interval.

group from intensive therapy while two tools did not provide any recommendations. Detailed results are available in online supplementary appendix II.

Case two portrayed a patient with low risk of CVD, intended to be a patient that was not a suitable candidate for intensive therapy. After applying the tool to the patient case, two risk scores predicted 'no intensive therapy recommendation', and another three tools categorised the patient into low cardiovascular risk or low benefit group. However, another two prediction models classified this patient into a high benefit group or a benefit with less harm group potentially recommending intensive therapy while two tools did not provide any recommendations.

The risk predictions and therapeutic recommendations from the tools were compared with the recommendations from the clinicians in this study for both patient cases. Recommendations from three of the tools matched the expected therapy recommendations for both cases (well-performing cases); three other tools did not differentiate the two patient cases for therapy recommendations (two tools recommended standard therapy, and one estimated intensive therapy for both cases); one tool recommended the opposite of clinicians' recommendations for both cases; and the final two tools only displayed risk and benefit without predicting a recommendation for any therapy.

In terms of usability, the mean (SD) time required to calculate a score/use the tool was 1.3 (±1.1) min. Only one risk model was an equation format for which investigators took longer than 5 min to calculate the risk. Three investigators responded that inputting the patient information into the risk score was easy or somewhat easy (78%; median [IQI]=4 [3–4]), and the output was easy or somewhat easy to understand (56%; median [IQI]=3 [2–4]). However, despite favourable ease of use or understandable output, 74% of the time the investigators disagreed or strongly disagreed about recommending the tool for healthcare providers making clinical decisions (median [IQI]=2 [1.0–1.5]).

DISCUSSION

We found that although many submissions used the primary composite outcome from the SPRINT Trial, along with similar candidate variables, in their risk prediction models, findings differed substantially. This is most likely the result of employing varying approaches in building the risk score or prediction models by different investigators. The numerous steps that are required when developing a clinical risk score create multiple subjective decision points that may allow for divergent results. For example, researchers must make choices about the statistical modelling approach, statistical thresholds allowed for inclusion and exclusion of model variables, ways to quantify risk and benefit (absolute risk reduction, absolute differences in risk benefit, etc) approach to scoring, methods to assess model performance and interpret results of their internal validation testing of competing models to choose what they consider the best model. These choices are not governed by strict statistical rules, resulting in greater subjectivity and varying judgement in model development processes. Furthermore, although most of the models used similar candidate variables and the same outcome, we found that disparate prediction models resulted with even minute changes in variables or

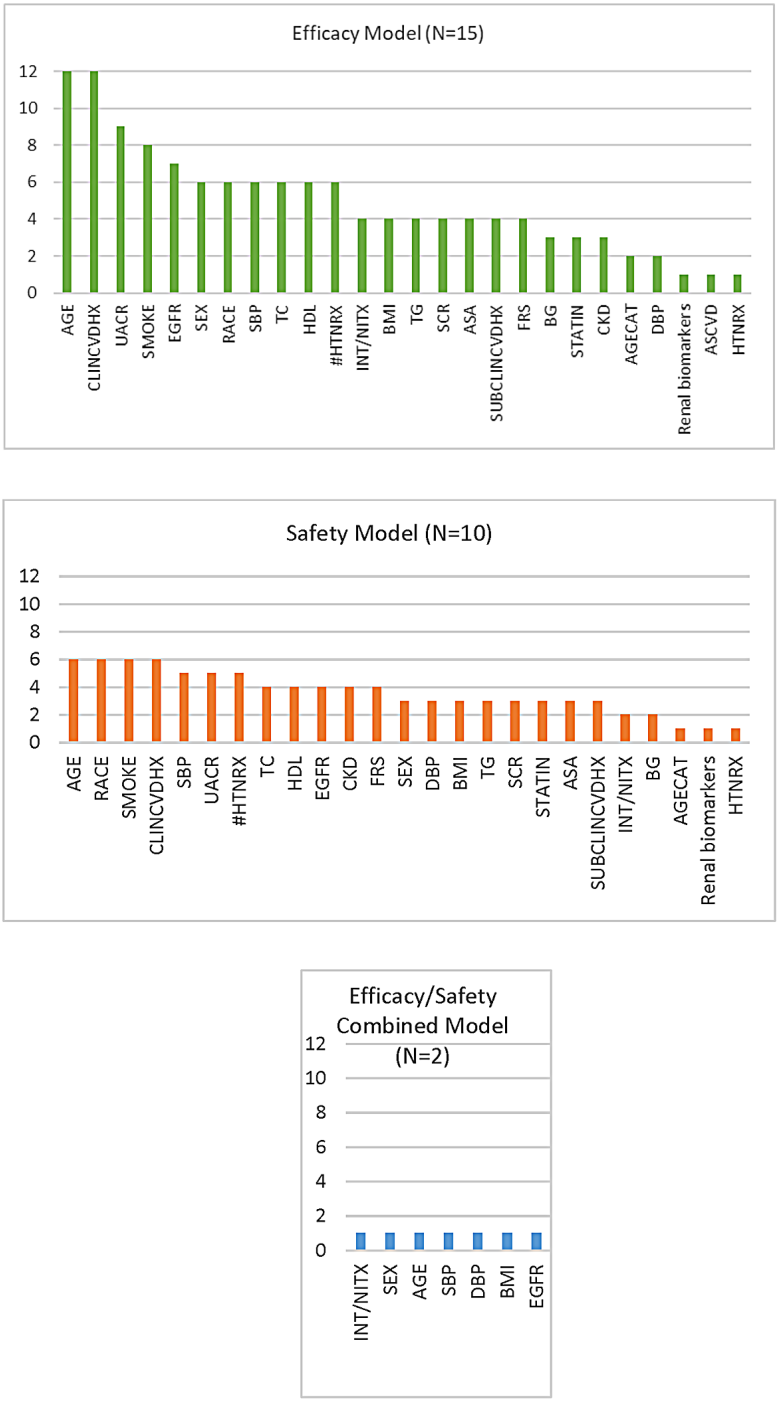

Figure 2 This figure is a bar chart that shows the frequency of variables included in the efficacy, safety and combined efficacy/safety models for the submissions included in the systematic evaluation. The x-axis lists the variables (with abbreviations defined in the footnote) and the y-axis shows the number of models that included each variable in their final prediction models. AGECAT, age category; ASA, daily aspirin use; ASCVD, atherosclerotic cardiovascular disease risk; BG, serum glucose; BMI, body mass index; CKD, indicator of eGFR <60 mL/min/1.73m^2; CLINCVDHX, history of clinical cardiovascular disease; DBP, diastolic blood pressure; EGFR, estimated glomerular filtration rate; FRS, indicator whether 10-year Framingham Risk Score is >15%; HDL, high-density lipoprotein cholesterol; HTNRX, number of distinct antihypertensive agents prescribed; INT/NITX, treatment assignment (either intensive or standard treatment); SBP, systolic blood pressure; SCR, serum creatinine; STATIN, on any statin medication; SUBCLINCVDHX, history of subclinical cardiovascular disease; TC, total cholesterol; TG, triglycerides; UACR, urine albumin/creatinine ratio.

approaches. Our systematic evaluation highlights the diversity of approaches that may be taken to solve the same problem, under the same rules of engagement. Our study which collates these approaches can be foundational for researchers who wish to further examine this research question using the SPRINT data set.

Table 3 Prediction model performance measures

Performance measures	Efficacy model		Safety model		Efficacy/safety combined model	
	Abstract, N	%	Abstract, N	%	Abstract, N	%
Total number of abstracts	23	100%	14	100%	4	100%
Number of abstracts that reported any model performance measures	14	61%	9	64%	1	25%
Discrimination measures						
C-statistics from development	6	26%	5	36%	–	–
Median (IQI, range)*	0.70	(IQI: 0.69–0.71, range: 0.68–0.72)	0.68	(IQI: 0.68–0.70, range: 0.62–0.72)	–	–
Median (IQI, range) for the best-case scenario†	0.71	(IQI: 0.70–0.77, range: 0.68–0.85)	0.69	(IQI: 0.68–0.78, range: 0.62–0.85)		
Median (IQI, range) for the worst-case scenario‡	0.69	(IQI: 0.63–0.70, range: 0.59–0.72)	0.62	(IQI: 0.61–0.68, range: 0.59–0.69)		
C-statistics from internal validation	7	30%	4	29%	-	-
Median	0.69	(IQI: 0.69–0.71, range: 0.64–0.73)	0.68	(IQI: 0.66–0.72, range: 0.65–0.78)	-	-
C-statistics from external validation	–	–	–	–	–	–
Calibration measures	6	26%	5	36%	–	–
Internal validation	13	57%	9	64%	3	75%
Bootstrapping	7	30%	6	43%	–	–
Cross-validation	5	22%	2	14%	1	25%
Split-sample	1	4%	1	7%	2	50%
External validation	2	9%	1	7%	–	–
Correlation between efficacy and safety models	1	4%	–	–	–	–

This table shows number of abstracts that reported efficacy, safety or combined prediction model. One abstract may report both efficacy and safety models separately, and this abstract was included both in the efficacy model abstract and in the safety model abstract.
*In case of multiple C-statistics from one abstract, the median of the ranges was used to summarise the data (two abstracts reported multiple C-statistics).
†Best-case scenario is using the highest C-statistics in case the abstract provided ranges of C-statistics from multiple different models.
‡Worst-case scenario is using the highest C-statistics in case the abstract provided ranges of C-statistics from multiple different models.

These differences became most noticeable and clinically relevant when we applied the available tools to a high-risk and a low-risk SPRINT-eligible patient case. We found that there were few prediction models that created readily available tools that we could assess with the cases, and these tools provided wide-ranging absolute and relative risk estimates and recommendations for managing the hypothetical patients. Only about half of the tools provided the expected recommendation of 'intensive treatment' for the high-risk patient, and 'standard treatment' for the low-risk patient. Given that the cases were chosen to test whether the tools could discriminate between more obvious risk scenarios rather than examine more challenging patients in the grey zone, their poor performance raises concern. The well-performing tools all conducted internal validations, and in addition, one tool conducted external validation, whereas only half of the poorly performing tools conducted internal validations. Also, most of the well-performing tools considered both efficacy and safety outcomes together for clinical recommendations. These characteristics of well-performing tools suggest the need for robust research methods when building clinical prediction models.

There are many steps in developing a clinical prediction rule or risk score.[11] The Transparent Reporting of multivariable prediction model for Individual Prognosis of Diagnosis (TRIPOD) statement checklist includes specification of predictors, outcomes, and model building and performance as key methods steps to report. TRIPOD also states that some form of internal validation is a necessary

part of model development, and strongly recommends external validation.[11] We found that overall only half of the submissions (13/29, 57%) reported internal validation, and even fewer conducted an external validation. In fact, the two published risk scores have both conducted internal validation, and both also conducted external validation with the same Action to Control Cardiovascular Risk in Diabetes Study data set. It is possible that other research teams may not have published their work yet in order to complete their validation, or given the short time line for the competition, may not have had access to a similar external data source with which to conduct external validation. Since most tools were not externally validated, this may in part explain the poor performance of the tools in our high-risk and low-risk patient cases, and the unwillingness of recommending the tool for healthcare providers making clinical decisions. Our study reviewed only the abstracts submitted to the SPRINT Challenge, therefore, the insufficient quality of the abstracts may have limited reviewers from access to all the necessary information, including validation methods that were not included due to word count limits of the submission. Moreover, these SPRINT Challenge submissions did not undergo a standardised peer-review process. Therefore, the quality of the abstracts submitted may be lower than those in peer-reviewed publications, which may have impacted our study findings.

While we found that the most common method used in developing the tools was the traditional approach of choosing variables based on both clinical and statistical significance, many teams instead chose to employ a data-driven, machine-learning approach. At the present time, it is difficult to determine which approach is better. When comparing the model performance of the five submissions with the same study purpose, the same data and the same outcomes, the C-statistics using machine-learning techniques and traditional approaches appeared similar (0.69–0.73 for the machine-learning approach vs 0.64–0.69 for the traditional approach). Moreover, not all these studies conducted external validation or made tools available for our use, therefore, it is difficult to determine which model performs better than the other. When we compared the C-statistics of well-performing models and poorly performing models based on the hypothetical vignettes, the C-statistics were very similar (around 0.70 for both) although a smaller number of studies from the poorly performing models conducted internal validation. As more of the submissions' full methods and results are made publicly accessible through publication, researchers will be able to further examine the benefits and drawbacks of each of the methodological strategies. It is important to note that this study reviewed SPRINT Challenge submissions only, and did not review clinical prediction models or clinical risk scores outside of the SPRINT Challenge. Future research can further evaluate prediction models outside of the SPRINT Challenge.

Just as few meeting abstracts get translated into publications, the SPRINT Challenge submissions may be experiencing the same fate, creating a new form of grey literature.[12] At 1 year after the SPRINT Challenge, few research teams (2/29, 7%) that created risk prediction models have published their results in the peer-reviewed literature.[13 14] Some investigators may have viewed the competition as preliminary work, or did not enter the competition with the intent to publish. In this research area, where 29 submissions addressed similar and important research questions, with diverse options for developing usable risk scores and tools, preprint publication may be a beneficial venue to garner valuable feedback for works in progress.[15]

Our systematic evaluation raises perhaps more questions than it provides answers. Part of our study's purpose was to prompt researchers to review what has been done to date, in order to stimulate further thinking about the next steps to take. We hope that by collating these results, research teams who invested substantial time and effort into the SPRINT Challenge competition will be able to more easily learn from each other about the different approaches taken by the competing teams, and explore why the results differed. Given that there are such different approaches possible, our study highlights the importance of prespecification of the methodological approach, or of declaring that a study is exploratory with multiple comparisons.[16] We hope this review stimulates researchers to take further steps in developing their clinical decision tools, including external validation, which was done infrequently in these submissions, but is recommended by TRIPOD, in order to improve clinical decision-making tools available for patients with hypertension.[11] Given the recent controversy over the 2017 American College of Cardiology/American Heart Association hypertension guidelines, further research investigating the risk/benefit balance of hypertensive treatment is essential.[17]

Furthermore, we anticipate seeing more data-sharing opportunities in the future with the recent interest in the open science movement. Therefore, our findings are likely to be of interest to researchers and clinicians, and that those organising future open science initiatives may also benefit from our systematic evaluation. We offer the following suggestions to organisers of open science competitions to enhance the experience and potential productivity of such future endeavours: (1) Incorporate a greater use of structured reporting of key design elements in the abstract submissions to permit better examination of study methods. (2) Allow a more liberal word count for submissions. (3) Provide a process to foster postcompetition dialogue among research groups. Only time will tell whether this type of open science initiative truly advances science. We believe that our systematic evaluation provides a useful reflection of the initial impact and output of this data-sharing effort as a step forward in this process.

Author affiliations
[1]Pharmacy Department, Western University of Health Sciences, Pomona, California, USA

[2]ICES, Toronto, Ontario, Canada
[3]Institute for Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada
[4]VA Greater Los Angeles Healthcare System, Los Angeles, California, USA
[5]University Health Network, Toronto, Ontario, Canada
[6]Division of Cardiology, Schulich Heart Centre, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada
[7]Section of General Internal Medicine, Department of Medicine, Yale University School of Medicine, New Haven, Connecticut, USA
[8]Department of Health Policy and Management, Yale University School of Public Health, New Haven, Connecticut, USA
[9]Center for Outcomes Research and Evaluation, Yale–New Haven Hospital, New Haven, Connecticut, USA
[10]Department of Environmental Health Sciences, Yale University School of Public Health, New Haven, Connecticut, USA
[11]Collaboration for Research Integrity and Transparency, Yale Law School, New Haven, Connecticut, USA
[12]Section of Cardiovascular Medicine, Department of Medicine, Yale University School of Medicine, New Haven, Connecticut, USA

Twitter @HeartRPh

Contributors CAJ and DTK conceived the study idea. CAJ coordinated the systematic review. CAJ and JJA designed the search strategy. JJA, JS and CAJ screened title and abstracts for inclusion. JJA, SA and JDW acquired the data from the submissions, and CAJ acted as the arbitrator. DTK, JSR and HMK reviewed the cases for clinical recommendations. MK, JJA, SA extracted data related to applicability and applied the relevant tools to the cases. JJA and CAJ performed the data analysis and wrote the first draft of the manuscript. All authors interpreted the data analysis and critically revised the manuscript. CAJ is the guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests DTK is supported by a Mid-Career Investigator Award from the Heart and Stroke Foundation of Canada (HSFC), Ontario Provincial Office. HMK was a recipient of a research grant, through Yale, from Medtronic and the U.S. Food and Drug Administration to develop methods for post-market surveillance of medical devices; is a recipient of research agreements with Medtronic and Johnson & Johnson (Janssen), through Yale, to develop methods of clinical trial data sharing; works under contract with the Centers for Medicare & Medicaid Services to develop and maintain performance measures that are publicly reported; received payment from the Arnold & Porter Law Firm for work related to the Sanofi clopidogrel litigation and from the Ben C.Martin Law Firm for work related to the Cook IVC filter litigation; chairs a Cardiac Scientific Advisory Board for United Health; is a participant/participant representative of the IBM Watson Health Life Sciences Board; is a member of the Advisory Board for Element Science and the Physician Advisory Board for Aetna; and is the founder of Hugo, a personal health information platform. In the past 36 months, JSR received research support through Yale University from Medtronic, Inc. and the Food and Drug Administration (FDA) to develop methods for postmarket surveillance of medical devices (U01FD004585), from the Centers of Medicare and Medicaid Services (CMS) to develop and maintain performance measures that areused for public reporting (HHSM-500-2013-13018I), and from the Blue Cross Blue Shield Association to better understand medical technology evaluation, and he currently receives research support through Yale University from Johnson and Johnson to develop methods of clinical trial data sharing, from the Food and Drug Administration to establish Yale-Mayo Clinic Center for Excellence in Regulatory Science and Innovation (CERSI) program (U01FD005938), from the Agency for Healthcare Research and Quality (R01HS022882), from the National Heart, Lung and Blood Institute of the National Institutes of Health (NIH) (R01HS025164), and from the Laura and John Arnold Foundation to establish the Good Pharma Scorecard at Bioethics International and to establish the Collaboration for Research Integrity and Transparency (CRIT) at Yale. In the past 36 months, JDW has received research support through the Meta Research Innovation Center at Stanford (METRICS) and the Collaboration for Research Integrity and Transparency (CRIT) at Yale from the Laura and John Arnold Foundation.

Patient consent for publication Not required.

Ethics approval This study was reviewed by the Institutional Review Board of Western University of Health Sciences.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Data are available within the tables and appendices. No additional data are available.

REFERENCES

1. Drazen JM, Morrissey S, Malina D, *et al*. The Importance - and the Complexities - of Data Sharing. *N Engl J Med* 2016;375:1182–3.
2. Wright JT, Williamson JD, Whelton PK, *et al*. SPRINT Research Group. A Randomized Trial of Intensive versus Standard Blood-Pressure Control. *N Engl J Med* 2015;373:2103–16.
3. Groves T, Godlee F. Open science and reproducible research. *BMJ* 2012;344:e4383.
4. Ross JS, Krumholz HM. Ushering in a new era of open science through data sharing: the wall must come down. *JAMA* 2013;309:1355–6.
5. Burns NS, Miller PW. Learning What We Didn't Know - The SPRINT Data Analysis Challenge. *N Engl J Med* 2017;376:2205–7.
6. Krumholz HM, Gross CP, Blount KL, *et al*. Sea change in open science and data sharing: leadership by industry. *Circ Cardiovasc Qual Outcomes* 2014;7:499–504.
7. Strom BL, Buyse ME, Hughes J, *et al*. Data Sharing — Is the Juice Worth the Squeeze? *N Engl J Med Overseas Ed* 2016;375:1608–9.
8. Bierer BE, Crosas M, Pierce HH. Data authorship as an incentive to data sharing. *N Engl J Med* 2017;2017.
9. The International Consortium of Investigators for Fairness of Trial Data Sharing. *N Engl J Med* 2016;375:405–7.
10. Moons KG, de Groot JA, Bouwmeester W, *et al*. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11:e1001744.
11. Collins GS, Reitsma JB, Altman DG, *et al*. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:55–63.
12. Basu S, Sussman JB, Rigdon J, *et al*. Benefit and harm of intensive blood pressure treatment: derivation and validation of risk models using data from the SPRINT and ACCORD trials. *PLoS Med* 2017;14:e1002410.
13. Patel KK, Arnold SV, Chan PS, *et al*. Personalizing the Intensity of Blood Pressure Control: Modeling the Heterogeneity of Risks and Benefits From SPRINT (Systolic Blood Pressure Intervention Trial). *Circ Cardiovasc Qual Outcomes* 2017;10:e003624.
14. Scherer RW, Ugarte-Gil C, Schmucker C, *et al*. Authors report lack of time as main reason for unpublished research presented at biomedical conferences: a systematic review. *J Clin Epidemiol* 2015;68:803–10.
15. Lauer MS, Krumholz HM, Topol EJ. Time for a prepublication culture in clinical research? *Lancet* 2015;386:2447–9.
16. Whelton PK, Carey RM, Aronow WS, *et al*. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension* 2018;71:1269–324.
17. Munafo MR, Nosek BA, Bishop DVM, *et al*. A manifesto for reproducible science. *Nature Hum Behav* 2017.