BMJ Open

Submissions from the SPRINT Data Analysis Challenge on Clinical Risk Prediction: A Systematic Review and Applicability

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-025936
Article Type:	Research
Date Submitted by the Author:	09-Aug-2018
Complete List of Authors:	Jackevicius, Cynthia; Western University of Health Sciences, An, JaeJin; Western University of Health Sciences, Ko, Dennis T.; Inst Clin Evaluat Sci Ross, Joseph; Yale University School of Medicine, Internal Medicine Angraal, Suveen; Yale University School of Medicine, Center for Outcomes Research and Evaluation Wallach, Joshua; Yale University, Collaboration for Research Integrity and Transparency; Yale University Center for Outcomes Research and Evaluation, Koh, Maria; Institute for Clinical Evaluative Sciences Song, Jeeeun; Western University of Health Sciences Krumholz, Harlan
Keywords:	open science, SPRINT, Clinical Prediction, Risk Prediction

Submissions from the SPRINT Data Analysis Challenge on Clinical Risk Prediction: A Systematic Review and Applicability

Cynthia A. Jackevicius, BScPhm, PharmD, MSc, Professor [1-5], JaeJin An, BPharm, PhD, Assistant Professor[1], Dennis T. Ko, MD, MSc, Senior Scientist[2,4,6], Joseph S. Ross, MD, MHS, Associate Professor[7,8], Suveen Angraal, MBBS, Post-Doctoral Associate[7], Joshua D. Wallach, PhD, Post-Doctoral Research Fellow[7,9], Maria Koh, MSc, Epidemiologist[2], Jeeeun Song, BSc, Student[1], Harlan M. Krumholz, MD, SM, Professor[7,10,11]

Department of Pharmacy Practice and Administration, College of Pharmacy, Western University of Health Sciences, Pomona, CA[1]; Institute for Clinical Evaluative Sciences, Toronto, Canada[2]; Veterans Affairs Greater Los Angeles Healthcare System, Los Angeles, CA[3]; Institute of Health Policy, Management and Evaluation, Faculty of Medicine, University of Toronto, Toronto, Canada[4]; University Health Network, Toronto, Canada[5]; Division of Cardiology, Schulich Heart Centre, Sunnybrook Health Sciences Centre, University of Toronto, Toronto, Canada[6]; Center for Outcomes Research and Evaluation (CORE), Yale-New Haven Hospital, New Haven, CT[7]; Department of Internal Medicine, Section of General Internal Medicine, Yale School of Medicine, New Haven, CT[8]; Collaboration for Research Integrity and Transparency (CRIT), Yale Law School, New Haven, CT[9];

Department of Medicine, Section of Cardiovascular Medicine[10]; Department of Epidemiology and Public Health, Section of Health Policy and Administration, New Haven, CT[11]

Corresponding author: Cynthia Jackevicius, BScPhm, PharmD, MSc, Western University of Health Sciences, College of Pharmacy, 309 E. Second St., Pomona, CA, 91766 Phone: 909-469-5527 Fax: 909-469-5539 Email: cjackevicius@westernu.edu

Abstract (300 words)

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

Design Systematic review and applicability study.

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 2 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 were produced by the 29 SPRINT Challenge competition submissions for clinical risk prediction. Of the 9 evaluable risk prediction tools, clinical performance was suboptimal. Our findings may be used to stimulate researchers to further optimize the development of risk prediction tools in SPRINT-eligible populations, as well as to inform the conduct of future similar open science projects.

Strengths and Limitations

- Unique systematic examination of clinical risk prediction submissions to the SPRINT Data 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 the published data from the SPRINT trial, a multi-national, randomized, controlled, open-label trial that was terminated early after 3.3 years upon showing intensive blood pressure therapy improved clinical outcomes more than standard blood pressure therapy in 9,361 hypertensive patients without prior stroke or diabetes.[2] Health professionals, researchers and scientists from all over the world were invited to analyze the SPRINT trial dataset 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.[4,5] Given that the SPRINT Challenge was a highly publicized 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.[4-9] The next step is to evaluate what the effort of the SPRINT Challenge produced. Therefore, our objective was to conduct a systematic review that collates, and systematically characterizes 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 hypothesized 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. Characterizing and disseminating the range of approaches and the findings that resulted from crowdsourcing on this topic using a systematic review 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 was screened in duplicate by 2 investigators (JA, JS) independently to determine whether the submissions met the inclusion and exclusion criteria. Discrepancies between the investigators were reviewed by a third investigator (CJ) with further discussion resolved by consensus as needed.

Data Abstraction

Data were extracted based on a standardized 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 (JA, JDW, and SA). Reviewers were first trained on a common set of 3 submissions, then iteratively a second set of 2 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 (JA, JDW, SA) received 2/3 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 modeling approach, inclusion of variables in the model, how risk and benefit was 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 and low risk of adverse clinical outcomes as well as high and low risk of adverse therapy effects were created by one clinician investigator (DK) 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 2 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) 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 (Appendix II). 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 was applied and extracted by 3 investigators (JA, 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 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 synthesized quantitatively using descriptive statistics, including mean, median, standard deviation, interquartile intervals (IQI), , or proportions as appropriate for the data. Risk estimates and recommendations from the tools/scores based on the case scenarios were also summarized 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 v9.2 (Cary, NC). This study was reviewed by the Institutional Review Board of Western University of Health Sciences.

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. (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 2 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 modeling. The most frequent statistical approach was a traditional multivariable Cox proportional hazard (PH) model alone (11/29, 38%), followed by both machine learning and a Cox PH approach combined (9/29, 31%). The most novel approach to create the prediction model was to use machine learning, either without 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 one-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 4 abstracts made predictions for the combined outcome (both efficacy and safety). The vast majority of the efficacy models (87%; 20 of 23) 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 (64% each; 9 of 14) followed by injurious fall or bradycardia (43% each, 6 of 14).

A median (IQI) of 21 (18 to 27) candidate variables were used to construct the 23 efficacy models, with 15 models reporting a median of 7 (5 to 9) variables in the final efficacy prediction models. A median of 20 (18 to 27) candidate variables tested in the safety models, with a median of 10 (5 to 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 4 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 (CVD) for the efficacy models, and race for the safety models. (Figure 2) Many of these common predictors for efficacy 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 1 of 4 of the combined efficacy/safety models did so. (Table 3) The most frequent performance measure for the 23 efficacy models was the C-statistic; 6 abstracts (26%) reported C-statistics from the model development phase and 7 abstracts (39%) from the internal validation phase. The median (IQI) C-statistic from internal validation was 0.69 (0.64 to 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 to 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 regularization or Least Absolute Shrinkage and Selection Operator (LASSO)) reported C-statistics ranges from 0.69 to 0.73, and three submissions using traditional methods (Cox proportional hazards model, or Fine Gray Cox proportional hazards model) reported C-statistics ranges from 0.64 to 0.69.

Although 7 submissions developed web-based risk prediction tools and 8 developed clinical scores, only 9 of these submissions were available in a usable format in order to apply to the patient cases. These included 3 clinical scores, 3 risk stratification algorithms, 2 web-based calculators, and 1 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% up to 13.1%. Only 2 of the 9 tools explicitly predicted intensive therapy recommendation considering both benefit and risk, while 2 other prediction tools categorized 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 3 tools categorized the patient into either a low benefit or no significant benefit group from intensive therapy while 2 tools did not provide any recommendations. Detailed results are available in Appendix II.

Case 2 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, 2 risk scores predicted “no intensive therapy recommendation", and another 3 tools categorized the patient into low CV risk or low benefit group. However, another 2 prediction models classified this patient into a high benefit group or a benefit with less harm group potentially recommending intensive therapy while 2 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 3 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 (2 tools recommended standard therapy, and 1 estimated intensive therapy for both cases); 1 tool recommended the opposite of clinicians' recommendations for both cases; and the final 2 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) minutes. Only one risk model was an equation format for which investigators took longer than 5 minutes 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 to4)), and the output was easy or somewhat easy to understand (56%; median (IQI) = 3 (2 to 4)). However, despite favorable 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 to 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 modeling 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 judgment 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 approaches. Our systematic review 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 dataset.

These differences became most noticeable and clinically relevant when we applied the available tools to a high 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 gray 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 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 2 published risk scores have both conducted internal validation, and both also conducted external validation with the same Action to Control Cardiovascular Risk in Diabetes (ACCORD) study dataset. It is possible that other research teams may not have published their work yet in order to complete their validation. Since most tools were not externally validated, this may in part explain the poor performance of the tools in our high 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 the all necessary information, including validation methods that were not included due to word count limits of the submission.

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 to 0.73 for machine learning vs. 0.64 to 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 another. 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.

Just as few meeting abstracts get translated into publications, the SPRINT Challenge submissions may be experiencing the same fate.[14] At one 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.[12,13] While 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 review 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 pre-specification of the methodological approach, or of declaring that a study is exploratory with multiple comparisons.[17]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 ACC/AHA hypertension guidelines, further research investigating the risk/benefit balance of hypertensive treatment is essential.[16]

Furthermore, we anticipate that those organizing future open science initiatives may also benefit from our systematic review. We offer the following suggestions to enhance the experience and potential productivity of such future endeavors: 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; and 3) provide a process to foster post-competition dialogue amongst 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.

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. The SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. N Engl J Med 2015;373:2103-116.
3. Burns NS, Miller PW. Learning what we didn't know – the SPRINT Data Analysis Challenge. N Engl J Med 2017;376:2205-7.
4. Groves T, Godleee G. Open science and reproducible research. BMJ 2012:344:e4383.
5. 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.
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, Knoppers BM. Data sharing – is the juice worth the squeeze? N Engl J Med 2016;17:1608-9.
8. Bierer BE, Crosas M, Pierce HH. Data authorship as an incentive to data sharing. N Engl J Med 2017; March 29, 2017DOI: 10.1056/NEJMsb1616595.
9. The International Consortium of Investigators for Fairness of Trial Data Sharing. N Engl J Med 2016;375:405-7.
10. Moons KGM, de Groot JAH, 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, Moons KGM. 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, Meerpohl JJ. Authors report lack of time as main reason for unpublished research presented at biomedical conferences: a systematic review. *J Clin Epidemiol.* 2015;68(7):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, Casey DE Jr, Collins KJ, Dennison Himmelfarb C, DePalma SM, Gidding S, Jamerson KA, Jones DW, MacLaughlin EJ, Muntner P, Ovbiagele B, Smith SC Jr, Spencer CC, Stafford RS, Taler SJ, Thomas RJ, Williams KA Sr, Williamson JD, Wright JT Jr. 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: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension. 2017;λλ:eλλλλ–eλλλλ.

17. Munafo MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. Nature Hum Behav 2017; DOI: 10.1038/s41562-016-0021.

Contributors: CAJ and DK conceived the study idea. CAJ coordinated the systematic review. CAJ and JA designed the search strategy. JA, JS and CAJ screened title and abstracts for inclusion. JA, SA, and JW acquired the data from the submissions, and CAJ acted as the arbitrator. DK, JSR, and HMK reviewed the cases for clinical recommendations. MK, JA, SA extracted data related to applicability and applied the relevant tools to the cases. JA and CAJ performed the data analysis. CAJ and JA 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: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests: Dr. Ko is supported by a Mid-Career Investigator Award from the Heart and Stroke Foundation of Canada (HSFC), Ontario Provincial Office. Drs. Ross and Krumholz receive support from Medtronic, Johnson and Johnson, and the Food and Drug Administration to develop methods to enhance postmarket surveillance of medical devices and the Centers for Medicare and Medicaid Services to develop and maintain performance measures that are used for public reporting. Dr. Krumholz is supported by a National Heart, Lung, and Blood Institute Cardiovascular Outcomes Center Award (1U01HL105270-04). Dr. Krumholz chairs a scientific advisory board for UnitedHealthcare. Dr. Krumholz 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. Dr. Ross is supported by the National Institute on Aging (grant K08 AG032886) and by the American Federation for Aging Research through the Paul B. Beeson Career Development Award Program. In the past 36 months, Dr. Wallach 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. No other disclosures are reported. The other authors report no disclosures or conflicts.

Ethical approval: Not required.

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

Transparency: The lead author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Summary Box

What is already known on this topic

143 entries were submitted to the SPRINT Challenge competition

The team that won first place developed a weighted risk-benefit calculator for examining whether intensive treatment would be beneficial for individual patients with hypertension.

Approximately one-quarter of entries were benefit-risk calculators

What this study adds

While a diversity of approaches were used and diverse results were produced by the 29 SPRINT Challenge submissions that focused on clinical risk prediction, few of these submissions underwent both internal and external validation processes that is recommended by current risk prediction methods standards.

Clinical performance of the 9 evaluable risk prediction tools using hypothetical case vignette scenarios was suboptimal.

Our findings may be used by researchers to stimulate future work in this field, and by open science organizers to improve the conduct of open science projects.

Table 1. Characteristics of Prediction Models

Characteristic	N	%
Study Population (N=29)	29	
Overall Cohort	26	90%
Others (Patients without CKD, Patients without Primary Endpoint, 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 PH Model Only	11	38%
Multivariable Cox PH 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%

CKD = Chronic Kidney Disease

*Myocardial infarction, acute coronary syndrome, stroke, heart failure, or death from cardiovascular causes

**Machine learning techniques include Least Absolute Shrinkage and Selection Operator (LASSO), Generalized, Unbiased, Interaction Detection and Estimation (GUIDE) Regression Tree, Weighted k-nearest Neighbor Model, Support Vector Machines, Supervised Learning, Elastic Net Regularization, 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)

Table 2. Variables Used in the Prediction Models

	Efficacy Model (Abstract, N=23)	Safety Model (Abstract, N=14)	Efficacy/Safety Combined Model (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%)	-	-

Note: This table shows the number of abstracts reporting an efficacy, a safety, or a combined prediction model.

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.

Abbreviation: IQI = interquartile interval

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 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%	-	-	-	-

$In case of multiple C-statistics from one abstract, the median of the ranges were used to summarize the data (2 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

Note: This table shows number of abstracts 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.

Abbreviation: IQI = interquartile interval

Box. Two Hypothetical Patient Case Vignettes

#	Case
1	**55 yo white M with history of smoking, and prior myocardial infarction, BP 140/90, on aspirin, statin, and beta blocker and ACE inhibitor for his prior MI. Creatinine 1.1.**
2	**60 yo white female, non-smoker, normal lipids, on one blood pressure medication, SBP 130/90, creatinine of 1.0.**

Figure Legends

Figure 1

This figure illustrates the selection process of the submissions included in the systematic review and the reasons for exclusion.

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 review. 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.

Figure 1. Flow Diagram of Selection of Abstracts

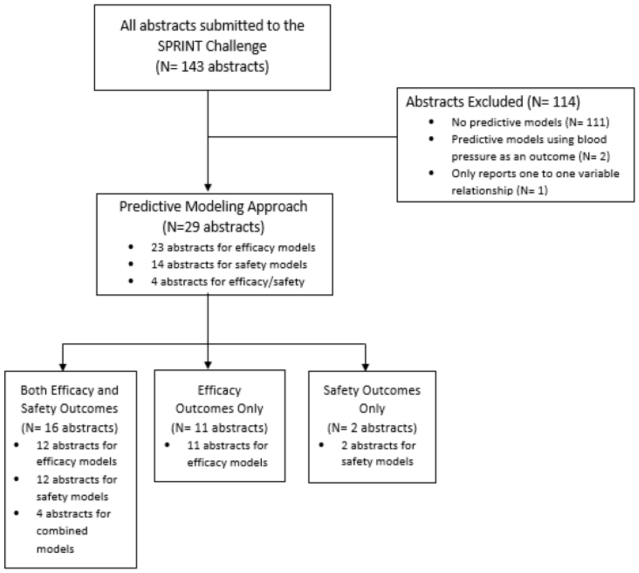
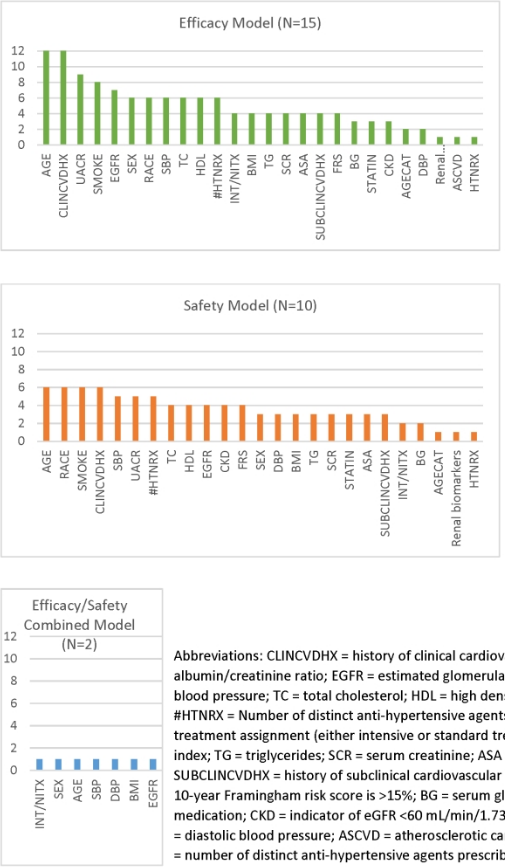

All abstracts submitted to the SPRINT Challenge (N= 143 abstracts)
Abstracts Excluded (N= 114)
No predictive models (N= 111)
Predictive models using blood pressure as an outcome (N= 2)
Only reports one to one variable relationship (N= 1)
Predictive Modeling Approach (N=29 abstracts)
23 abstracts for efficacy models
14 abstracts for safety models
4 abstracts for efficacy/safety
Both Efficacy and Safety Outcomes (N= 16 abstracts)
12 abstracts for efficacy models
12 abstracts for safety models
4 abstracts for combined models
Efficacy Outcomes Only (N= 11 abstracts)
11 abstracts for efficacy models
Safety Outcomes Only (N= 2 abstracts)
2 abstracts for safety models

Figure 1

215x279mm (300 x 300 DPI)

Figure 2. Frequency of Variables Included in the Prediction Models



Abbreviations: CLINCVDHX = history of clinical cardiovascular disease; UACR = urine albumin/creatinine ratio; EGFR = estimated glomerular filtration rate; SBP = systolic blood pressure; TC = total cholesterol; HDL = high density lipoprotein-cholesterol; #HTNRX = Number of distinct anti-hypertensive agents prescribed; INT/NITX = treatment assignment (either intensive or standard treatment); BMI = body mass index; TG = triglycerides; SCR = serum creatinine; ASA = daily aspirin use; SUBCLINCVDHX = history of subclinical cardiovascular disease; FRS = indicator whether 10-year Framingham risk score is >15%; BG = serum glucose; STATIN = on any statin medication; CKD = indicator of eGFR <60 mL/min/1.73m^2; AGECAT = age category; DBP = diastolic blood pressure; ASCVD = atherosclerotic cardiovascular disease risk; HTNRX = number of distinct anti-hypertensive agents prescribed

Figure 2

215x279mm (300 x 300 DPI)

Appendix I. List of Abstracts (Author, Titles, Investigator Information) Included

#	Title	Investigator	Investigator Degree	Number of Co-Investigators	Institution	Institution Location
1	Should all patients be under intensive treatment?	Wenwen Zhang		0	Takeda Pharmaceuticals	Cambridge, MA United States
2	Individual patient data from SPRINT modeled for benefit harm balance demonstrates equivalence for blood pressure targets of 120 and 140 mmHg	Hélène Aschmann		0	University of Zurich	Zurich, ZH Switzerland
3	Individualizing treatment choices in SPRINT trial	João Pedro Ferreira	MD, PhD	2	Centre Hospitalier Universitaire de Nancy	Ludres, 54 France
4	Personalized antihypertensive therapy: using individual variation in population-level statistics to guide clinical decisions	Anish Patnaik		3	McGovern Medical School	Austin, TX United States
5	To Treat Intensively or Not – Individualized Decision Making Support Tool	Noa Dagan	MD, MPH	0	Clalit Research Institute	Tel Aviv, TA Israel
6	A Machine-Learning Model for Personalized Trial Data Exploration	Jochen Lennerz	MD, PhD	2	Massachusetts General Hospital and Harvard Medical School	MA, United States
7	Clinical Prediction Scores of Benefit and Harm from Intensive Blood Pressure Management	Jaejin An	BPharm, PhD	1	Western University of Health Sciences College of Pharmacy	Pomona, CA United States
8	Blood pressure-lowering treatment based on cardiovascular risk compared with systolic blood pressure	Johan Sundstrom	MD PhD	0	Uppsala University	Uppsala, C Sweden
9	Uplift Modeling to Personalize Intensive Blood Pressure Control	Francis Wilson	MD MSCE	0	Yale School of Medicine	New Haven, CT United States

10	Multivariate analysis enables personalized prediction of adverse heart and kidney outcomes	Gel Dinstag		2	Tel Aviv	Tel Aviv, TA Israel
11	Risk-Benefit Assessment of Intensive Blood-Pressure Control	Mikko Venäläinen	MSc	3	CompBiomedTurku	Turku, 19 Finland
12	Exploring heterogeneous treatment effects for stratified blood pressure treatment	Ludovic Trinquart		1	BUSPH Biostatistics	Boston, CA United States
13	Development and Validation of a Clinical Decision Score to Maximize Benefit and Minimize Harm from Intensive Blood Pressure Treatment	Sanjay Basu	MD, PhD	5	Stanford University	Stanford, CA United States
14	Personalized Balance of Benefits and Risks of Hypertension Treatment	Lin Li		1	Biostat Solutions, Inc.	Rockville, MD United States
15	The Treatment Effect of Intensive Blood Pressure Lowering May Follow an Inverted U-shaped Curve Related to Baseline Cardiovascular Risk	Marco Huesch	MBBS, PhD	0	Penn State's Milton S. Hershey Medical Center	Hershey, PA United States
16	Individualizing SPRINT. Going Beyond the Crowd	Nicole Jaspers	MD	5	UMC Utrecht	Utrecht, UT Netherlands
17	Identification of patients with high blood pressure who would benefit from intensive treatment	Yang Xie	PhD, MD	11	UT Southwestern Medical Center	Dallas, TX United States
18	Estimating personalized responses to lower systolic blood pressure targets: a machine learning-based causal analysis of the SPRINT Trial	Aron Baum	PhD	2	Icahn School of Medicine at Mount Sinai	New York, NY United States
19	Personalized blood pressure therapy in hypertensive patients: an analysis of the SPRINT trial	Jan van den Brand	PhD	0	Radboud University Medical Center	Nigmegen, GE Netherlands
20	Features that Predict Poor Outcomes in Hypertensive Non-Diabetic Patients – What Matters Most?	Ronilda Lacson	MD, PhD	5	Brigham and Women's Hospital	Boston, MA United States

21	Identifying Patients Who Do Not Benefit from Intensive Blood-Pressure Control in the Systolic Blood Pressure Intervention Trial (SPRINT)	David Cheng		0	Harvard School of Public Health	Boston, MA United States
22	Using Machine Learning to Personalize Blood Pressure Treatment	Kaveh Danesh		0	University of California, Berkeley	Berkeley, CA United States
23	Individualizing benefit and harm of intensive vs standard blood pressure control: an analysis of SPRINT data	Jacob Udell	MD, MPH	0	University of Toronto	Toronto, Canada
24	Machine learning identifies hypertension patients who do not benefit from intensive treatment	Ljubomir Buturovic		1	Clinical Persona Inc.	East Palo Alto, CA United States
25	Identifying a subgroup with a favorable benefit and risk balance under the intensive treatment	Yan Sun		1	Abbvie Inc	Lake Bluff, IL United States
26	Balancing Benefit and Harm of Intensive Antihypertensive Therapy	Maria Koh		5	Institute for Clinical Evaluative Sciences	Toronto, ON Canada
27	Development of a Prediction Rule for Benefit and Harm of Intensive Blood Pressure Lowering: The SPRINT Score	Manan Pareek	MD, PhD	3	Odense University Hospital	Odense, 83 Denmark
28	Systolic Blood Pressure Intervention Trial (SPRINT) Selection Tool	Janine Bauman	BSN	1	The HOLMES (Health Outcomes Linkage with Medical Electronic System) Team	Cleveland, OH United States
29	Prediction Risk Factors for significant eGFR decrease in patients without CKD, and a Possible Point System	Fei Tang	PhD	0	University of Miami	Miami, FL United States

Appendix II. Case Study Comparisons

Case 1 – High CV Risk Patient

Risk Calculation from Web/App Tools or Equation Provided

ID	Efficacy Outcome	Safety Outcome	Efficacy and Safety Outcomes Combined	No. of Variables Used to Calculate the Risk	Time When Risk Calculated (in years)	AR of Efficacy from Standard Therapy (%)	AR of Efficacy from Intensive Therapy (%)	AR of Safety from Standard Therapy (%)	AR of Safety from Intensive Therapy (%)	ARR of Efficacy (Standard-Intensive, %)	ARI of Safety (Intensive-Standard, %)	Net Benefit (Benefit-Harm) from Intensive Therapy (%)	Interpretation/Recommendation for Intensive Therapy (Based on cutoff provided or NNH/NNT calculated)
6	-	-	Assume composite SPRINT and SAE outcome	5	Not Specified	0.05	0.06	0.56	0.64				No specific recommendation is provided
28	MI, ACS, Stroke, HF, CVD death, Death, AKI	Hypotension, Syncope, Bradycardia, ELYTE, fall, OHYPO-SX, OHYPO-ASX, Albuminuria	-	22	3.3								Color coding to differentiate difference between treatments, 5 levels
16	SPRINT composite outcome	-	-	8	5	2.76	2.1			0.67			iNNT>100 - Low benefit group

Risk Calculation from Clinical Scores Developed

ID	Efficacy Outcome	Safety Outcome	Efficacy and Safety Outcomes Combined	No. of Variables Used to Calculate the Risk	Time When Risk Calculated (in years)	Benefit Score	Harm Score	Benefit and Harm Combined Score	ARR of Efficacy Outcome (Standard - Intensive, %)	ARI of Safety Outcome (Intensive - Standard, %)	Net Benefit (Benefit-Harm) from Intensive Therapy (%)	Interpretation/Recommendation for Intensive Therapy (Based on cutoff provided or NNH/NNT calculated)
7	SPRINT composite outcome	Composite of Hypotension, Syncope, Bradycardia, ELYTE, fall, AKI	-	9	3.3			4	2	2	0	Recommend Intensive Therapy
27	SPRINT composite outcome	Composite of Hypotension, Syncope, ELYTE, fall, AKI	-	9 for Efficacy/ 7 for Safety	Not Specified	5	4		-3			Recommend Intensive Therapy
23	SPRINT composite outcome	Composite of Hypotension, Syncope, Bradycardia, ELYTE, fall, AKI	-	9	3.3			quartile 2	1.29	1.62		Low benefit group. No specific recommendations.

Risk Category Classified from the Submission

ID	Efficacy Outcome	Safety Outcome	No. of Variables Used to Calculate the Risk	Name the Variables Used to Categorize the Risk	Time When Risk Calculated (in years)	AR of Efficacy from Standard Therapy (%)	AR of Efficacy from Intensive Therapy (%)	AR of Safety from Standard Therapy (%)	AR of Safety from Intensive Therapy (%)	ARR of Efficacy (Standard-Intensive, %)	ARI of Safety (Intensive-Standard, %)	HR of Outcome (Intensive vs. Standard)	Interpretation/Recommendation for Intensive Therapy (HR of Intensive vs. Standard)
14	-	Hypotension, AKI	3	Framingham score, kidney disease, total cholesterol	Not Specified			Hypotension (3%), kidney disease (5%)	Hypotension (4%), kidney disease (7%)			HR benefit = 0.74; HR Safety = 1.28 for hypotension, 1.46 for Kidney Disease	Subgroup 1 (Low Harm, Benefit)
15	SPRINT composite outcome	-	3	clinical CVD, age, ascvd risk	Not Specified	13.1	11.6	3.5	6.4	1.5	3		Group D (High CV Risk but No Benefit)
17	SPRINT composite outcome	-	3		Not Specified							HR of benefit = 0.66	High risk

Case 2 – Low CV Risk Patient

Risk Calculation from Web/App Tools or Equation Provided													
ID	**Efficacy Outcome**	**Safety Outcome**	**Efficacy and Safety Outcomes Combined**	**No. of Variables Used to Calculate the Risk**	**Time When Risk Calculated (in years)**	**AR of Efficacy from Standard Therapy (%)**	**AR of Efficacy from Intensive Therapy (%)**	**AR of Safety from Standard Therapy (%)**	**AR of Safety from Intensive Therapy (%)**	**ARR of Efficacy (Standard - Intensive, %)**	**ARI of Safety (Intensive - Standard, %)**	**Net Benefit (Benefit-Harm) from Intensive Therapy (%)**	**Interpretation/Recommendation for Intensive Therapy (Based on cutoff provided or NNH/NNT calculated)**
6	-	-	Assume composite SPRINT and SAE outcome	5	Not Specified	0.06	0.07	0.53	0.79				No specific recommendation is provided
28	MI, ACS, Stroke, HF, CVD death, Death, AKI	Same as above	-	22	3.3								Color coding to differentiate difference between treatments, 5 levels
16	SPRINT composite outcome	-	-	8	5	0.99	0.75			0.24			iNNT>100 - Low benefit group

Risk Calculation from Clinical Scores Developed

ID	Efficacy Outcome	Safety Outcome	Efficacy and Safety Outcomes Combined	No. of Variables Used to Calculate the Risk	Time When Risk Calculated (in years)	Benefit Score	Harm Score	Benefit and Harm Combined Score	ARR of Efficacy Outcome (Standard - Intensive, %)	ARI of Safety Outcome (Intensive - Standard, %)	Net Benefit (Benefit-Harm) from Intensive Therapy (%)	Interpretation/Recommendation for Intensive Therapy (Based on cutoff provided or NNH/NNT calculated)
7	SPRINT composite outcome	Composite of Hypotension, Syncope, Bradycardia, ELYTE, fall, AKI	-	9	3.3			0	2	3.5	-1.5	Recommend Standard Therapy
27	SPRINT composite outcome	Composite of Hypotension, Syncope, ELYTE, fall, AKI	-		Not Specified	0	0		-0.5			Recommend Standard Therapy
23	SPRINT composite outcome	Composite of Hypotension, Syncope, Bradycardia, ELYTE, fall, AKI	-	9	3.3			quartile 1	0.82	0.97		Low benefit group. No specific recommendations.

Risk Category Classified from the Submission													
ID	**Efficacy Outcome**	**Safety Outcome**	**No. of Variables Used to Calculate the Risk**	**Name the Variables Used to Categorize the Risk**	**Time When Risk Calculated (in years)**	**AR of Efficacy from Standard Therapy (%)**	**AR of Efficacy from Intensive Therapy (%)**	**AR of Safety from Standard Therapy (%)**	**AR of Safety from Intensive Therapy (%)**	**ARR of Efficacy (Standard - Intensive, %)**	**ARI of Safety (Intensive - Standard, %)**	**HR of Outcome (Intensive vs. Standard)**	**Interpretation/Recommendation for Intensive Therapy (HR of Intensive vs. Standard)**
14	-	Hypotension, AKI	3	Framingham score, kidney disease, total cholesterol	Not Specified			Hypotension (3%), kidney disease (5%)	Hypotension (4%), kidney disease (7%)			HR benefit = 0.74; HR Safety = 1.28 for hypotension, 1.46 for Kidney Disease	Subgroup 1 (Low Harm, Benefit)
15	SPRINT composite outcome	-	3	clinical CVD, age, ascvd risk	Not Specified	2.8	1.9	1.2	2.2	0.9	1		Group A (Low CV risk but higher Benefit)
17	SPRINT composite outcome	-	3		Not Specified							HR of benefit = 0.83	Low risk

AR=absolute risk; ARR=absolute risk reduction; ARI=absolute risk increase; NNH=number needed to harm; NNT=number needed to treat; SAE=serious adverse events; MI=myocardial infarction; ACS=acute coronary syndrome; HF=heart failure; CVD=cardiovascular diseases; ELYTE=Electrolyte abnormality, fall=Injurious fall, OHYPO-SX=Orthostatic Hypotension with dizziness, OHYPO-ASX= Orthostatic hypotension without dizziness, AKI=acute kidney injury; ASCVD=Atherosclerotic Cardiovascular Disease;

MOOSE Checklist for Meta-analyses of Observational Studies

Item No	Recommendation	Reported on Page No
Reporting of background should include		
1	Problem definition	4
2	Hypothesis statement	4
3	Description of study outcome(s)	7-8
4	Type of exposure or intervention used	6
5	Type of study designs used	6
6	Study population	6
Reporting of search strategy should include		
7	Qualifications of searchers (eg, librarians and investigators)	Title page
8	Search strategy, including time period included in the synthesis and key words	6
9	Effort to include all available studies, including contact with authors	6
10	Databases and registries searched	6
11	Search software used, name and version, including special features used (eg, explosion)	6
12	Use of hand searching (eg, reference lists of obtained articles)	-
13	List of citations located and those excluded, including justification	Appendix I
14	Method of addressing articles published in languages other than English	-
15	Method of handling abstracts and unpublished studies	6
16	Description of any contact with authors	-
Reporting of methods should include		
17	Description of relevance or appropriateness of studies assembled for assessing the hypothesis to be tested	6-8
18	Rationale for the selection and coding of data (eg, sound clinical principles or convenience)	6-8
19	Documentation of how data were classified and coded (eg, multiple raters, blinding and interrater reliability)	6-8
20	Assessment of confounding (eg, comparability of cases and controls in studies where appropriate)	7
21	Assessment of study quality, including blinding of quality assessors, stratification or regression on possible predictors of study results	6-8
22	Assessment of heterogeneity	-
23	Description of statistical methods (eg, complete description of fixed or random effects models, justification of whether the chosen models account for predictors of study results, dose-response models, or cumulative meta-analysis) in sufficient detail to be replicated	8
24	Provision of appropriate tables and graphics	Tables 1-3, Figs 1-2
Reporting of results should include		
25	Graphic summarizing individual study estimates and overall estimate	-
26	Table giving descriptive information for each study included	Table 1, Figure 2
27	Results of sensitivity testing (eg, subgroup analysis)	-
28	Indication of statistical uncertainty of findings	-

Item No	Recommendation	Reported on Page No
Reporting of discussion should include		
29	Quantitative assessment of bias (eg, publication bias)	-
30	Justification for exclusion (eg, exclusion of non-English language citations)	-
31	Assessment of quality of included studies	Table 2
Reporting of conclusions should include		
32	Consideration of alternative explanations for observed results	14-16
33	Generalization of the conclusions (ie, appropriate for the data presented and within the domain of the literature review)	16-17
34	Guidelines for future research	-
35	Disclosure of funding source	20

From: Stroup DF, Berlin JA, Morton SC, et al, for the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Group. Meta-analysis of Observational Studies in Epidemiology. A Proposal for Reporting. *JAMA*. 2000;283(15):2008-2012. doi: 10.1001/jama.283.15.2008.

BMJ Open

Submissions from the SPRINT Data Analysis Challenge on Clinical Risk Prediction: A Cross-Sectional Evaluation

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-025936.R1
Article Type:	Research
Date Submitted by the Author:	13-Dec-2018
Complete List of Authors:	Jackevicius, Cynthia; Western University of Health Sciences, An, JaeJin; Western University of Health Sciences, Ko, Dennis T.; Inst Clin Evaluat Sci Ross, Joseph; Yale University School of Medicine, Internal Medicine Angraal, Suveen; Yale University School of Medicine, Center for Outcomes Research and Evaluation Wallach, Joshua; Yale University, Collaboration for Research Integrity and Transparency; Yale University Center for Outcomes Research and Evaluation, Koh, Maria; Institute for Clinical Evaluative Sciences Song, Jeeeun; Western University of Health Sciences Krumholz, Harlan
<b>Primary Subject Heading</b>:	Research methods
Secondary Subject Heading:	Health services research, Cardiovascular medicine
Keywords:	open science, SPRINT, Clinical Prediction, Risk Prediction

Submissions from the SPRINT Data Analysis Challenge on Clinical Risk Prediction: A Cross-Sectional Evaluation

Cynthia A. Jackevicius, BScPhm, PharmD, MSc, Professor [1-5], JaeJin An, BPharm, PhD, Assistant Professor[1], Dennis T. Ko, MD, MSc, Senior Scientist[2,4,6], Joseph S. Ross, MD, MHS, Associate Professor[7,8], Suveen Angraal, MD, Post-Doctoral Associate[7], Joshua D. Wallach, PhD, MS, Assistant Professor[7,9, 10], Maria Koh, MSc, Epidemiologist[2], Jeeeun Song, BSc, Student[1], Harlan M. Krumholz, MD, SM, Professor[7,11,12]

Department of Pharmacy Practice and Administration, College of Pharmacy, Western University of Health Sciences, Pomona, CA[1]; ICES, Toronto, Canada[2]; Veterans Affairs Greater Los Angeles Healthcare System, Los Angeles, CA[3]; Institute of Health Policy, Management and Evaluation, Faculty of Medicine, University of Toronto, Toronto, Canada[4]; University Health Network, Toronto, Canada[5]; Division of Cardiology, Schulich Heart Centre, Sunnybrook Health Sciences Centre, University of Toronto, Toronto, Canada[6]; Center for Outcomes Research and Evaluation (CORE), Yale-New Haven Hospital, New Haven, CT[7]; Department of Internal Medicine, Section of General Internal Medicine, Yale School of Medicine, New Haven, CT[8]; Collaboration for Research Integrity and Transparency (CRIT), Yale Law School, New Haven, CT[9]; Department of Environmental Health Sciences, Yale School of Public Health[10]; Department of Medicine, Section of Cardiovascular Medicine[11]; Department of Epidemiology and Public Health, Section of Health Policy and Administration, New Haven, CT[12]

Corresponding author: Cynthia Jackevicius, BScPhm, PharmD, MSc, Western University of Health Sciences, College of Pharmacy, 309 E. Second St., Pomona, CA, 91766 Phone: 909-469-5527 Fax: 909-469-5539 Email: cjackevicius@westernu.edu

Abstract (300 words)

Objectives To collate and systematically characterize 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 2 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 were produced by the 29 SPRINT Challenge competition submissions for clinical risk prediction. Of the 9 evaluable risk prediction tools, clinical performance was suboptimal. By collating an overview of the range of approaches taken, researchers may further optimize 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

- Unique systematic examination of clinical risk prediction submissions to the SPRINT Data 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 the published data from the SPRINT trial, a multi-national, randomized, controlled, open-label trial that was terminated early after a median of 3.3 years of follow-up upon showing intensive blood pressure therapy improved clinical outcomes more than standard blood pressure therapy in 9,361 hypertensive patients without prior stroke or diabetes.[2] Health professionals, researchers and scientists from all over the world were invited to analyze the SPRINT trial dataset 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 publicized 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 characterizes 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 hypothesized 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. Characterizing 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 was screened in duplicate by 2 investigators (JA, JS) independently to determine whether the submissions met the inclusion and exclusion criteria. Discrepancies between the investigators were reviewed by a third investigator (CJ) with further discussion resolved by consensus as needed.

Data Abstraction

Data were extracted based on a standardized 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 (JA, JDW, and SA). Reviewers were first trained on a common set of 3 submissions, then iteratively a second set of 2 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 (JA, JDW, SA) received 2/3 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 modeling approach, inclusion of variables in the model, how risk and benefit was 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 and low risk of adverse clinical outcomes as well as high and low risk of adverse therapy effects were created by one clinician investigator (DK) 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 2 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) 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 was applied and extracted by 3 investigators (JA, 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 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 synthesized quantitatively using descriptive statistics, including mean, median, standard deviation, interquartile intervals (IQI), or proportions as appropriate for the data. Risk estimates and recommendations from the tools/scores based on the case scenarios were also summarized 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 v9.2 (Cary, NC). This study was reviewed by the Institutional Review Board of Western University of Health Sciences.

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. (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 2 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 modeling. The most frequent statistical approach was a traditional multivariable Cox proportional hazard (PH) model alone (11/29, 38%), followed by both machine learning and a Cox PH approach combined (9/29, 31%). The most novel approach to create the prediction model was to use machine learning, either without 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 one-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 4 abstracts made predictions for the combined outcome (both efficacy and safety). The vast majority of the efficacy models (87%; 20 of 23) 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 (64% each; 9 of 14) followed by injurious fall or bradycardia (43% each, 6 of 14).

A median (IQI) of 21 (18 to 27) candidate variables were used to construct the 23 efficacy models, with 15 models reporting a median of 7 (5 to 9) variables in the final efficacy prediction models. A median of 20 (18 to 27) candidate variables tested in the safety models, with a median of 10 (5 to 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 4 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 (CVD) for the efficacy models, and race for the safety models. (Figure 2) Many of these common predictors for efficacy 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 1 of 4 of the combined efficacy/safety models did so. (Table 3) The most frequent performance measure for the 23 efficacy models was the C-statistic; 6 abstracts (26%) reported C-statistics from the model development phase and 7 abstracts (39%) from the internal validation phase. The median (IQI) C-statistic from internal validation was 0.69 (0.64 to 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 to 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 regularization or Least Absolute Shrinkage and Selection Operator (LASSO)) reported C-statistics ranges from 0.69 to 0.73, and three submissions using traditional methods (Cox proportional hazards model, or Fine Gray Cox proportional hazards model) reported C-statistics ranges from 0.64 to 0.69.

Although 7 submissions developed web-based risk prediction tools and 8 developed clinical scores, only 9 of these submissions were available in a usable format in order to apply to the patient cases. These included 3 clinical scores, 3 risk stratification algorithms, 2 web-based calculators, and 1 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% up to 13.1%. Only 2 of the 9 tools explicitly predicted intensive therapy recommendation considering both benefit and risk, while 2 other prediction tools categorized 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 3 tools categorized the patient into either a low benefit or no significant benefit group from intensive therapy while 2 tools did not provide any recommendations. Detailed results are available in Appendix II.

Case 2 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, 2 risk scores predicted "no intensive therapy recommendation", and another 3 tools categorized the patient into low CV risk or low benefit group. However, another 2 prediction models classified this patient into a high benefit group or a benefit with less harm group potentially recommending intensive therapy while 2 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 3 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 (2 tools recommended standard therapy, and 1 estimated intensive therapy for both cases); 1 tool recommended the opposite of clinicians' recommendations for both cases; and the final 2 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) minutes. Only one risk model was an equation format for which investigators took longer than 5 minutes 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 to4)), and the output was easy or somewhat easy to understand (56%; median (IQI) = 3 (2 to 4)). However, despite favorable 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 to 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 modeling 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 judgment 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 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 dataset.

These differences became most noticeable and clinically relevant when we applied the available tools to a high 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 gray 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 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 2 published risk scores have both conducted internal validation, and both also conducted external validation with the same Action to Control Cardiovascular Risk in Diabetes (ACCORD) study dataset. It is possible that other research teams may not have published their work yet in order to complete their validation, or given the short timeline 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 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 the all 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 standardized 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 to 0.73 for machine learning vs. 0.64 to 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 another. 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 score 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 one 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 pre-specification 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 ACC/AHA 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 organizing future open science initiatives may also benefit from our systematic evaluation. We offer the following suggestions to organizers of open science competitions to enhance the experience and potential productivity of such future endeavors: 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; and 3) provide a process to foster post-competition dialogue amongst 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.

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. The SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. N Engl J Med 2015;373:2103-116.

3. Groves T, Godlee G. 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, Knoppers BM. Data sharing – is the juice worth the squeeze? N Engl J Med 2016;17:1608-9.

8. Bierer BE, Crosas M, Pierce HH. Data authorship as an incentive to data sharing. N Engl J Med 2017; March 29, 2017DOI: 10.1056/NEJMsb1616595.

9. The International Consortium of Investigators for Fairness of Trial Data Sharing. N Engl J Med 2016;375:405-7.

10. Moons KGM, de Groot JAH, 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, Moons KGM. 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, Meerpohl JJ. Authors report lack of time as main reason for unpublished research presented at biomedical conferences: a systematic review. *J Clin Epidemiol.* 2015;68(7):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, Casey DE Jr, Collins KJ, Dennison Himmelfarb C, DePalma SM, Gidding S, Jamerson KA, Jones DW, MacLaughlin EJ, Muntner P, Ovbiagele B, Smith SC Jr, Spencer CC, Stafford RS, Taler SJ, Thomas RJ, Williams KA Sr, Williamson JD, Wright JT Jr. 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: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension. 2017;λλ:eλλλλ–eλλλλ.

17. Munafo MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. Nature Hum Behav 2017; DOI: 10.1038/s41562-016-0021.

Contributors: CAJ and DK conceived the study idea. CAJ coordinated the systematic review. CAJ and JA designed the search strategy. JA, JS and CAJ screened title and abstracts for inclusion. JA, SA, and JW acquired the data from the submissions, and CAJ acted as the arbitrator. DK, JSR, and HMK reviewed the cases for clinical recommendations. MK, JA, SA extracted data related to applicability and applied the relevant tools to the cases. JA and CAJ performed the data analysis. CAJ and JA 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: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests: Dr. Ko is supported by a Mid-Career Investigator Award from the Heart and Stroke Foundation of Canada (HSFC), Ontario Provincial Office. Drs. Ross and Krumholz receive support from Medtronic, Johnson and Johnson, and the Food and Drug Administration to develop methods to enhance postmarket surveillance of medical devices and the Centers for Medicare and Medicaid Services to develop and maintain performance measures that are used for public reporting. Dr. Krumholz is supported by a National Heart, Lung, and Blood Institute Cardiovascular Outcomes Center Award (1U01HL105270-04). Dr. Krumholz chairs a scientific advisory board for UnitedHealthcare. Dr. Krumholz 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. Dr. Ross is supported by the National Institute on Aging (grant K08 AG032886) and by the American Federation for Aging Research through the Paul B. Beeson Career Development Award Program. In the past 36 months, Dr. Wallach 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. No other disclosures are reported. The other authors report no disclosures or conflicts.

Ethical approval: Not required.

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

Transparency: The lead author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Summary Box

What is already known on this topic

143 entries were submitted to the SPRINT Challenge competition

The team that won first place developed a weighted risk-benefit calculator for examining whether intensive treatment would be beneficial for individual patients with hypertension.

Approximately one-quarter of entries were benefit-risk calculators

What this study adds

While a diversity of approaches were used and diverse results were produced by the 29 SPRINT Challenge submissions that focused on clinical risk prediction, few of these submissions underwent both internal and external validation processes that is recommended by current risk prediction methods standards.

Clinical performance of the 9 evaluable risk prediction tools using hypothetical case vignette scenarios was suboptimal.

Our findings may be used by researchers to stimulate future work in this field, and by open science organizers to improve the conduct of open science projects.

Table 1. Characteristics of Prediction Models

Characteristic	N	%
Study Population (N=29)	29	
Overall Cohort	26	90%
Others (Patients without CKD, Patients without Primary Endpoint, 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 PH Model Only	11	38%
Multivariable Cox PH 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%

CKD = Chronic Kidney Disease
*Myocardial infarction, acute coronary syndrome, stroke, heart failure, or death from cardiovascular causes
**Machine learning techniques include Least Absolute Shrinkage and Selection Operator (LASSO), Generalized, Unbiased, Interaction Detection and Estimation (GUIDE) Regression Tree, Weighted k-nearest Neighbor Model, Support Vector Machines, Supervised Learning, Elastic Net Regularization, 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)

Table 2. Variables Used in the Prediction Models

	Efficacy Model (Abstract, N=23)	Safety Model (Abstract, N=14)	Efficacy/Safety Combined Model (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%)	-	-

Note: This table shows the number of abstracts reporting an efficacy, a safety, or a combined prediction model.
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.

Abbreviation: IQI = interquartile interval

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 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%	-	-	-	-

$In case of multiple C-statistics from one abstract, the median of the ranges were used to summarize the data (2 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

Note: This table shows number of abstracts 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.

Abbreviation: IQI = interquartile interval

Box. Two Hypothetical Patient Case Vignettes

#	Case
1	**55 yo white M with history of smoking, and prior myocardial infarction, BP 140/90, on aspirin, statin, and beta blocker and ACE inhibitor for his prior MI. Creatinine 1.1.**
2	**60 yo white female, non-smoker, normal lipids, on one blood pressure medication, SBP 130/90, creatinine of 1.01.**

Figure Legends

Figure 1

This figure illustrates the selection process of the submissions included in the systematic evaluation and the reasons for exclusion.

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.

Figure 1. Flow Diagram of Selection of Abstracts

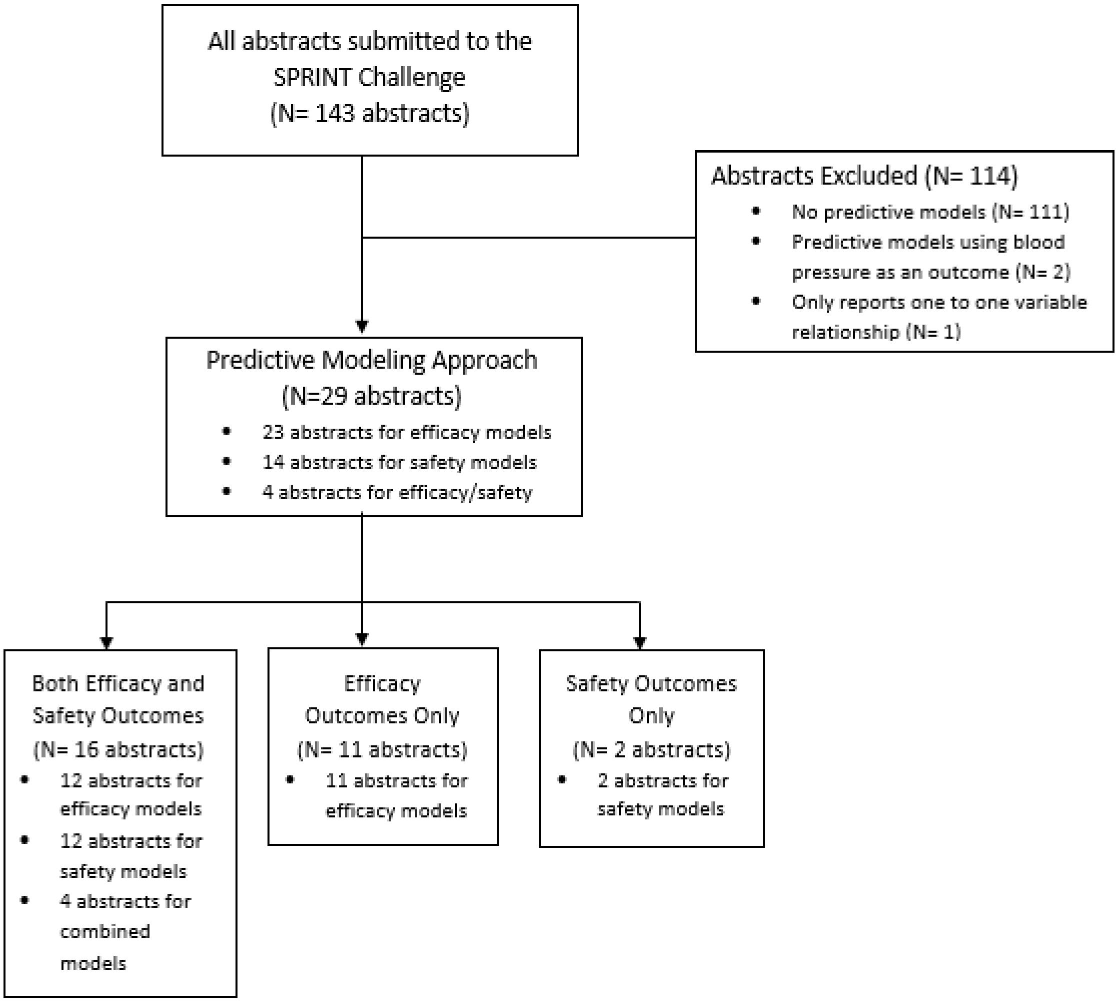
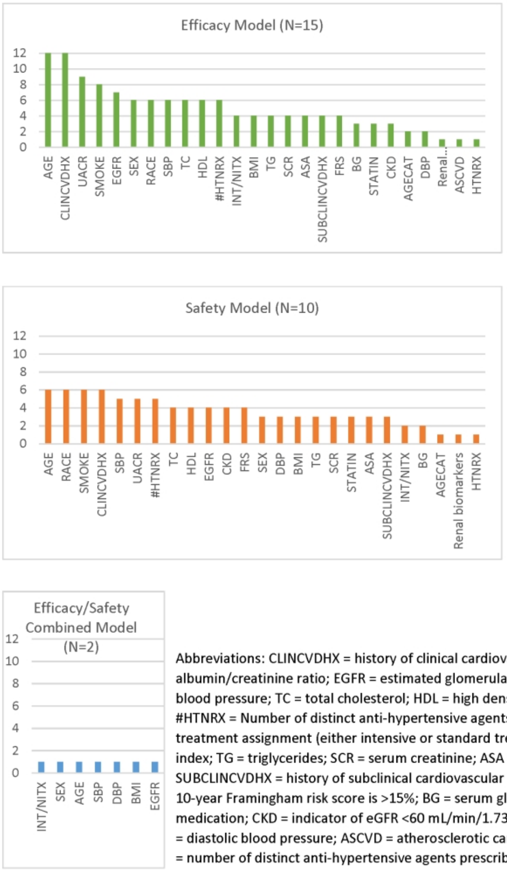

All abstracts submitted to the SPRINT Challenge
(N= 143 abstracts)
Abstracts Excluded (N= 114)
• No predictive models (N= 111)
• Predictive models using blood pressure as an outcome (N= 2)
• Only reports one to one variable relationship (N= 1)
Predictive Modeling Approach
(N=29 abstracts)
• 23 abstracts for efficacy models
• 14 abstracts for safety models
• 4 abstracts for efficacy/safety
Both Efficacy and Safety Outcomes
(N= 16 abstracts)
• 12 abstracts for efficacy models
• 12 abstracts for safety models
• 4 abstracts for combined models
Efficacy Outcomes Only
(N= 11 abstracts)
• 11 abstracts for efficacy models
Safety Outcomes Only
(N= 2 abstracts)
• 2 abstracts for safety models

Figure 2. Frequency of Variables Included in the Prediction Models



Abbreviations: CLINCVDHX = history of clinical cardiovascular disease; UACR = urine albumin/creatinine ratio; EGFR = estimated glomerular filtration rate; SBP = systolic blood pressure; TC = total cholesterol; HDL = high density lipoprotein-cholesterol; #HTNRX = Number of distinct anti-hypertensive agents prescribed; INT/NITX = treatment assignment (either intensive or standard treatment); BMI = body mass index; TG = triglycerides; SCR = serum creatinine; ASA = daily aspirin use; SUBCLINCVDHX = history of subclinical cardiovascular disease; FRS = indicator whether 10-year Framingham risk score is >15%; BG = serum glucose; STATIN = on any statin medication; CKD = indicator of eGFR <60 mL/min/1.73m^2; AGECAT = age category; DBP = diastolic blood pressure; ASCVD = atherosclerotic cardiovascular disease risk; HTNRX = number of distinct anti-hypertensive agents prescribed

Figure 2

215x279mm (300 x 300 DPI)

Appendix I. List of Abstracts (Author, Titles, Investigator Information) Included

#	Title	Investigator	Investigator Degree	Number of Co-Investigators	Institution	Institution Location
1	Should all patients be under intensive treatment?	Wenwen Zhang		0	Takeda Pharmaceuticals	Cambridge, MA United States
2	Individual patient data from SPRINT modeled for benefit harm balance demonstrates equivalence for blood pressure targets of 120 and 140 mmHg	Hélène Aschmann		0	University of Zurich	Zurich, ZH Switzerland
3	Individualizing treatment choices in SPRINT trial	João Pedro Ferreira	MD, PhD	2	Centre Hospitalier Universitaire de Nancy	Ludres, 54 France
4	Personalized antihypertensive therapy: using individual variation in population-level statistics to guide clinical decisions	Anish Patnaik		3	McGovern Medical School	Austin, TX United States
5	To Treat Intensively or Not – Individualized Decision Making Support Tool	Noa Dagan	MD, MPH	0	Clalit Research Institute	Tel Aviv, TA Israel
6	A Machine-Learning Model for Personalized Trial Data Exploration	Jochen Lennerz	MD, PhD	2	Massachusetts General Hospital and Harvard Medical School	MA, United States
7	Clinical Prediction Scores of Benefit and Harm from Intensive Blood Pressure Management	Jaejin An	BPharm, PhD	1	Western University of Health Sciences College of Pharmacy	Pomona, CA United States
8	Blood pressure-lowering treatment based on cardiovascular risk compared with systolic blood pressure	Johan Sundstrom	MD PhD	0	Uppsala University	Uppsala, C Sweden
9	Uplift Modeling to Personalize Intensive Blood Pressure Control	Francis Wilson	MD MSCE	0	Yale School of Medicine	New Haven, CT United States

10	Multivariate analysis enables personalized prediction of adverse heart and kidney outcomes	Gel Dinstag		2	Tel Aviv	Tel Aviv, TA Israel
11	Risk-Benefit Assessment of Intensive Blood-Pressure Control	Mikko Venäläinen	MSc	3	CompBiomedTurku	Turku, 19 Finland
12	Exploring heterogeneous treatment effects for stratified blood pressure treatment	Ludovic Trinquart		1	BUSPH Biostatistics	Boston, CA United States
13	Development and Validation of a Clinical Decision Score to Maximize Benefit and Minimize Harm from Intensive Blood Pressure Treatment	Sanjay Basu	MD, PhD	5	Stanford University	Stanford, CA United States
14	Personalized Balance of Benefits and Risks of Hypertension Treatment	Lin Li		1	Biostat Solutions, Inc.	Rockville, MD United States
15	The Treatment Effect of Intensive Blood Pressure Lowering May Follow an Inverted U-shaped Curve Related to Baseline Cardiovascular Risk	Marco Huesch	MBBS, PhD	0	Penn State's Milton S. Hershey Medical Center	Hershey, PA United States
16	Individualizing SPRINT. Going Beyond the Crowd	Nicole Jaspers	MD	5	UMC Utrecht	Utrecht, UT Netherlands
17	Identification of patients with high blood pressure who would benefit from intensive treatment	Yang Xie	PhD, MD	11	UT Southwestern Medical Center	Dallas, TX United States
18	Estimating personalized responses to lower systolic blood pressure targets: a machine learning-based causal analysis of the SPRINT Trial	Aron Baum	PhD	2	Icahn School of Medicine at Mount Sinai	New York, NY United States
19	Personalized blood pressure therapy in hypertensive patients: an analysis of the SPRINT trial	Jan van den Brand	PhD	0	Radboud University Medical Center	Nigmegen, GE Netherlands
20	Features that Predict Poor Outcomes in Hypertensive Non-Diabetic Patients – What Matters Most?	Ronilda Lacson	MD, PhD	5	Brigham and Women's Hospital	Boston, MA United States

21	Identifying Patients Who Do Not Benefit from Intensive Blood-Pressure Control in the Systolic Blood Pressure Intervention Trial (SPRINT)	David Cheng		0	Harvard School of Public Health	Boston, MA United States
22	Using Machine Learning to Personalize Blood Pressure Treatment	Kaveh Danesh		0	University of California, Berkeley	Berkeley, CA United States
23	Individualizing benefit and harm of intensive vs standard blood pressure control: an analysis of SPRINT data	Jacob Udell	MD, MPH	0	University of Toronto	Toronto, Canada
24	Machine learning identifies hypertension patients who do not benefit from intensive treatment	Ljubomir Buturovic		1	Clinical Persona Inc.	East Palo Alto, CA United States
25	Identifying a subgroup with a favorable benefit and risk balance under the intensive treatment	Yan Sun		1	Abbvie Inc	Lake Bluff, IL United States
26	Balancing Benefit and Harm of Intensive Antihypertensive Therapy	Maria Koh		5	Institute for Clinical Evaluative Sciences	Toronto, ON Canada
27	Development of a Prediction Rule for Benefit and Harm of Intensive Blood Pressure Lowering: The SPRINT Score	Manan Pareek	MD, PhD	3	Odense University Hospital	Odense, 83 Denmark
28	Systolic Blood Pressure Intervention Trial (SPRINT) Selection Tool	Janine Bauman	BSN	1	The HOLMES (Health Outcomes Linkage with Medical Electronic System) Team	Cleveland, OH United States
29	Prediction Risk Factors for significant eGFR decrease in patients without CKD, and a Possible Point System	Fei Tang	PhD	0	University of Miami	Miami, FL United States

Appendix II. Case Study Comparisons

Case 1 – High CV Risk Patient

Risk Calculation from Web/App Tools or Equation Provided													
ID	**Efficacy Outcome**	**Safety Outcome**	**Efficacy and Safety Outcomes Combined**	**No. of Variables Used to Calculate the Risk**	**Time When Risk Calculated (in years)**	**AR of Efficacy from Standard Therapy (%)**	**AR of Efficacy from Intensive Therapy (%)**	**AR of Safety from Standard Therapy (%)**	**AR of Safety from Intensive Therapy (%)**	**ARR of Efficacy (Standard-Intensive, %)**	**ARI of Safety (Intensive-Standard, %)**	**Net Benefit (Benefit-Harm) from Intensive Therapy (%)**	**Interpretation/Recommendation for Intensive Therapy (Based on cutoff provided or NNH/NNT calculated)**
6	-	-	Assume composite SPRINT and SAE outcome	5	Not Specified	0.05	0.06	0.56	0.64				No specific recommendation is provided
28	MI, ACS, Stroke, HF, CVD death, Death, AKI	Hypotension, Syncope, Bradycardia, ELYTE, fall, OHYPO-SX, OHYPO-ASX, Albuminuria	-	22	3.3								Color coding to differentiate difference between treatments, 5 levels
16	SPRINT composite outcome	-	-	8	5	2.76	2.1			0.67			iNNT>100 - Low benefit group

Risk Calculation from Clinical Scores Developed

ID	Efficacy Outcome	Safety Outcome	Efficacy and Safety Outcomes Combined	No. of Variables Used to Calculate the Risk	Time When Risk Calculated (in years)	Benefit Score	Harm Score	Benefit and Harm Combined Score	ARR of Efficacy Outcome (Standard - Intensive, %)	ARI of Safety Outcome (Intensive - Standard, %)	Net Benefit (Benefit -Harm) from Intensive Therapy (%)	Interpretation/Recommendation for Intensive Therapy (Based on cutoff provided or NNH/NNT calculated)
7	SPRINT composite outcome	Composite of Hypotension, Syncope, Bradycardia, ELYTE, fall, AKI	-	9	3.3			4	2	2	0	Recommend Intensive Therapy
27	SPRINT composite outcome	Composite of Hypotension, Syncope, ELYTE, fall, AKI	-	9 for Efficacy/ 7 for Safety	Not Specified	5	4		-3			Recommend Intensive Therapy
23	SPRINT composite outcome	Composite of Hypotension, Syncope, Bradycardia, ELYTE, fall, AKI	-	9	3.3			quartile 2	1.29	1.62		Low benefit group. No specific recommendations.

Risk Category Classified from the Submission

ID	Efficacy Outcome	Safety Outcome	No. of Variables Used to Calculate the Risk	Name the Variables Used to Categorize the Risk	Time When Risk Calculated (in years)	AR of Efficacy from Standard Therapy (%)	AR of Efficacy from Intensive Therapy (%)	AR of Safety from Standard Therapy (%)	AR of Safety from Intensive Therapy (%)	ARR of Efficacy (Standard-Intensive, %)	ARI of Safety (Intensive-Standard, %)	HR of Outcome (Intensive vs. Standard)	Interpretation/Recommendation for Intensive Therapy (HR of Intensive vs. Standard)
14	-	Hypotension, AKI	3	Framingham score, kidney disease, total cholesterol	Not Specified			Hypotension (3%), kidney disease (5%)	Hypotension (4%), kidney disease (7%)			HR benefit = 0.74; HR Safety = 1.28 for hypotension, 1.46 for Kidney Disease	Subgroup 1 (Low Harm, Benefit)
15	SPRINT composite outcome	-	3	clinical CVD, age, ascvd risk	Not Specified	13.1	11.6	3.5	6.4	1.5	3		Group D (High CV Risk but No Benefit)
17	SPRINT composite outcome	-	3		Not Specified							HR of benefit = 0.66	High risk

Case 2 – Low CV Risk Patient

Risk Calculation from Web/App Tools or Equation Provided													
ID	**Efficacy Outcome**	**Safety Outcome**	**Efficacy and Safety Outcomes Combined**	**No. of Variables Used to Calculate the Risk**	**Time When Risk Calculated (in years)**	**AR of Efficacy from Standard Therapy (%)**	**AR of Efficacy from Intensive Therapy (%)**	**AR of Safety from Standard Therapy (%)**	**AR of Safety from Intensive Therapy (%)**	**ARR of Efficacy (Standard - Intensive, %)**	**ARI of Safety (Intensive - Standard, %)**	**Net Benefit (Benefit-Harm) from Intensive Therapy (%)**	**Interpretation/Recommendation for Intensive Therapy (Based on cutoff provided or NNH/NNT calculated)**
6	-	-	Assume composite SPRINT and SAE outcome	5	Not Specified	0.06	0.07	0.53	0.79				No specific recommendation is provided
28	MI, ACS, Stroke, HF, CVD death, Death, AKI	Same as above	-	22	3.3								Color coding to differentiate difference between treatments, 5 levels
16	SPRINT composite outcome	-	-	8	5	0.99	0.75			0.24			iNNT>100 - Low benefit group

Risk Calculation from Clinical Scores Developed

ID	Efficacy Outcome	Safety Outcome	Efficacy and Safety Outcomes Combined	No. of Variables Used to Calculate the Risk	Time When Risk Calculated (in years)	Benefit Score	Harm Score	Benefit and Harm Combined Score	ARR of Efficacy Outcome (Standard - Intensive, %)	ARI of Safety Outcome (Intensive - Standard, %)	Net Benefit (Benefit-Harm) from Intensive Therapy (%)	Interpretation/Recommendation for Intensive Therapy (Based on cutoff provided or NNH/NNT calculated)
7	SPRINT composite outcome	Composite of Hypotension, Syncope, Bradycardia, ELYTE, fall, AKI	-	9	3.3			0	2	3.5	-1.5	Recommend Standard Therapy
27	SPRINT composite outcome	Composite of Hypotension, Syncope, ELYTE, fall, AKI	-		Not Specified	0	0		-0.5			Recommend Standard Therapy
23	SPRINT composite outcome	Composite of Hypotension, Syncope, Bradycardia, ELYTE, fall, AKI	-	9	3.3			quartile 1	0.82	0.97		Low benefit group. No specific recommendations.

Risk Category Classified from the Submission													
ID	Efficacy Outcome	Safety Outcome	No. of Variables Used to Calculate the Risk	Name the Variables Used to Categorize the Risk	Time When Risk Calculated (in years)	AR of Efficacy from Standard Therapy (%)	AR of Efficacy from Intensive Therapy (%)	AR of Safety from Standard Therapy (%)	AR of Safety from Intensive Therapy (%)	ARR of Efficacy (Standard - Intensive, %)	ARI of Safety (Intensive - Standard, %)	HR of Outcome (Intensive vs. Standard)	Interpretation/Recommendation for Intensive Therapy (HR of Intensive vs. Standard)
14	-	Hypotension, AKI	3	Framingham score, kidney disease, total cholesterol	Not Specified			Hypotension (3%), kidney disease (5%)	Hypotension (4%), kidney disease (7%)			HR benefit = 0.74; HR Safety = 1.28 for hypotension, 1.46 for Kidney Disease	Subgroup 1 (Low Harm, Benefit)
15	SPRINT composite outcome	-	3	clinical CVD, age, ascvd risk	Not Specified	2.8	1.9	1.2	2.2	0.9	1		Group A (Low CV risk but higher Benefit)
17	SPRINT composite outcome	-	3		Not Specified							HR of benefit = 0.83	Low risk

AR=absolute risk; ARR=absolute risk reduction; ARI=absolute risk increase; NNH=number needed to harm; NNT=number needed to treat; SAE=serious adverse events; MI=myocardial infarction; ACS=acute coronary syndrome; HF=heart failure; CVD=cardiovascular diseases; ELYTE=Electrolyte abnormality, fall=Injurious fall, OHYPO-SX=Orthostatic Hypotension with dizziness, OHYPO-ASX= Orthostatic hypotension without dizziness, AKI=acute kidney injury; ASCVD=Atherosclerotic Cardiovascular Disease;

MOOSE Checklist for Meta-analyses of Observational Studies

Item No	Recommendation	Reported on Page No
Reporting of background should include		
1	Problem definition	4
2	Hypothesis statement	4
3	Description of study outcome(s)	7-8
4	Type of exposure or intervention used	6
5	Type of study designs used	6
6	Study population	6
Reporting of search strategy should include		
7	Qualifications of searchers (eg, librarians and investigators)	Title page
8	Search strategy, including time period included in the synthesis and key words	6
9	Effort to include all available studies, including contact with authors	6
10	Databases and registries searched	6
11	Search software used, name and version, including special features used (eg, explosion)	6
12	Use of hand searching (eg, reference lists of obtained articles)	-
13	List of citations located and those excluded, including justification	Appendix I
14	Method of addressing articles published in languages other than English	-
15	Method of handling abstracts and unpublished studies	6
16	Description of any contact with authors	-
Reporting of methods should include		
17	Description of relevance or appropriateness of studies assembled for assessing the hypothesis to be tested	6-8
18	Rationale for the selection and coding of data (eg, sound clinical principles or convenience)	6-8
19	Documentation of how data were classified and coded (eg, multiple raters, blinding and interrater reliability)	6-8
20	Assessment of confounding (eg, comparability of cases and controls in studies where appropriate)	7
21	Assessment of study quality, including blinding of quality assessors, stratification or regression on possible predictors of study results	6-8
22	Assessment of heterogeneity	-
23	Description of statistical methods (eg, complete description of fixed or random effects models, justification of whether the chosen models account for predictors of study results, dose-response models, or cumulative meta-analysis) in sufficient detail to be replicated	8
24	Provision of appropriate tables and graphics	Tables 1-3, Figs 1-2
Reporting of results should include		
25	Graphic summarizing individual study estimates and overall estimate	-
26	Table giving descriptive information for each study included	Table 1, Figure 2
27	Results of sensitivity testing (eg, subgroup analysis)	-
28	Indication of statistical uncertainty of findings	-

Item No	Recommendation	Reported on Page No
Reporting of discussion should include		
29	Quantitative assessment of bias (eg, publication bias)	-
30	Justification for exclusion (eg, exclusion of non-English language citations)	-
31	Assessment of quality of included studies	Table 2
Reporting of conclusions should include		
32	Consideration of alternative explanations for observed results	14-16
33	Generalization of the conclusions (ie, appropriate for the data presented and within the domain of the literature review)	16-17
34	Guidelines for future research	-
35	Disclosure of funding source	20

From: Stroup DF, Berlin JA, Morton SC, et al, for the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Group. Meta-analysis of Observational Studies in Epidemiology. A Proposal for Reporting. *JAMA*. 2000;283(15):2008-2012. doi: 10.1001/jama.283.15.2008.