Received: 3 April 2019 | Revised: 12 December 2019 | Accepted: 14 January 2020

DOI: 10.1002/sim.8489

RESEARCH ARTICLE

Statistics in Medicine WILEY

On the aggregation of published prognostic scores for causal inference in observational studies

Tri-Long Nguyen[1,2,3] | **Gary S. Collins**[4] | **Fabio Pellegrini**[5] | **Karel G.M. Moons**[2,6] | **Thomas P.A. Debray**[2,6]

[1]Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

[2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

[3]Department of Pharmacy, Nîmes University Hospital Centre, Nîmes, France

[4]National Institute for Health Research Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK

[5]Biogen International GmbH, Zug, Switzerland

[6]Cochrane Netherlands, University Medical Center Utrecht, Utrecht, The Netherlands

Correspondence
Tri-Long Nguyen, Section of Epidemiology, Department of Public Health, University of Copenhagen, CSS, Øster Farimagsgade 5, DK-1014 Copenhagen, Denmark.
Email: long@sund.ku.dk

Funding information
The Netherlands Organization for Health Research and Development, 91617050, 91215058

As real world evidence on drug efficacy involves nonrandomized studies, statistical methods adjusting for confounding are needed. In this context, prognostic score (PGS) analysis has recently been proposed as a method for causal inference. It aims to restore balance across the different treatment groups by identifying subjects with a similar prognosis for a given reference exposure ("control"). This requires the development of a multivariable prognostic model in the control arm of the study sample, which is then *extrapolated* to the different treatment arms. Unfortunately, large cohorts for developing prognostic models are not always available. Prognostic models are therefore subject to a dilemma between overfitting and parsimony; the latter being prone to a violation of the assumption of no unmeasured confounders when important covariates are ignored. Although it is possible to limit overfitting by using penalization strategies, an alternative approach is to adopt evidence synthesis. Aggregating previously published prognostic models may improve the generalizability of PGS, while taking account of a large set of covariates—even when limited individual participant data are available. In this article, we extend a method for prediction model aggregation to PGS analysis in nonrandomized studies. We conduct extensive simulations to assess the validity of model aggregation, compared with other methods of PGS analysis for estimating marginal treatment effects. We show that aggregating existing PGS into a "meta-score" is robust to misspecification, even when elementary scores wrongfully omit confounders or focus on different outcomes. We illustrate our methods in a setting of treatments for asthma.

KEYWORDS
aggregation, causal inference, observational study, prognostic score, regression modelling

1 | INTRODUCTION

Nonrandomized studies are an important source of evidence for causal inference, and often used to assess the safety and effectiveness of certain treatments.[1] Because nonrandomized studies are observational by nature, differences in individual

outcomes cannot directly be attributed to the treatment exposure, and the estimation of comparative treatment effects is therefore prone to confounding bias.[2]

Over the past few decades, several methods for addressing confounding in nonrandomized studies have been proposed. Amongst these, prognostic score (PGS) analysis (also referred to as "disease-risk score analysis") has recently gained popularity in the medical literature.[3-9] Initially formalized by Hansen[9] in the case of binary treatments, the theory of the PGS has since been extended to the case of general treatment regimes.[6] This approach can be seen as the "prognostic analogue of the propensity score" and is aimed at achieving a prognostic balance across the different treatment arms or exposure groups.[6,9] Briefly, PGS analysis involves the development of a multivariable model predicting an efficacy or safety endpoint (eg, mortality) in patients under a specific exposure (often the control exposure). This model is then applied to the patients of the remaining treatment exposure groups, to estimate their most likely outcome if they had received the control exposure.[6,9] In other words, the PGS analysis develops a prediction model in a set of control individuals, and then *extrapolates* its predictions to individuals of other exposure groups. Hansen[9] (in binary exposures), then Nguyen and Debray[6] (in general treatment regimes), showed that conditioning on this PGS allows to estimate treatment effects that are free from confounding bias. Yet, a key requirement for PGS analysis is that the developed prognostic model provides sufficiently accurate predictions of the (potential) outcome in the control and treated populations. This implies that the prognostic model should properly account for covariates that are associated with the outcome under the control exposure, particularly if their distribution varies across the treatment groups.[6]

Unfortunately, developing accurate prognostic models is subject to dilemmas when few control individuals are available. On the one hand, considering too numerous covariates is likely to result in overfitting, which impedes the generalizability of the prognostic model to other treatment groups. On the other hand, many parsimonious modelling strategies such as penalization and variable selection (eg, least absolute shrinkage and selection operator [LASSO] regression) typically lead to biased outcome predictions, which in turn might introduce (unmeasured) confounding in PGS analysis. To avoid this problem, it has been suggested to use prognostic models developed from large external cohorts of control individuals.[3-5,8] However, there is a growing evidence that published prognostic models tend to have limited transportability across different settings and populations, especially when studies adopt different selection criteria, different variable and outcome definitions, or different measurement methods. For this reason, it is recommended to tailor or update existing models before implementing them in new populations.[10,11]

Rather than focusing on a single external prognostic model, a much more effective strategy might be to utilize and combine multiple previously published prognostic models. In particular, by combining multiple prediction models it becomes possible to account for a large number of covariates without having to estimate their individual effects in the data at hand—which may limit both the risk of overfitting and that of unmeasured confounding in PGS analysis. Debray et al[12] previously showed that aggregating multiple published prognostic models, rather than developing or tailoring/updating a single model, helps improve predictive accuracy in new settings (ie, transportability), particularly when model development samples are relatively small. In this article, we extend this aggregation approach to PGS analysis for causal inference purposes in nonrandomized studies.

The article is structured as follows: Section 2 presents an overview of the framework of PGS analysis; Section 3 describes a method for aggregating published prognostic models within the context of PGS analysis; Section 4 displays an illustrative study; Section 5 reports two series of simulations; finally, Section 6 opens a discussion on our proposed approach to PGS analysis.

2 | PGS ANALYSIS FOR CAUSAL INFERENCE: AN OVERVIEW

2.1 | Causal estimands

Following Rubin's counterfactual model,[2] for individual i, let Z_i be the treatment status (we consider a binary treatment exposure in this article: $Z_i = 1$, treated individual; $Z_i = 0$, control individual), $Y_{(1)i}$ and $Y_{(0)i}$ the *potential* outcomes if Z_i were set to be equal to 1 and 0, respectively, and $\boldsymbol{X}_i \in \mathcal{X}$ a set of measurable prognostic covariates.

Under the "stable unit treatment value assumption", individual potential responses $Y_{(\cdot)i}$ are not influenced by other units $j, \forall j \neq i$.[13] Hence, for individual i the effect caused by the treatment, say "individual treatment effect" (ITE), is defined as:

$$\mathrm{ITE}_i = Y_{(1)i} - Y_{(0)i}.$$

In practice, ITE_i can never be measured—which is referred to as the "fundamental problem of causal inference"[14]—as only $Y_{(1)i}$ or only $Y_{(0)i}$ can be observed, but never both. Under the "consistency rule", the *observed* outcome Y_i is usually defined as: $Y_i = ZY_{(1)i} + (1-Z)Y_{(0)i}$.[15]

An quantity of interest to be estimated is the Average Treatment effect in the Entire population (ATE); that is, the average difference in outcome that would be caused if all individuals were to receive the treatment vs if all individuals were not to receive the treatment (but the control):

$$\text{ATE} = E(Y_1) - E(Y_0) = E(Y_1 - Y_0) = E(\text{ITE}).$$

(Index i disappears since potential outcomes and ITEs are averaged across individuals.)

Another quantity of interest is the Average Treatment effect in the Treated population (ATT); that is, the average difference in outcome that would be caused if the treated population were treated ("factual" condition) vs if the treated population were not treated but under the control exposure ("counterfactual" condition):

$$\text{ATT} = E(Y_1|Z=1) - E(Y_0|Z=1) = E(Y_1 - Y_0|Z=1) = E(\text{ITE}|Z=1).$$

As opposed to the ATE, here the expectation is iterated over the treated population instead of the entire population. The ATT is often an estimand of primary interest since it refers to the causal effect of the treatment in the specific population that it is intended for.[16] Hence, we focus on the estimation of this estimand in the current article.

In nonrandomized studies, treatment allocation (eg, drug prescription) is often related to several (un)measured covariates. As a consequence, no causal effect can be directly inferred from the average difference in the observed outcomes across the treatment arms.[2] (In other words, there is no reason for $E(Y|Z=1) - E(Y|Z=0)$ to coincide with the ATT or the ATE.) In response to this issue, multivariable scoring methods for addressing confounding have been proposed, such as the propensity score analysis and its prognostic analogue.[6,9,17,18]

2.2 | Prognostic scores

Briefly, PGS analysis aims to achieve a prognostic balance between different exposure groups. It thereby approximates a design in which treated and control individuals would have a similar distribution of prognosis if they were left untreated (ie, under the control exposure).[9] Note, nonrandomized clinical datasets generally include more control than treated units, which is the reason why the PGS is often defined with respect to the control exposure; in contrary case (ie, less control than treated units), "treated" can be relabeled as "control" and vice-versa. Although PGS analysis originally focused on causal inference with binary exposures, we recently discussed how to investigate general treatment regimes.[6] For the sake of simplicity, we here focus on binary treatment exposures.

Let $\psi(\boldsymbol{X}) \equiv E(Y_0|\boldsymbol{X})$ be a PGS such that $Y_0 \perp \boldsymbol{X} \,|\, \psi(\boldsymbol{X})$.[9] Hansen[9] showed that, under the assumption of no "hidden bias" given $\boldsymbol{X}$ (ie, $Y_0 \perp Z \,|\, \boldsymbol{X}$; that is, $\boldsymbol{X}$ captures all confounders), conditioning on $\psi(\boldsymbol{X})$ leads to $Y_0 \perp Z \,|\, \psi(\boldsymbol{X})$. From this, under the rule of consistency (ie, $Y = ZY_1 + (1-Z)Y_0$) and the assumption of "positivity" conditional on the PGS (ie, $0 < \Pr(Z=1|\,\psi(\boldsymbol{X})) < 1$ with certainty), the ATT can be identified as follows. (For mathematical proofs, see Hansen.[9])

$$\text{ATT} = E_{\psi(\boldsymbol{X})|Z=1}\{E(Y|Z=1,\psi(\boldsymbol{X})) - E(Y|Z=0,\psi(\boldsymbol{X}))\}. \tag{1}$$

$E_{\psi(\boldsymbol{X})\,|\,Z=1}$ denotes the total expectation over all $\psi(\boldsymbol{X})$ in the treated population. Thus, conditioning on the PGS via subclassification or matching allows an unbiased estimation of the ATT.[6,9] (Note, this identification of the ATT holds regardless of the presence/absence of effect modifiers, which is not the case for the ATE.[6,9]).

The main advantages of PGS analysis are the following. First, as opposed to the propensity score (defined as $e(\boldsymbol{X}) \equiv \Pr(Z=1|\boldsymbol{X})$), the PGS does not require to predict the treatment allocation, and is therefore particularly useful when exposures are rare or have a large number of categories.[6,19,20] Second, PGS analysis does not require the strong positivity assumption needed in propensity score analysis, $\Pr\{0 < \Pr(Z=1|\boldsymbol{X}) < 1\} = 1$,[18] but instead the relaxed assumption $\Pr\{0 < \Pr(Z=1|\,\psi(\boldsymbol{X})) < 1\} = 1$.[9] Rather than requiring a stochastic treatment allocation across *all covariate values*, PGS analysis only requires a stochastic treatment allocation across all values of the PGS.

2.3 | Practical issues in PGS analysis

In practice, neither the propensity score, $e(\boldsymbol{X})$, nor the PGS, $\psi(\boldsymbol{X})$, is known; they are both estimated using the available nonrandomized data. Whilst the propensity score model, $\widehat{e}(\boldsymbol{X}) = \Pr(Z = 1|\boldsymbol{X}, \widehat{\alpha})$, is fitted to the entire sample (ie, including both treated and control individuals), the PGS model, $\widehat{\psi}(\boldsymbol{X}) = E(Y_0|\boldsymbol{X}, \widehat{\beta})$, is fitted *only to the control arm*, as a regression model fitted to both arms would estimate a confused mixture of $E(Y_0|\boldsymbol{X})$ and $E(Y_1|\boldsymbol{X})$.[9] (Though it may be possible to use both arms by stratifying all unknown PGS parameters across the treatment groups, for instance by means of interaction terms, such an approach is likely to be at risk of model-dependence. For instance, see Groenwold et al[21] or van Klaveren et al.[22])

Two practical issues should be considered in PGS analysis:

(A) After fitting $\widehat{\psi}(\boldsymbol{X})$ to the control arm, $Z = 0$, the estimated PGS model is to be *extrapolated* to the treated arm, $Z = 1$.
(B) Contrary to propensity score analysis where the balance property of $\widehat{e}(\boldsymbol{X})$ can be evaluated in the entire sample, the balance property of $\widehat{\psi}(\boldsymbol{X})$ is assessable only in the control arm, since Y_0 is not observable in the treated arm (ie, "fundamental problem of causal inference").

Practical issue (A) arises from the PGS model being estimated only in the control patients. This implies that the sample size for estimating a PGS model is smaller than for propensity score modelling. The estimated PGS model is then applied to all patients of the study to predict their potential outcome Y_0. Since treated patients do not contribute to the estimation of the prognostic model, it is possible that model predictions do not transport well to the treated arm. This may affect the validity of the subsequent treatment effect estimation.

Practical issue (B) arises from the difficulty to evaluate the accuracy of the PGS model, because Y_0 is not observed in treated patients. As a result, this accuracy can be assessed only in the control patients.

In this sense, (A) is a matter of modelling (ie, to be ensured, the generalizability of $\widehat{\psi}(\boldsymbol{X})$ requires an appropriate modelling strategy); whilst (B) is a matter of diagnostic (ie, to be ensured, the balance property of $\widehat{\psi}(\boldsymbol{X})$ needs proper diagnostic methods). Some authors have proposed the use of a "dry-run" analysis in response to (B)[23]; however, there remains, to our best knowledge, a paucity of the literature addressing (A).

It is well-recognized that a prognostic model derived in one particular sample may not generalize well to other samples from the same (control) or different (treated) populations. Overfitting is likely to happen when the development sample is relatively small (compared to the number of candidate predictors); and transportability issues are likely to occur when the case-mix of the development sample is too narrow.[24-27] In PGS analysis, it is of particular importance that $\widehat{\psi}(\boldsymbol{X})$ properly accounts for covariates that substantially affect prognosis and may have different distributions across the treatment exposure groups. Although $\widehat{\psi}(\boldsymbol{X})$ can be derived in the same cohort containing the treated and untreated individuals, it has been shown through simulations that $\widehat{\psi}(\boldsymbol{X})$ obtained from a separate sample of control individuals should be more efficient.[9] Hereupon, follows the discussion that "these difficulties [related to same-sample estimation] may be substantially avoided if an alternate sample of controls, perhaps historical controls, are available for the determination of the PGS."[9] The use of external, historical scores has therefore gained increased consideration, as they often improve the performance of PGS analysis.[3-5,8]

Unfortunately historical controls are not always easy to obtain, and models derived from historical controls may not necessarily calibrate well across different studies or treatment groups (for instance, due to differences in variable definitions or measurement methods across cohort studies). Furthermore, the use of a single, historical PGS model may not always be straightforward, particularly when multiple competing models exist. Hence, instead of considering one external, historical PGS model at face validity, we introduce the idea of aggregating multiple published PGS models, and tailoring them to the control arm of the nonrandomized treatment study.

3 | PROPOSED METHOD: AGGREGATION OF PUBLISHED PGSS

Debray et al[12] previously proposed aggregating published prediction models to help improve their accuracy and generalizability across different settings and populations. The approach bears some similarities with penalization during prediction model development, as it loosely imposes (in this case, previously estimated) predictive associations to the data at hand. The authors showed through clinical datasets and simulations that the aggregation of multiple models into one "meta-model" tends to outperform redevelopment or updating of a single prognostic model, particularly

when development datasets are relatively small compared to the number of predictors being studied.[12] We focus on an approach called "stacked regressions",[12] that is, a linear combination of published models—that we extend to the PGS analysis.

Let $\widehat{\boldsymbol{\Psi}}$ be the set of M published PGS models, $\widehat{\boldsymbol{\Psi}} = (\widehat{\psi}_1, \widehat{\psi}_2, \dots, \widehat{\psi}_M)$, wherein, using generalized linear models with $g()$ link function:

$$\widehat{\psi}_m(\boldsymbol{X}) = E(Y_0|\boldsymbol{X}, \widehat{\beta}_m) = g^{-1}\left(\widehat{\beta}_{0,m} + \sum_{p=1}^{P} \widehat{\beta}_{p,m} X_p\right).$$

Stacked regressions incorporate each prediction $g(\widehat{\psi}_m)$ as a predictor in a generalized linear regression model that is to be estimated in the control arm of the nonrandomized study sample.

For instance, for samples with binary outcomes, the meta-model is estimated by optimizing the following log-likelihood function in the N control individuals with Z=0:

$$l(\boldsymbol{\vartheta}) = \sum_{i=1}^{N} Y_{(0)i}\left(\theta_0 + \sum_{m=1}^{M} \theta_m g(\widehat{\psi}_m(\boldsymbol{X}))\right) - \log\left(1 + \exp\left(\theta_0 + \sum_{m=1}^{M} \theta_m g(\widehat{\psi}_m(\boldsymbol{X}))\right)\right).$$

In this expression, the constraint $\theta_m \geq 0$ was proposed to allow the omission of literature models $\widehat{\psi}_m(\boldsymbol{X})$ with strong collinearity.[12] Further, $\boldsymbol{X}$ is defined as the set of all covariates across the M published PGS models. The estimation of the parameters leads to the meta-model $\widehat{\Phi}$:

$$\widehat{\Phi}(\boldsymbol{X}) = E(Y_0|\widehat{\boldsymbol{\Psi}}, \widehat{\theta}) = E(Y_0|\boldsymbol{X}, \widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \dots, \widehat{\beta}_M), \widehat{\theta}) = g^{-1}\left\{\widehat{\theta}_0 + \sum_{m=1}^{M} \widehat{\theta}_m\left(\widehat{\beta}_{0,m} + \sum_{p=1}^{P} \widehat{\beta}_{p,m} X_p\right)\right\}.$$

Since $\widehat{\psi}_1, \widehat{\psi}_2, \dots, \widehat{\psi}_M$ originate from different studies and settings, they are likely to include many covariates. This implies that the stacked model, which only involves one unknown parameter for each literature model (rather than for each covariate), allows one to adjust for many confounders even when very few data are at hand, thereby minimizing both the risk of unmeasured confounders and that of overfitting.

The PGS model derived from the meta-model, $\widehat{\Phi}(\boldsymbol{X})$, is then used for conditioning (eg, matching or subclassification). Replacing $\psi(\boldsymbol{X})$ by $\widehat{\Phi}(\boldsymbol{X})$ in Equation (1), one can estimate the ATT as:

$$\widehat{\text{ATT}} = E_{\widehat{\Phi}(\boldsymbol{X})|Z=1}\{E(Y|Z = 0, \widehat{\Phi}(\boldsymbol{X})) - E(Y|Z = 0, \widehat{\Phi}(\boldsymbol{X}))\}.$$

The next section presents an illustrative case in which we estimate the ATT, using a "meta-score", external (ie, previously published) PGS, or same-sample PGS.

4 | ILLUSTRATIVE CASE STUDY

4.1 | Methods

To illustrate our approach, we performed a subgroup analysis of the ACCURATE trial.[28] (The data that support the findings of the current study are available from the investigators of this trial upon reasonable request.)

Briefly, this pragmatic trial investigated three asthma interventions to reduce the 1-year risk of exacerbation in asthmatic patients (see Data S1, Section 1.1). The severity of clinical manifestations of asthma was classified into controlled asthma (Ca), partly Ca (PCa), and uncontrolled asthma categories. Here, we defined Ca as an Asthma Control Questionnaire (ACQ) score equal or below 0.75, and PCa as $0.75 < \text{ACQ} \leq 1.50$.[28]

We considered PGS analysis in the ACCURATE trial to estimate the comparative efficacy of PCa ("treatment") vs Ca ("control exposure") in a subgroup of female asthmatic patients. A naïve estimation of the causal effect would likely suffer from confounding bias given the cluster-randomization,[29,30] subgroup analysis,[31,32] and exclusion of some randomized patients (see Data S1, Section 1.1).

We conducted three strategies of PGS analysis. First, we derived three same-sample (logistic) PGS models using maximum likelihood (ML) estimation, ridge regression, and LASSO regression. These scores were developed from the "control" female patients allocated to Ca, and included seven covariates: ACQ score; current smoking status (yes/no); chronic sinusitis complaints (yes/no); hospitalization for asthma ever (yes/no); steroid courses for asthma within last year (yes/no); predicted prebronchodilator forced expiratory volume in 1 second (FEV1); fraction of exhaled nitric oxide. (We chose these covariates as they had been previously identified as prognostic variables,[33] and thus, potential confounders.)

Second, we used three external, published asthma prediction models as PGSs, which predicted different but related outcomes (Data S1, Section 1.2). The model reported by Schatz et al[34] predicted the risk of asthma hospitalization at 1 year, given three covariates: hospitalization for asthma ever (yes/no); steroid courses for asthma within last year (yes/no); income. The model presented by Eisner et al. estimated the risk of unscheduled care visit at 1 year, given two covariates: severity of asthma score and asthma control test.[35] The TENOR model, reported by Miller et al,[36] predicted the risk of

TABLE 1 Covariates and outcomes considered in different PGSs

	Same-sample score	Schatz score	Eisner score	TENOR score	Meta-score
Covariates	*7*	*3*	*2*	*10*	*15*
ACQ score	X				
Currently smoking	X				
Chronic sinusitis complaint	X				
Hospitalization for asthma	X	X			X
Steroid course for asthma within last year	X	X			X
Predicted FEV1	X				
Fraction of exhaled nitric oxide	X				
Income		X			X
Severity of asthma score			X		X
Asthma control test			X		X
Age				X	X
Sex				X	X
Race				X	X
Body mass index				X	X
Predicted FVC				X	X
History of pneumonia				X	X
Diabetes				X	X
Cataracts				X	X
Intubation for asthma				X	X
Steroid bursts in the prior 3 months				X	X
Outcomes					
Exacerbation at 1 year	X				X
Hospitalization at 1 year		X			
Unscheduled care visit at 1 year			X		
Emergency department visit or hospitalization at 6 months				X	

Abbreviations: ACQ, Asthma Control Questionnaire; FEV1, prebronchodilator forced expiratory volume in 1 s; FVC; postbronchodilator forced capacity volume; PGS, prognostic score.

emergency department visit or hospitalization at 6 months, given 10 covariates: younger age; female sex; nonwhite race; body mass index > or = 35 kg m^{-2}; postbronchodilator percent predicted forced vital capacity <70%; history of pneumonia; diabetes; cataracts; intubation for asthma; and three or more steroid bursts in the prior 3 months.

Finally, using the proposed method of stacked regressions, we aggregated the three external scores (Schatz score, Eisner score, and TENOR score) into a "meta-score" in the "control" arm of study data, and compared this meta-score with the same-sample PGS.

All covariate sets and outcomes considered in each of the PGS are reported in Table 1.

We performed a full matching analysis on the PGSs to estimate the ATT.[6] To address the issue of missing data (which concerned only the same-sample PGSs), we used multiple imputation as described by Leyrat et al[37] in propensity score analysis, who recommend to apply Rubin's rules on the estimated treatment effects across the imputed datasets, and not over the scores. We therefore created 10 imputed datasets. In each imputed dataset, we fit three PGSs (based on ML, ridge, and LASSO) to the control arm, matched on each on these three scores (separately), and estimated the treatment effect. We then averaged the treatment effect estimates obtained from the 10 imputed datasets.

We computed standard errors of treatment effect estimates (and 95% confidence intervals [CI]) by bootstrapping (1000 iterations): we resampled the initial dataset with replacement and performed in each bootstrap sample all aforementioned analyses, including multiple imputations.[38] We estimated the standard errors as the standard deviations of the estimates across the 1000 iterations.

4.2 | Results

In the subgroup of 281 female patients, the active "treatment" (ie, strategy aiming at PCa) was allocated to 134 individuals (47.7%). Table 2 summarizes baseline characteristics and predicted outcome risks across the two groups. Twenty-three (15.6%) and 21 (15.7%) patients experienced an exacerbation at 1 year in the "control" and in the "treatment" group, respectively. Thus, a naïve approach ignoring the potential for confounding bias would estimate the treatment effect to be equal to 0.001, on an absolute risk difference scale. Note that given the few outcomes observed in the control arm, the same-sample PGSs—in particular, the one estimated by ML—may have suffered more from overfitting issues (ratio of events per variable [EPV] = 3.3) than did the "meta-score" (EPV = 7.7).[39,40]

After full matching on the same-sample PGSs derived by ML, ridge, and LASSO regressions, the ATT was estimated to be equal to 0.021 (95% CI: −0.054 to 0.097), 0.024 (95% CI: −0.052 to 0.099), and 0.022 (95% CI: −0.052 to 0.096),

TABLE 2 Baseline characteristics and predicted outcome risks in the subgroup of female patients (N = 281)

	Ca	**PCa**
	n = 147 (42.7%)	**n = 134 (37.7%)**
ACQ score	1.22 (SD: 1)	1.02 (SD: 0.96)
Currently smoking	21 (14.5%)	22 (16.9%)
Chronic sinusitis complaint	20 (13.7%)	17 (13.2%)
Hospitalization for asthma	14 (9.5%)	15 (11.2%)
Steroid course for asthma within last year	39 (26.5%)	27 (20.1%)
Predicted FEV1 (%)	92.29 (SD: 14.33)	92.35 (SD: 16.75)
Fraction of exhaled nitric oxide	26.40 (SD: 30.96)	21.83 (SD: 21.18)
Schatz score (logit scale)	−3.32 (SD: 0.61)	−3.30 (SD: 0.56)
Eisner score (logit scale)	−1.06 (SD: 0.48)	−1.11 (SD: 0.45)
TENOR score (logit scale)	−3.75 (SD: 0.87)	−3.81 (SD: 0.89)
Exacerbation at 1 year	23 (15.6%)	21 (15.7%)

Note: Mean and SD are reported for continuous variables; frequency and percentage for binary variables.
Abbreviations: ACQ, Asthma Control Questionnaire; Ca, aiming at controlled asthma; FEV1, prebronchodilator forced expiratory volume in 1 s; PCa, aiming at partially controlled asthma.

respectively. After full matching on the (historical) Schatz, Eisner, and TENOR scores, the estimated ATT was equal to 0.018 (95% CI: −0.069 to 0.105), 0.027 (95% CI: −0.054 to 0.108), and − 0.002 (95% CI: −0.088 to 0.085), respectively. After full matching on the meta-score (ie, aggregation and simultaneous updating of the Schatz, Eisner, and TENOR scores), the estimated ATT was equal to 0.019 (95% CI, −0.060; 0.098). In summary, the evaluated approaches yielded an absolute treatment effect ranging from 0 to 3%, with considerably large CI.

Given these differences in estimated treatment effects, we conducted two simulation studies to better inform the performance of each method.

5 | SIMULATION STUDIES

5.1 | Simulation study 1

5.1.1 | Data generation

We conducted a series of simulations in which we estimated the ATT of a (fictive) treatment Z. We considered nine covariates $\boldsymbol{X} = (X_1, \dots, X_9)$ and a binary outcome Y (see Box 1). To generate data with realistic distributions and correlation structures, we took the ACCURATE study as a reference,[28] and generated new, simulated, datasets by random sampling from its posterior distribution (Data S1, Sections 1.3 and 1.4).

Briefly, we estimated the conditional distributions of Y_0, X_1, …, X_9 in the ACCURATE study. The conditional distributions were then combined into a Markov chain where Monte Carlo sampling (using 1000 iterations) was used to generate new individuals that resemble those observed in the ACCURATE cohort (see Box 1).

We defined four populations A, B, C, and D with different patient characteristics and focusing on different but related outcomes.

Population D (and its corresponding outcome definition) was of primary interest for estimation of the ATT of treatment Z such that $Y^D_{0,k} = Y_{0,k}$. Samples of population D were, however, small and restricted to $N^D_k = 100$.

Populations A, B, and C represented external historical cohorts, from which previous prognostic models could have been derived and published in the literature. These external cohorts had larger sample sizes ($N^{(A,B,C)}_k \geq 500$) but adopted outcome definitions that did not necessarily correspond to the outcome of primary interest ($Y^A_{0,k}$, $Y^B_{0,k}$, $Y^C_{0,k} \neq Y_{0,k}$). Furthermore, since in clinical practice inclusion criteria are likely to differ across studies, we defined bounds (ie, truncation) to impose slightly different covariate distributions across the four populations.

In samples of population D, we generated the potential outcome $Y_{1,k}$ of primary interest that would be observed were individuals to receive the fictive treatment (see Box 1). We defined $Y_{1,k}$ using a nonlinear nonmonotonic function of $Y_{0,k}$ (see Box 1). In samples of population D, we generated the treatment status Z_k according to a logistic model mimicking a clinical decision (ie, true propensity model; see Box 1). This model set the treatment prevalence at 35 to 40%. Finally, we generated the observed outcome Y_k using the rule of consistency.

BOX 1 Definition of variables used in simulations

$\boldsymbol{X} = (X_1, \dots, X_9)$ and Y_0 are generated in samples of populations (A, B, C, and D), using a Gibbs sampler approximating joint distributions observed in the ACCURATE cohort[28]:

X_1: age (continuous, years).

X_2: sex (binary, 0 = male, 1 = female).

X_3: baseline ACQ score (continuous, ranges from 0 to 6).

X_4: current smoking status (binary, 0 = no, 1 = yes).

X_5: chronic sinusitis complaints (binary, 0 = no, 1 = yes).

X_6: hospitalization for asthma ever (binary, 0 = no, 1 = yes).

X_7: steroid courses for asthma within last year (binary, 0 = no, 1 = yes).

X_8: predicted FEV1 (continuous, %).

X_9: fraction of exhaled nitric oxide (continuous, ppb).

Y_0: occurrence of exacerbation at 1 year following the definition of the American Thoracic Society (binary, 0 = no, 1 = yes); potential outcome of primary interest under $Z = 0$.

Y_1: binary potential outcome of primary interest under $Z = 1$, generated in population D from a distribution Bernoulli($p(Y_1)$), where $p(Y_1) = \frac{1}{1+\exp\left\{3-0.7\log\left(\frac{\sin(Y_0+u+0.4)}{1-\sin(Y_0+u+0.4)}\right)\right\}}$ and u is a random number drawn from a Uniform distribution $U(0, 1)$.

Z: fictive treatment (binary, 0 = control, 1 = treatment), generated in population D from a distribution Bernoulli($e(\boldsymbol{X})$), where $e(\boldsymbol{X}) = \frac{1}{1+\exp\{0.25-0.05\log(X_1)-0.25X_2+0.25X_3+0.5X_4+0.75X_5+X_6+1.25X_7\}}$.

Y: binary observed outcome of primary interest, generated in population D using the rule of consistency: $Y = ZY_1 + (1-Z)Y_0$

Y_0^A: binary potential outcome of secondary interest under $Z = 0$, generated in population A from a distribution Bernoulli($p(Y_0^A)$), where $p(Y_0^A) = \frac{1}{1+\exp\left\{2-0.9\log\left(\frac{\sin\left(0.2+\frac{Y_0+u}{2}\right)}{1-\sin\left(0.2+\frac{Y_0+u}{2}\right)}\right)\right\}}$ and u is a random number drawn from a Uniform distribution $U(0, 1)$.

Y_0^B: binary potential outcome of secondary interest under $Z = 0$, generated in population B from a distribution Bernoulli($p(Y_0^B)$), where $p(Y_0^B) = \frac{1}{1+\exp\left\{1.5-1.5\log\left(\frac{\sin\left(0.15+\frac{Y_0+u}{2}\right)}{1-\sin\left(0.15+\frac{Y_0+u}{2}\right)}\right)\right\}}$ and u is a random number drawn from a Uniform distribution $U(0, 1)$.

Y_0^C: binary potential outcome of secondary interest under $Z = 0$, generated in population C from a distribution Bernoulli($p(Y_0^C)$), where $p(Y_0^C) = \frac{1}{1+\exp\left\{2-2\log\left(\frac{\sin\left(0.2+\frac{Y_0+u}{2}\right)}{1-\sin\left(0.2+\frac{Y_0+u}{2}\right)}\right)\right\}}$ and u is a random number drawn from a Uniform distribution $U(0, 1)$.

For each population (A, B, C, and D), we used this procedure to generate $K = 5000$ samples. Across all $K = 5000$ simulated samples of population D, the true ATT corresponded, on an absolute risk difference scale, to the empirical expectation:

$$\frac{1}{K}\sum_{k=1}^{K}\left\{\frac{\sum_{i=1}^{N_k^D}(Y_{(1)i,k} - Y_{(0)i,k})I(Z_{i,k} = 1)}{\sum_{i=1}^{N_k^D} I(Z_{i,k} = 1)}\right\} = 0.087.$$

5.1.2 | Statistical analysis

In each simulation k, we compared three strategies of PGS analysis for estimating the ATT. First, we fitted a logistic regression model to the control arm of sample D_k to predict the potential outcome $Y_{0,k}$ in this same sample (ie, same-sample PGS). We included all covariates ($X_{1,k}$ to $X_{9,k}$) into the regression model, such that no hidden bias was introduced (ie, "utopian" case). Second, we used the score derived from sample A_k and applied it as a PGS in sample D_k (ie, external PGS). Finally, we performed the proposed method of stacked regressions to aggregate external PGSs obtained from samples A_k, B_k, and C_k into a meta-score which predicts $Y_{0,k}$ in sample D_k (ie, aggregated external PGSs). To assess the robustness of the use of external PGSs (either individually or after aggregation), we considered different scenarios (see Box 2).

BOX 2 Definition of scenarios used in simulations

Scenario 1: All historical models predict Y_0 and include covariates $\boldsymbol{X} = (X_1, \dots, X_9)$.

Scenario 2: All historical models predict Y_0. Historical model A includes covariates $(X_1, X_2, X_3, X_4, X_5, X_6)$, historical model B includes covariates $(X_1, X_2, X_3, X_7, X_8, X_9)$, and historical model C includes covariates $(X_4, X_5, X_6, X_7, X_8, X_9)$.

Scenario 3: Historical model A predicts Y_0^A, historical model B predicts Y_0^B, and historical model C predicts Y_0^C. All historical models include covariates $\boldsymbol{X} = (X_1, \dots, X_9)$.

Scenario 4: Historical model A predicts Y_0^A and includes $(X_1, X_2, X_3, X_4, X_5, X_6)$, historical model B predicts Y_0^B and includes $(X_1, X_2, X_3, X_7, X_8, X_9)$, and historical model C predicts Y_0^C and includes covariates $(X_4, X_5, X_6, X_7, X_8, X_9)$.

Regardless of whether the PGSs were externally or same-sample derived, we considered three approaches to regression: ML, ridge, and LASSO regressions.

To evaluate the impact of the sample size of historical cohorts from which external scores were derived, we defined samples of populations A, B, and C as including either $N_k^{(A,B,C)} = 500$, 1000, or 2000 subjects.

Therefore, at each iteration $k \in (1, 2, \dots, K = 5000)$ we had: three same-sample PGSs derived from D_k; 36 external PGSs derived from sample A_k; 36 meta-PGSs aggregating models derived from samples A_k, B_k, and C_k.

We performed full matching analysis on the PGSs.[6] As opposed to classical matching methods, full matching allows the entire sample to be preserved.[41,42] It creates strata within which control and treated units are paired in an optimal way such that the overall distance between them is minimized. Since each stratum (or "pair") can include unequal numbers of treated and controls, individuals are to be weighted.[41,42] Let S_i denotes the stratum in which individual i falls; to estimate the ATT, the individual weight W_i^{ATT} is:

$$W_i^{\text{ATT}}(S_i = s) = Z_i + (1 - Z_i)\left\{\frac{\sum_i I(S_i = s)I(Z_i = 1)}{\sum_i I(Z_i = 1)}\right\}\left\{\frac{\sum_i I(Z_i = 0)}{\sum_i I(S_i = s)I(Z_i = 0)}\right\}.$$

We estimated the ATT as the weighted average difference in outcome across the two treatment arms. We compared the performance of the three PGS analyses, by measuring the bias and mean-squared error (MSE) of estimates.

5.1.3 | Results

Across all $K = 5000$ simulations, the mean treatment prevalence in population D was equal to 37.1%. The mean outcome prevalence was equal to 15.3% in control individuals and 16.9% in treated individuals, thereby leading to a mean naïve absolute risk difference of 0.016 (empirical SD: 0.077). In populations A, B, and C (ie, historical control cohorts), the mean outcome prevalence was equal to 10.9%, 12.3%, and 10.6%, respectively.

Using the different methods of PGS, the performances (bias, SD, and MSE) of ATT estimates across all scenarios were reported in Table 3.

TABLE 3 Performance of different PGS analyses

PGS	Outcome of external score	Covariates of external score	Modelling method	$N_k^{(A,B,C)} = 500$		$N_k^{(A,B,C)} = 1000$		$N_k^{(A,B,C)} = 2000$	
				Relative bias	100× MSE	Relative bias	100× MSE	Relative bias	100× MSE
Same-sample PGS			ML	−14.9%	0.6509	−14.9%	0.6509	−14.9%	0.6509
			Ridge	−2.3%	0.6070	−2.3%	0.6070	−2.3%	0.6070
			LASSO	−14.9%	0.6692	−14.9%	0.6692	−14.9%	0.6692
External PGS	Identical	Identical	ML	−13.8%	0.6235	−10.3%	0.6043	−8.0%	0.5987
			Ridge	−11.5%	0.6247	−8.0%	0.5928	−6.9%	0.5920
			LASSO	−12.6%	0.6197	−10.3%	0.5956	−8.0%	0.6047
		Different	ML	−41.4%	0.8078	−39.1%	0.7728	−39.1%	0.7769
			Ridge	−41.4%	0.8050	−39.1%	0.7820	−37.9%	0.7640
			LASSO	−42.5%	0.8040	−39.1%	0.7801	−39.1%	0.7686
	Different	Identical	ML	−48.3%	0.9012	−39.1%	0.8083	−26.4%	0.7057
			Ridge	−43.7%	0.8815	−33.3%	0.7707	−21.8%	0.6958
			LASSO	−64.4%	1.0745	−52.9%	0.9468	−33.3%	0.7664
		Different	ML	−62.1%	1.0218	−58.6%	0.9860	−51.7%	0.9065
			Ridge	−63.2%	1.0452	−58.6%	0.9799	−51.7%	0.9055
			LASSO	−74.7%	1.1631	−70.1%	1.1072	−60.9%	0.9829
Aggregated external PGSs	Identical	Identical	ML	−6.9%	0.5975	−5.7%	0.5851	−5.7%	0.5729
			Ridge	−3.4%	0.5759	−3.4%	0.5861	−3.4%	0.5675
			LASSO	−4.6%	0.5775	−4.6%	0.5911	−4.6%	0.5713
		Different	ML	−5.7%	0.5898	−4.6%	0.5767	−3.4%	0.5869
			Ridge	−4.6%	0.5859	−3.4%	0.5848	−3.4%	0.5805
			LASSO	−5.7%	0.5931	−3.4%	0.5706	−3.4%	0.5835
	Different	Identical	ML	−13.8%	0.6337	−9.2%	0.6146	−5.7%	0.5956
			Ridge	−9.2%	0.6235	−4.6%	0.5854	−2.3%	0.5884
			LASSO	−17.2%	0.6644	−9.2%	0.6118	−5.7%	0.6020
		Different	ML	−12.6%	0.6308	−9.2%	0.6111	−5.7%	0.6024
			Ridge	−11.5%	0.6228	−8.0%	0.6068	−5.7%	0.6003
			LASSO	−19.5%	0.6585	−11.5%	0.6202	−8.0%	0.6202

Abbreviations: LASSO, least absolute shrinkage and selection operator; ML, maximum likelihood; MSE, mean-squared error; $N_k^{(A,B,C)}$, sample size for external prognostic score derivation; PGS, prognostic score.

Though it was considered under "utopian" conditions (ie, no misspecification nor hidden bias), we found that the use of same-sample PGSs could lead to a biased estimation of the ATT, in particular when the PGS was derived by ML or LASSO regression (Table 3). This bias may have arisen from overfitting when using ML, whilst it might be due to residual confounding given variable selection when using LASSO regression. In comparison, same-sample PGSs derived by ridge regression performed better (Table 3).

The use of external PGSs (derived from samples of population A) that included all covariates and predicted the outcome of primary interest yielded a performance comparable to that of same-sample PGSs, when the sample size for external PGS derivation was relatively small, $N_k^A = 500$ (Figure 1). This performance in estimating the ATT improved when the size of samples A_k was larger, $N_k^A \in \{1000, 2000\}$ (Figures 2 and 3). We showed that external PGSs including a limited set of covariates—thus, potentially omitting important confounders—resulted in biased estimates, regardless of

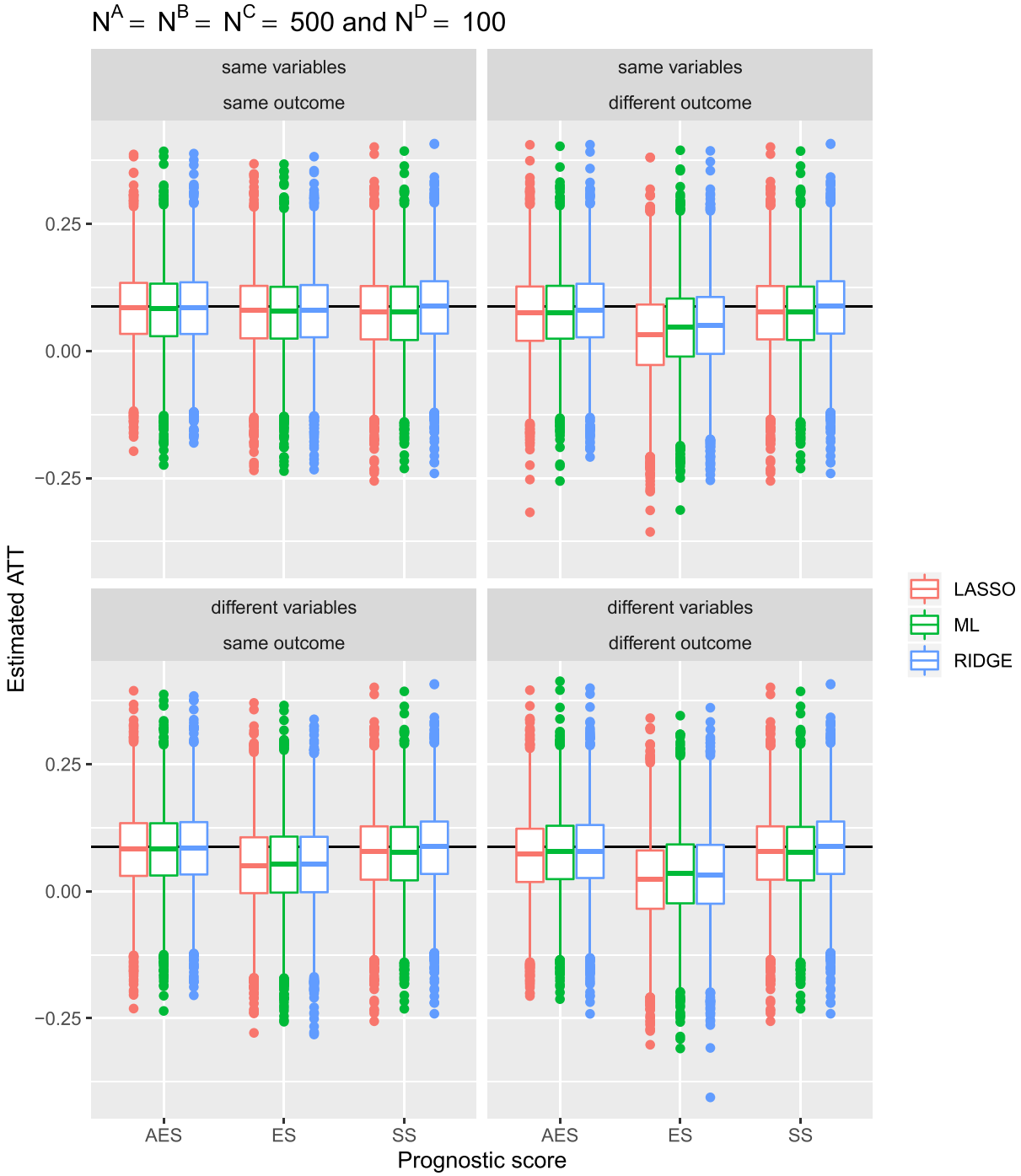

FIGURE 1 Performance of different approaches to PGS analysis (sample size for external score derivation, $N_k^{(A,B,C)} = 500$). The dotted line refers to the "true" ATT. AES, aggregated external prognostic scores; ES, external prognostic scores; LASSO, least absolute shrinkage and selection operator; ML, maximum likelihood; PGS, prognostic score; SS, same-sample prognostic scores [Colour figure can be viewed at wileyonlinelibrary.com]

whether the outcome of sample A_k was identical to the primary outcome of sample D_k (Figures 1-3). In addition, considering a different but related secondary outcome (Y_0^A instead of Y_0) also led to bias, even when no confounder was missing in the external PGSs (Figures 1-3). The use of external PGSs derived by ridge regression generally outperformed those obtained by ML and LASSO regression across the different scenarios (Figures 1-3).

As shown in Figures 1-3, the use of aggregated external PGSs after stacked regressions led to unbiased estimates of the ATT. This result appeared consistent across all (unfavorable) scenarios. When all three external PGSs (derived from samples of populations A, B, and C) included all covariates and predicted the outcome of primary interest, the

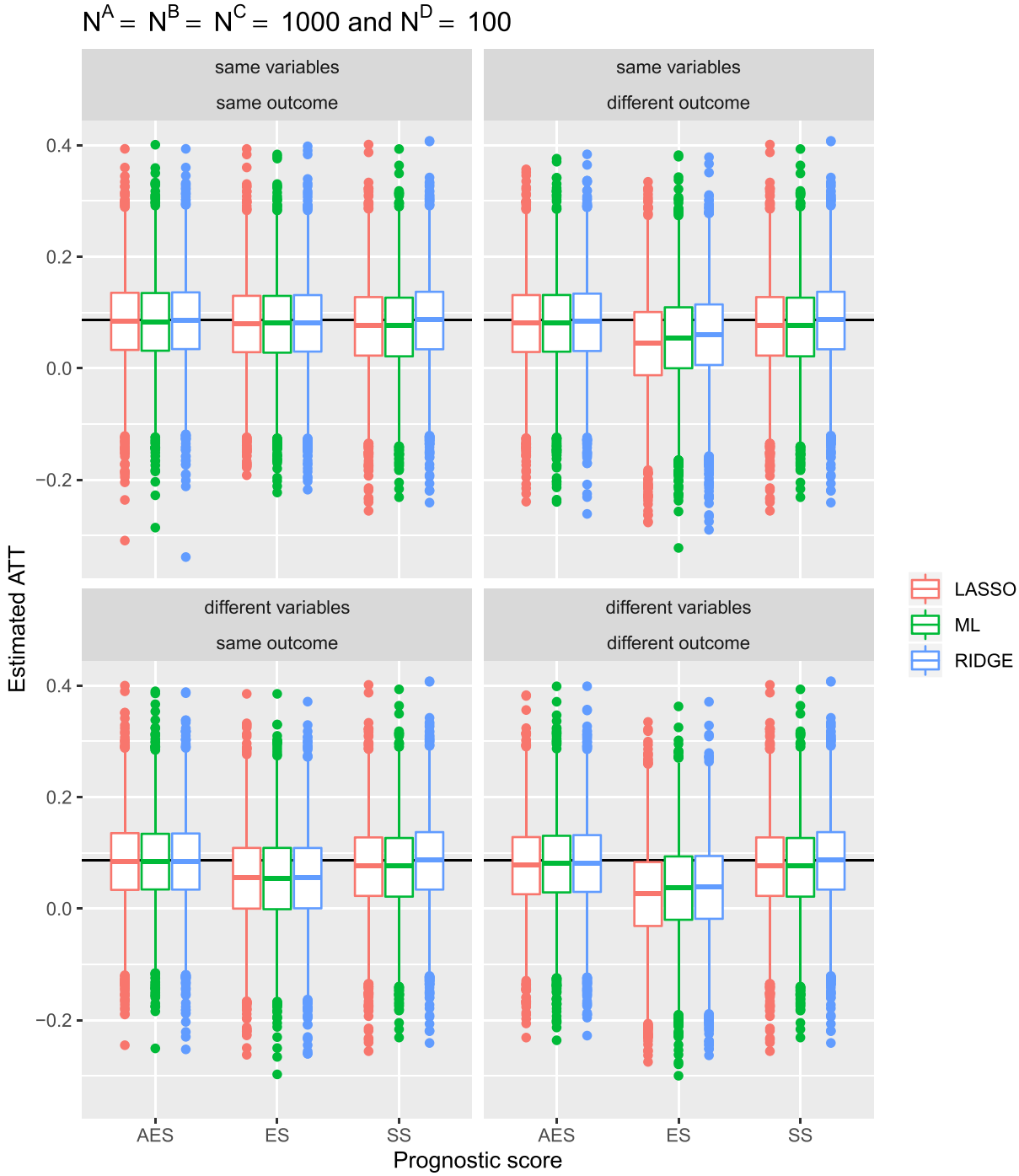

FIGURE 2 Performance of different approaches to PGS analysis (sample size for external score derivation, $N_k^{(A,B,C)} = 1000$). The dotted line refers to the "true" ATT. AES, aggregated external prognostic scores; ES, external prognostic scores; LASSO, least absolute shrinkage and selection operator; ML, maximum likelihood; PGS, prognostic score; SS, same-sample prognostic scores [Colour figure can be viewed at wileyonlinelibrary.com]

use of a meta-PGS outperformed all other methods in terms of bias, variance, and mean square error (Table 3). We observed that the aggregation of external PGSs derived by ridge regression led to the best performance (Table 3). In the most unfavorable scenario, in which the three external PGSs included different covariate sets and predicted secondary outcomes (Y_0^A, Y_0^B, and Y_0^C instead of Y_0), the aggregation of PGSs still allowed an unbiased estimation of the ATT, especially when the historical cohorts were large (Figures 2 and 3). Even in this worst-case scenario, the aggregation of external PGSs yielded a performance comparable to that of "utopian" same-sample PGSs derived by ridge regression (Table 3).

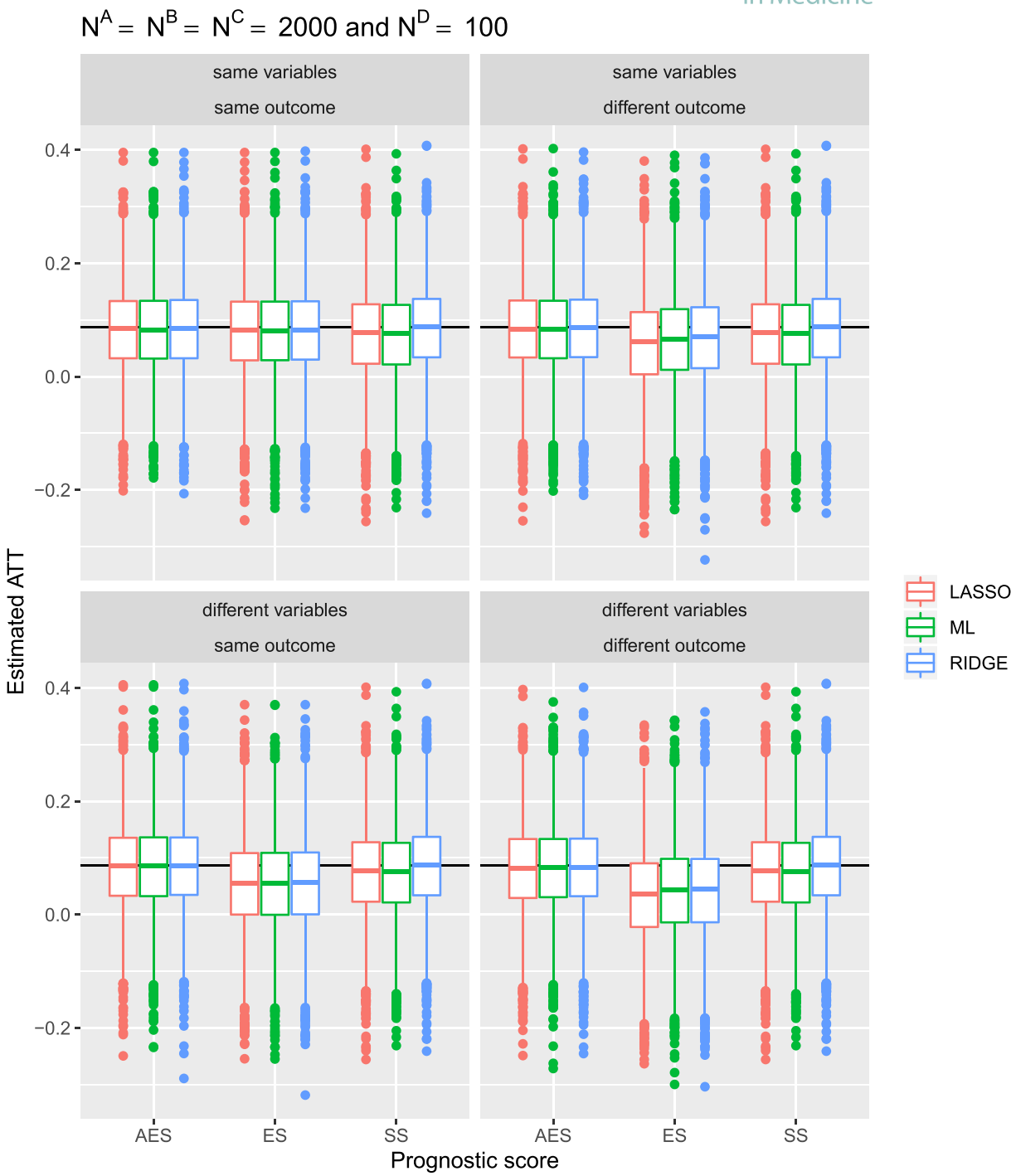

FIGURE 3 Performance of different approaches to PGS analysis (sample size for external score derivation, $N_k^{(A,B,C)} = 2000$). The dotted line refers to the "true" ATT. AES, aggregated external prognostic scores; ES, external prognostic scores; LASSO, least absolute shrinkage and selection operator; ML, maximum likelihood; PGS, prognostic score; SS, same-sample prognostic scores [Colour figure can be viewed at wileyonlinelibrary.com]

Additional results from simulations on larger sample sizes ($N_k^D = 1000$) are reported in Data S1, Section 1.5. Results were more modest, but comparable to those described above.

5.2 | Simulation study 2

We conducted a second series of simulations which favored the traditional same-sample PGS over stacked regressions of external PGSs (for details, see Data S1, Section 2). We considered a binary outcome and 20 independent covariates (16 of

these affected the outcome). Further, we defined the treatment under investigation as ineffective, that is, null treatment effect.

We considered three scenarios to evaluate the performance of stacked regressions. For all scenarios, we generated five historical cohorts with 50 events (ie, rare outcome).

- Scenario A: The historical cohorts varied systematically with respect to their baseline risk. More specifically, there was between-study heterogeneity in the intercept term of the data generation model from each historical cohort.
- Scenario B: The historical cohorts varied systematically with respect to their baseline risk and the overall magnitude of their predictor effects. More specifically, there was between-study heterogeneity in the intercept term and the overall slope of the data generation model from each historical cohort.
- Scenario C: The historical cohorts varied systematically with respect to their baseline risk and their individual predictor effects. More specifically, there was between-study heterogeneity in the intercept term and the coefficients of the data generation model from each historical cohort.

The historical cohorts were used to derive the "previously published" PGSs following a backward selection procedure (ie, risk of omitting important confounders). Finally, we generated a nonrandomized study where the covariate distribution was unbalanced between treated and control patients (ie, confounding). We used this sample to estimate the ATT.

Results demonstrated that stacked regressions helped reduce bias and stabilize treatment effect estimates (see Data S1, Section 2.3). Stacked regressions outperformed traditional PGS analysis even when historical cohorts differed with respect to baseline risk and relative strength of predictor effects (scenarios A and B). When historical cohorts were less related to the study population (scenario C), the aggregation of published PGSs showed a benefit over the traditional estimator only when the study sample at hand was relatively small.

6 | DISCUSSION

We have proposed and assessed an extension of stacked regressions to PGS analysis, when multiple historical scores are available. We demonstrated through extensive simulations that this approach facilitated an unbiased estimation of causal treatment effects in nonrandomized studies, even if the historical PGSs omitted important covariates and/or predicted different (but related) outcomes.

Since the treatment assignment mechanism in observational studies is often based on clinical decisions (which are likely to vary across practitioners), to have an exhaustive set of confounders at disposal would suppose (i) knowing, comprehensively, the underlying mechanism of treatment assignment and (ii) recording and accessing to all recognized confounders—a scenario which could be deemed "utopian". A major advantage of PGS analysis (over propensity score analysis) is that it no longer requires to predict the treatment allocation and adopts a less restrictive positivity assumption. However, PGS analysis (propensity score analysis alike) assumes the absence of unmeasured confounding ("hidden bias"), and therefore requires to adjust for as many (possible) confounders as possible.[6,9] This may become problematic when relatively few patient-level data are available.

Although we show that ridge regression may improve the estimation of PGSs in small samples, further improvement is possible by considering evidence synthesis strategies. In particular, aggregation of multiple published PGSs allows to account for a wide set of confounding covariates (at a limited expense in terms of required degrees of freedom), and thus diminishes the potential for hidden bias in nonrandomized studies. Across our simulations, we show that aggregation outperforms traditional penalization methods and enables an unbiased treatment effect estimation, even when the PGSs include different covariates or predict different outcomes from that being investigated. Aggregation is most advantageous when the sample available for treatment effect estimation is small, likely because same-sample PGS is prone to overfitting (when using ML regression) or residual confounding (when selecting covariates by LASSO regression).

Interestingly, prognostic research is increasing, and prediction models are becoming abundant in the medical literature.[27,43] Nowadays, clinical observational datasets often include prognostic models aiming at predicting patient outcomes that are used in daily practice routine. For instance, in intensive care units some prognostic models (eg, simplified acute physiology score II[44]) are routinely recorded (therefore available in observational datasets), whilst the individuals model variables are not necessarily reported (eg, natremia level or patient temperature).

Stacked regressions require no raw (individual participant) data from the historical cohorts; they only need access to the previously published models. These may be available as part of the study results and be presented as estimated regression coefficients, or (adjusted) relative risks or odds ratios. It is also possible to combine literature models, which are presented as score charts or risk scores (eg, TENOR risk score in our illustrative example), or which have been implemented in a web app. To critically appraise previously published prognostic models that may be relevant for aggregation, we recommend the use of PROBAST and CHARMS tools.[45-47] Both were developed to help researchers assess the risk of bias and applicability of published prognostic models, and thus infer whether their inclusion would be beneficial. If considered at low risk of bias and applicable to the question at hand, published prediction models can be externally validated in the available data. Because stacked regressions involve a default update of the common intercept and slope, the inclusion of miscalibrated models may not necessarily be problematic.

An important result from our simulation study is that using historical PGSs without considering any form of updating might be treacherous and lead to substantial bias. This practice has, however, often been recommended in the literature where it is common to assume that historical PGSs developed from very large samples ($N \geq 10\,000$) remain perfectly valid when applied to new patients.[3-5,8] It is, however, likely that historical scores are derived in different (but related) settings, under suboptimal conditions, which may include different covariate sets, consider different outcomes, or adopt different variable definitions. As a result, a single external PGS can be miscalibrated in the nonrandomized data at hand and result in biased treatment effect estimates. We therefore recommend to update published PGSs before their implementation, and propose to borrow strength across multiple PGSs by combining and updating them in the control arm of the study sample through stacked regressions. Recent extensions of stacked regressions allow to recalibrate multiple prediction models while concurrently revising specific covariates and adopting a penalized likelihood.[48]

As raised by an anonymous reviewer, stacked regressions bear some similarities with Super Learning.[49-51] Both approaches create a weighted sum of candidate models (possibly derived using various modelling methods, such as generalized linear regression, neural networks, or random forests). However, rather than combining multiple models that are developed in a single dataset, we propose to combine published models previously developed in different (and possibly larger) datasets. As a result, the candidate models for stacked regressions may exhibit more diversity and provide new information that cannot be derived from the data at hand. Because these candidate models are more likely to be miscalibrated, stacked regressions simultaneously update their predictions to the study population and estimate their optimal combination. In a way akin to Super Learning, the performance of stacked regressions could further be improved by expanding the set of candidate models with additional models derived from the nonrandomized data at hand.

Our study has to be considered in light of some limitations. First, we restricted our simulations to scenarios of small to medium sample sizes, and did not explore the benefit of aggregating PGSs over direct application of external PGSs derived from very large settings. The major interest of the latter resides in addressing the overfitting issues to which same-sample approaches may be prone. Given that aggregating multiple PGSs enables to gather a large amount of information from the medical literature at the expense of but a very few parameters to be estimated, we hypothesize that integrating PGSs originating from large historical cohorts into meta-scores may all the more contribute to the robustness of our approach. Second, we did not investigate the scenario in which meta-scores did not cover all true confounders. If the analyst recognizes that confounders are left aside, one strategy to adopt may be to include those confounders into the meta-score, along with the external PGSs. Finally, methods for estimating standard errors of treatment effect estimates obtained from our approach are needed; one of which might be the adoption of bootstrapping as we conducted in our illustrative study. We did not investigate this in the current simulation study as it requires substantial computational power.

In conclusion, this article proposes a method for aggregating PGSs for causal inference. Through extensive simulations, we demonstrate the robustness of this approach, in particular, to misspecification. We show and discuss how our method could, interestingly, limit both the bias due to unmeasured confounders and that related to overfitting in PGS analysis.

ACKNOWLEDGEMENTS

We thank the two anonymous reviewers for their comments and suggestions, which helped us improve our work. We gratefully acknowledge the ACCURATE study group and Rik Loijmans in particular for sharing of participant level data to illustrate and evaluate the proposed statistical methods. T.P.A.D. would like to acknowledge the scientific exchange of ideas and discussions with Biogen researchers which occurred while serving as a Biogen consultant. That collaboration resulted in part of the theoretical discussion presented in this article. T.P.A.D. was supported by the Netherlands Organization for Health Research and Development (91617050 and 91215058).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of the current study are available from the investigators of this trial upon reasonable request.

ORCID

Tri-Long Nguyen https://orcid.org/0000-0002-6376-7212
Gary S. Collins https://orcid.org/0000-0002-2772-2316
Thomas P.A. Debray https://orcid.org/0000-0002-1790-2719

REFERENCES

1. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ*. 1996;312(7040):1215-1218.
2. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;6:688-701.
3. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiol Drug Saf*. 2012;21(suppl 2):138-147.
4. Kumamaru H, Gagne JJ, Glynn RJ, Setoguchi S, Schneeweiss S. Comparison of high-dimensional confounder summary scores in comparative studies of newly marketed medications. *J Clin Epidemiol*. 2016;76:200-208.
5. Kumamaru H, Schneeweiss S, Glynn RJ, Setoguchi S, Gagne JJ. Dimension reduction and shrinkage methods for high dimensional disease risk scores in historical data. *Emerg Themes Epidemiol*. 2016;13:5.
6. Nguyen T-L, Debray TPA. The use of prognostic scores for causal inference with general treatment regimes. *Stat Med*. 2019;38(11):2013-2029.
7. Tadrous M, Gagne JJ, Sturmer T, Cadarette SM. Disease risk score as a confounder summary method: systematic review and recommendations. *Pharmacoepidemiol Drug Saf*. 2013;22(2):122-129.
8. Wyss R, Ellis AR, Brookhart MA, et al. Matching on the disease risk score in comparative effectiveness research of new treatments. *Pharmacoepidemiol Drug Saf*. 2015;24(9):951-961.
9. Hansen BB. The prognostic analogue of the propensity score. *Biometrika*. 2008;95(2):481-488.
10. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61(1):76-86.
11. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98(9):691-698.
12. Debray TP, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KG. Meta-analysis and aggregation of multiple published prediction models. *Stat Med*. 2014;33(14):2341-2362.
13. Rubin DB. Randomization analysis of experimental data: the Fisher randomization test comment. *J Am Stat Assoc*. 1980;75(371):591-593.
14. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81:945-960.
15. Pearl J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology*. 2010;21(6):872-875.
16. Pirracchio R, Carone M, Rigon MR, Caruana E, Mebazaa A, Chevret S. Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. *Stat Methods Med Res*. 2016;25(5):1938-1954.
17. Imai K, Van Dyk DA. Causal inference with general treatment regimes. *J Am Stat Assoc*. 2004;99:854-866.
18. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
19. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. *Am J Epidemiol*. 2011;174(5):613-620.
20. Hajage D, De Rycke Y, Chauvet G, Tubach F. Estimation of conditional and marginal odds ratios using the prognostic score. *Stat Med*. 2017;36(4):687-716.
21. Groenwold RH, Moons KG, Pajouheshnia R, et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *J Clin Epidemiol*. 2016;78:90-100.
22. van Klaveren D, Vergouwe Y, Farooq V, Serruys PW, Steyerberg EW. Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. *J Clin Epidemiol*. 2015;68(11):1366-1374.
23. Wyss R, Hansen BB, Ellis AR, et al. The "dry-run" analysis: a method for evaluating risk scores for confounding control. *Am J Epidemiol*. 2017;185(9):842-852.
24. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.
25. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*. 2009;338:b606.
26. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer; 2009.
27. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381.

28. Honkoop PJ, Loijmans RJB, Termeer EH, et al. Symptom- and fraction of exhaled nitric oxide-driven strategies for asthma control: a cluster-randomized trial in primary care. *J Allergy Clin Immunol*. 2015;135(3):682-688.
29. Hahn S, Puffer S, Torgerson DJ, Watson J. Methodological bias in cluster randomised trials. *BMC Med Res Methodol*. 2005;5:10.
30. Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *BMJ*. 2003;327(7418):785-789.
31. Groenwold RH, Donders AR, van der Heijden GJ, Hoes AW, Rovers MM. Confounding of subgroup analyses in randomized data. *Arch Intern Med*. 2009;169(16):1532-1534.
32. VanderWeele TJ, Knol MJ. Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions. *Ann Intern Med*. 2011;154(10):680-683.
33. Loymans RJ, Honkoop PJ, Termeer EH, et al. Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. *Thorax*. 2016;71(9):838-846.
34. Schatz M, Cook EF, Joshua A, Petitti D. Risk factors for asthma hospitalizations in a managed care organization: development of a clinical prediction rule. *Am J Manag Care*. 2003;9(8):538-547.
35. Eisner MD, Yegin A, Trzaskoma B. Severity of asthma score predicts clinical outcomes in patients with moderate to severe persistent asthma. *Chest*. 2012;141(1):58-65.
36. Miller MK, Lee JH, Blanc PD, et al. TENOR risk score predicts healthcare in adults with severe or difficult-to-treat asthma. *Eur Respir J*. 2006;28(6):1145-1155.
37. Leyrat C, Seaman SR, White IR, et al. Propensity score analysis with partially observed covariates: how should multiple imputation be used? *Stat Methods Med Res*. 2019;28(1):3-19.
38. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. *Stat Med*. 2018;37(14):2252-2266.
39. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373-1379.
40. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and cox regression. *Am J Epidemiol*. 2007;165(6):710-718.
41. Hansen BB. Full matching in an observational study of coaching for the SAT. *J Am Stat Assoc*. 2004;99:609-618.
42. Rosenbaum PR. A characterization of optimal designs for observational studies. *J R Stat Soc B*. 1991;53:597-610.
43. Bouwmeester W, Zuithoff NP, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med*. 2012;9(5):1-12.
44. Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *JAMA*. 1993;270(24):2957-2963.
45. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1-W33.
46. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58.
47. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11(10):e1001744.
48. Martin GP, Mamas MA, Peek N, Buchan I, Sperrin M. A multiple-model generalisation of updating clinical prediction models. *Stat Med*. 2018;37(8):1343-1358.
49. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6(1): Article 25. https://doi.org/10.2202/1544-6115.1309
50. Polley EC, van der Laan MJ. Super learner in prediction. In. *U.C. Berkeley Division of Biostatistics Working Paper Series*. Berkeley, CA: The Berkeley Electronic Press; 2010.
51. Polley EC, Rose S, van der Laan MJ. Super learning. In: van der Laan MJ, Rose S, eds. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer; 2011.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Nguyen T-L, Collins GS, Pellegrini F, Moons KG, Debray TP. On the aggregation of published prognostic scores for causal inference in observational studies. *Statistics in Medicine*. 2020;39:1440–1457. https://doi.org/10.1002/sim.8489