

Simulation-based Design of Pragmatic Trials in Psoriatic Arthritis Using Propensity Scores

Journal Title XX(X):1–??
©The Author(s) 2020 Reprints
and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/

SAGE

Sarah M. Weinstein¹, Laura C. Coates², Philip S. Helliwell³, Alexis Ogdie^{1,4,*}, Alisa J. Stephens-Shields^{1,*}

Abstract

Design of clinical trials requires careful decision-making across several dimensions, including endpoints, eligibility criteria, and subgroup enrichment. Clinical trial simulation can be an informative tool in trial design, providing empirical evidence by which to evaluate and compare the results of hypothetical trials with varying designs. In this paper, we introduce a novel simulation-based approach using observational data to inform the design of a future pragmatic trial. To account for likely confounding by indication, we utilize propensity score-adjusted models to simulate hypothetical trials under alternative endpoints and enrollment criteria. We apply our approach to the design of pragmatic trials in psoriatic arthritis, using observational data embedded within the Tight Control of Inflammation in Early Psoriatic Arthritis study to simulate hypothetical open-label trials comparing treatment with tumor necrosis factor- α inhibitors to methotrexate. We first validate our simulations of a trial with traditional enrollment criteria and endpoints against a recently published trial. Next, we compare simulated treatment effects in patient populations defined by traditional and broadened enrollment criteria, where the latter is consistent with a future pragmatic trial. In each trial, we also consider five candidate primary endpoints. Our results highlight how changes in the enrolled population and primary endpoints may qualitatively alter study findings and the ability to detect heterogeneous treatment effects between clinical subgroups. These considerations, among others, are important for designing a future pragmatic trial aimed at having high external validity with relevance for real-world clinical practice. Our approach may be generalized to the study of other conditions where existing trial data are limited or do not generalize well to real-world clinical practice, but where observational data are available.

Keywords

clinical trial simulation, pragmatic trial design, causal inference, propensity score modeling

Introduction

By design, pragmatic trials should reflect aspects of real-world clinical practice by enrolling heterogeneous populations, assigning treatments without blinding, and evaluating comparative effectiveness between existing treatments. By mirroring key aspects of real-world clinical practice, pragmatic trials provide crucial insights into clinical decision-making and treatment effect heterogeneity, but may be especially challenging to design.^{1,2} Ideally, data from prior trials may be used to select clinically meaningful endpoints with high responsiveness to the treatment in the target patient population. Since it is often not possible to find an existing trial whose enrollment criteria are broad enough to reflect the heterogeneity of a real-world clinical population, one strategy is to turn to observational data.

[Version: 2017/01/17 v1.20]

It is widely understood that confounding is inherent to observational data, such that certain therapies may be prescribed more often to patients with a poorer prognosis,

¹Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, USA

²Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, UK

³Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, UK

⁴Division of Rheumatology, Center for Pharmacoepidemiology Research and Training, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, USA * Drs. Ogdie and Stephens-Shields contributed equally.

Corresponding author:

Sarah M. Weinstein

Email: smweinst@penmedicine.upenn.edu

resulting in worse outcomes that may be more related to pretreatment prognosis than to a lack of treatment efficacy.³ Causal inference techniques, such as propensity score modeling, are often used to recover unbiased estimates of treatment effects in observational studies. With appropriate assumptions and models, treatment effect estimates may then be interpreted causally—as if they were derived from a randomized trial.⁴ Causal inference tools have generally not been considered in

prospective trial design; they are more commonly used for retrospective analysis of observational data to emulate trials, or to address noncompliance and missing data.^{5–8}

In this paper, we introduce a novel approach combining observational data, causal inference, and simulations to inform design of future pragmatic trials. We motivate and apply our method using observational data embedded within the Tight Control of Inflammation in Early Psoriatic Arthritis (TICOPA) study.⁹ Psoriatic arthritis is a highly heterogeneous disease, but previous trials have primarily enrolled homogeneous patient samples and have not considered treatment effect heterogeneity among subgroups.¹⁰ Using observational data, we simulate two hypothetical pragmatic trials in psoriatic arthritis to compare treatment with tumor necrosis factor (TNF)- α inhibitors to methotrexate in a setting that mimics real-world clinical practice. In brief, our hypothetical trials aim to:

Trial 1: Replicate an existing trial of TNF- α inhibitors versus methotrexate in patients with a minimum of both three tender and three swollen joints.^{11,12} Patients enrolled in Trial 1 would have been eligible for enrollment in traditional explanatory trials in psoriatic arthritis, and are considered to be clinically “severe” based on their high numbers of tender and swollen joints.

Trial 2: Evaluate the impact of broadened enrollment criteria in a hypothetical trial of patients who would receive TNF- α inhibitors versus methotrexate in a more representative population without minimum joint counts (i.e., including clinically “severe” and “non-severe” individuals). Notably, by using observational data, we are able to “enroll” participants in Trial 2 who would not have been eligible for participation in previous trials, but who are ultimately part of the target population.

In both Trials 1 and 2, we also compare five primary endpoints, allowing us to assess how trial findings may change with different choices of endpoint.

Methods

Overview of the TICOPA study

TICOPA was a randomized multicenter trial whose primary goal was to evaluate the effect of intensive management, characterized by scheduled rheumatology appointments

every four weeks, versus standard care, with appointments every 12 weeks, in the treatment of psoriatic arthritis.⁹ Although TICOPA was randomized with respect to appointment frequency, it is effectively an observational study with respect to the comparison of TNF- α inhibitors versus methotrexate. Assignment to these therapies was not randomized and was therefore inherently subject to confounding by indication.³

The original TICOPA trial included $n=206$ patients, 179 of whom consented to have their data included in additional studies like the present one. All 179 patients in this subset initiated treatment with methotrexate; 85 (47.5%) were assigned to standard care and 94 (52.5%) were assigned to intensive management. At subsequent visits, 44 patients (7 (8.2%) standard care and 37 (39.4%) intensive management) switched to TNF- α inhibitors, often due to a lack of improvement under treatment with methotrexate. In the current study, we are primarily interested in disease progression following initiation of a new treatment; thus, we re-align the TICOPA data to include $n=223$ unique treatment initiation start points, which double-counts the 44 individuals who switched from methotrexate to TNF- α inhibitors.

We classified patients with fewer than three tender or swollen joints at the time of treatment initiation as “nonsevere” and those with at least three of both as “severe,” where “severe” patients are typically enrolled in psoriatic arthritis trials. At the time of treatment initiation, 148 patients were classified as “severe” (with 127 (85.8%) receiving methotrexate; 21 (14.2%) TNF- α inhibitors) and 75 were considered “non-severe” (with 52 (65.3%) receiving methotrexate; 23 (30.7%) TNF- α inhibitors).

Notation

In the TICOPA data and in hypothetical simulated trials of n participants, let A be an n -dimensional vector of treatment assignments to either methotrexate or TNF- α inhibitors ($A_i = 0$ or $A_i = 1$, respectively, for $i = 1, \dots, n$). Let $X_0 = [X_{01}, \dots, X_{0K}]$ be an $n \times K$ matrix of covariates at the time of treatment initiation. Similarly, let $X_1 = [X_{11}, \dots, X_{1K}]$ denote the $n \times K$ matrix of the same covariates at follow-up (including change in covariates between baseline and follow-up), where the latter timepoint corresponds to 12 weeks after treatment initiation. Let Y denote a candidate endpoint derived from X_1 : the American College of Rheumatology 20% Response Criteria (ACR20), and four psoriatic arthritis-specific endpoints: the Psoriatic Arthritis Disease Activity Score, the

Disease Activity Index for Psoriatic Arthritis (DAPSA), clinical DAPSA, or the Routine Assessment of Patient Index Data 3, using formulas provided in Supplemental Table S1.^{9,13–15}

Confounding adjustment

Propensity score methodologies are widely used in observational studies to estimate an average causal effect of a binary treatment (A) on an outcome (Y) in the presence of confounders (X_0).^{16,17} The propensity score, $P(A = 1 | X_0) = \pi(X_0)$, summarizes a high-dimensional set of confounders; thus, they are especially useful in small samples, where the usual multiple regression approach limits the number of confounders that may be used for adjustment.^{18–21}

An average causal effect is defined as a contrast of mean potential outcomes Y^1 and Y^0 (e.g., $E[Y^1]/E[Y^0]$ or $E[Y^1] - E[Y^0]$), where Y^a denotes the outcome that would be observed for the i th individual if they were to receive treatment $A = a$. Although both potential outcomes are not observed for any individual, under the assumptions of positivity ($0 < \pi(X_0) < 1$ for all X_0), conditional ignorability ($(Y^1, Y^0) \perp A | X_0$), and consistency ($Y_i = Y_i^a$ for $A_i = a$), average treatment effects (ATEs) may be identified and estimated with observed data using methods such as multiple regression or propensity score adjustment, matching, or weighting.^{16,17}

As shown in **Model C**, we use propensity score adjustment to fit models of X_1 and derive outcomes in our simulated hypothetical trials. In small samples and when the overlap in the distribution of propensity scores is poor, a propensity score-adjusted regression model is preferable to matching, stratification, or weighting.^{20–22}

Target models and estimation

In this section, we define models for baseline and followup variables. Parameter estimates from these models are used to define the distributions from which participant-level measurements are randomly sampled in simulated trials.

Continuous follow-up and change variables are each assumed to follow a normal distribution with participantlevel means μ_i and variance σ^2 . Count variables are assumed to follow a negative binomial distribution with participantlevel means λ_i and dispersion parameter θ . Binary variables are assumed to follow a Bernoulli distribution with participant-level probabilities p_i . We

estimate participantspecific means ($\hat{\mu}_i$, $\hat{\lambda}_i$, and \hat{p}_i) using predictions from **Model A**, **Model C**, and **Model D** below.

Models for baseline data (X_0) Models for baseline covariates are fit separately in “severe” ($S = 1$) and “nonsevere” ($S = 0$) groups. We represent these models using the following notation:

$$E[X_{0k} | \mathbf{X}_{0\{1:(k-1)\}}, S = s] = g(\mathbf{X}_{0\{1:(k-1)\}}; \alpha^s), \quad (\text{Model A})$$

where $g()$ may be a standard link function or more complex nonlinear model. We select $g()$ by simultaneously fitting and comparing standard parametric and nonparametric models for these variables using the SuperLearner package in R.²³ After modeling a subset of variables based on their observed marginal distributions in the TICOPA population (e.g., sex as a binary variable distributed Bernoulli(0.5)), we use SuperLearner to select models for the remaining baseline variables, in accordance with the order in which those variables would be simulated.

Because the TICOPA dataset double-counts the 44 patients who switched treatment groups, we model the variance for continuous variables, σ^2 , by fitting a linear mixed model of predictions $\hat{\mu}_i$ on the corresponding observations in the TICOPA data; we then extract the error variance from this model to obtain σ^2 . The dispersion parameter for count variables, θ , is modeled by fitting a quasi-Poisson generalized linear model of $\hat{\lambda}_i$ on the corresponding observations in the TICOPA data, and extracting $\hat{\theta}$ from model output.

Propensity score models We assume the following model for the propensity score:

$$\pi(X_0) = P(A = 1 | X_0) = f(X_0; \theta), \quad (\text{Model B})$$

where $f()$ is an invertible link function (e.g., logit) involving a parameter θ .

With a large number of possible confounders and a relatively small sample size in the TICOPA data, we consider several approaches to modeling $\pi(X_0)$. Treatment effect estimates obtained from over-fit propensity score models can be unstable, and misspecified propensity score models can result in biased treatment effect estimates.^{24,25} We therefore use model selection to determine a propensity score model to simulate realistic, stable outcomes for

hypothetical trials. Considerations for the final choice of a model are described in **Model validation** below.

We consider three main approaches to selecting a propensity score model:

1. Semi-automated model selection. To facilitate automated selection of a model for $\pi(\mathbf{X}_0)$, we use the SuperLearner package in R to simultaneously fit multiple candidate models (random forest, stepwise regression, gradient boosting machines, and others).²³ While SuperLearner is often used to build ensemble models, we do not use this feature due to the modest sample size of the TICOPA data as well as the marginal advantages of ensemble propensity score models.²² We then select a model that both maximizes the cross-validated area under the receiver operating characteristic curve (AUC) for prediction of treatment group and that satisfies the positivity assumption (i.e., overlap between propensity score distributions of the two treatment groups).
2. Standardized mean differences (SMDs). Our second approach to model selection for the propensity score involves a descriptive analysis of the baseline TICOPA data to identify any covariates that are imbalanced across treatment groups. Such imbalance would need to be accounted for in simulating hypothetical randomized trials. For the k th baseline covariate,

$$\text{SMD} = \frac{\bar{X}_{0k}^{(A=0)} - \bar{X}_{0k}^{(A=1)}}{\hat{\sigma}_p},$$

where the numerator is the difference in sample means between the treatment groups and the denominator is the pooled sample standard deviation. We consider covariates for which the absolute value of the SMD ≥ 0.1 to be imbalanced.²⁶ After identifying an initial subset of variables based on the SMD, we consider further restricting the number of variables in this model based on collinearity and missingness.²⁴ Finally, we fit a multiple logistic regression model with binary treatment assignment as the outcome and

$\mathbf{X}_0 = \mathbf{X}_{0(\text{SMD})}$ as covariates to obtain an SMD-based model of $\pi(\mathbf{X}_0)$.

3. Directed acyclic graphs (DAGs). DAGs are commonly used in the causal inference literature to

conceptualize the relationship among treatments, confounders, and outcomes to identify a sufficient set of variables to control for confounding. DAG-based variable selection considers TICOPA study design and expert opinion, allowing us to identify a reasonably sized set of confounders.²⁷ This approach informs variable selection for a DAG-based model of $\pi(\mathbf{X}_0)$, using $\mathbf{X}_0 = \mathbf{X}_{0(\text{DAG})}$.

Models for follow-up data (\mathbf{X}_1) For the k th follow-up variable ($k = 1, \dots, K^*$), we consider the following models, which are fit within each treatment arm ($A = a$) and severity group ($S = s$):

$$E[X_{1k} | \pi(\mathbf{X}_0), A = a, S = s] = g\left(\beta_0^{a,s} + \beta_1^{a,s} \times \pi(\mathbf{X}_0)\right) \quad (\text{Model C})$$

$$E[X_{1k} | \mathbf{X}_0, A = a, S = s] = g\left(\gamma_0^{a,s} + \gamma_1^{a,s} \times \mathbf{X}_0\right), \quad (\text{Model D})$$

where $g()$ is an inverse logit link for binary variables, inverse log link for count variables, and identity link for continuous and change variables. In **Model C**, we use predictions of $\pi(\mathbf{X}_0)$ as the sole covariate in a simple regression model.

The multiple regression-based model, **Model D**, directly links observed baseline covariates (\mathbf{X}_0) with follow-up variables (\mathbf{X}_1) without the intermediate step of modeling the propensity score. For this model, we use DAG-based variable selection to identify $\mathbf{X}_{0(\text{MR})}$, a subset of \mathbf{X}_0 for inclusion in the multiple regression-based model for \mathbf{X}_1 . We only use DAG-based variable selection for our multiple regression model as we believe it is the best approach for specifically identifying confounders that are related to both the treatment and follow-up measurements, whereas SMD-based variable selection only looks at association with treatment, and our semi-automated model selection approach does not inherently prioritize confounding selection. We select $\mathbf{X}_{0(\text{MR})}$ to be a smaller subset of \mathbf{X}_0 than $\mathbf{X}_{0(\text{DAG})}$ due to constraints regarding the appropriate number of variables that may be used in a multiple logistic regression model as a function of the number of events and sample size.^{18,19}

We estimate the variance parameter for continuous followup variables, σ^2 , as the within-person variance component from fitting **Model C** and **Model D**. The negative

binomial dispersion parameter, θ , is estimated in the same way as described previously for X_0 .

The type of model used for simulations remains fixed between Trials 1 and 2 to ensure comparability between the trials. Baseline models (**Model A**) are then re-fit separately within severe and non-severe groups in the TICOPA population, and follow-up variable models (**Model C** and **Model D**) are fit separately within severe/non-severe and TNF- α inhibitors/methotrexate groups.

Trial simulation based on target models

Trials 1 and 2 are each simulated 1000 times, with $n = 200$ participants per iteration. For a given trial and model for follow-up variables, the j th iteration involves the following steps for $j = 1, \dots, 1000$:

1. Simulate the j th baseline dataset, $X_0^{(j)}$ by randomly sampling observations for $n = 200$ patients with parameter estimates obtained from realizations of **Model A**. In Trial 1, all 200 patients are simulated according to parameters from models fit to “severe” individuals in the TICOPA data. In Trial 2, we designate 100 patients to be simulated in each severity group, where separate parameters are used to simulate baseline data for “severe” and “non-severe” individuals..
2. Randomly assign $n/2 = 100$ patients to each treatment group ($A_i = 0$ for methotrexate or $A_i = 1$ for TNF- α inhibitors), independent of $X_0^{(j)}$.
3. Apply parameter estimates from a given realization of **Model B** to $X_0^{(j)}$ to obtain $\pi(\mathbf{X}_0^{(j)})$, propensity score estimates from the j th simulated trial.
4. For propensity score-based generation of followup variables, apply treatment ($A = a$) and severity ($S = s$) group-specific parameter estimates from **Model C** to $\pi(\mathbf{X}_0^{(j)})$ to obtain participant-level mean estimates for follow-up variables. For multiple regression-based generation of follow-up variables, apply group-specific estimates from **Model D** to $X_0^{(j)}$ to obtain these participant-level means. Then, randomly sample participant-level measurements from appropriate distributions with participant-level means and group-level parameter estimates (e.g., σ^2 and $\hat{\theta}$ for continuous and count variables,

respectively) to obtain $X_1^{(j)}$, the follow-up measurements for the j th simulated trial.

5. Derive one or more candidate trial endpoints, $Y^{(j)}$, by applying formulas from Supplemental Table S1 to simulated follow-up variables, $X_1^{(j)}$. For each endpoint, estimate a treatment effect as a contrast of the estimated average outcomes (e.g., risk ratio

$$RR_d = E^{(j)}[Y^{(j)} | A = 1] / E^{(j)}[Y^{(j)} | A = 0] \quad \text{or risk difference } RD_d = E^{(j)}[Y^{(j)} | A = 1] - E^{(j)}[Y^{(j)} | A = 0].$$

After repeating steps 1-5 for 1000 simulations, we plot distributions of treatment effect estimates and obtain percentile-based 95% confidence intervals.

Model validation

We first consider 1000 simulations of Trial 1 using ACR20 as the primary endpoint. In this initial set of simulations, the enrollment criteria (“severe” individuals, with ≥ 3 tender and ≥ 3 swollen joints at treatment initiation) and endpoint (ACR20) match those of a trial published by Mease et al. (2019).¹¹ We compare treatment effect estimates measured 12 weeks after treatment initiation in Mease et al.’s study to those from each candidate model used to simulate X_1 and select the model whose distribution of treatment effect estimates are closest to Mease et al.’s reported estimates. This step ensures credibility of our simulated treatment effect estimates.

Results

Models selected for confounding adjustment

For our semi-automated approach to propensity score model selection, we found that the random forest and stepwise regression models achieved the top two highest crossvalidated AUCs (0.78 [95% C.I. 0.70, 0.87] and 0.74 [95% C.I. 0.62, 0.85], respectively). After cross-validation, we assessed the overlap in propensity scores estimated from each model trained in the TICOPA sample and found a complete lack of overlap for the random forest model (Supplemental Figure S1), suggesting a positivity violation. This finding is consistent with earlier work discussing problems that can arise when propensity score models are optimized to achieve high predictive accuracy.^{22,28} Thus, we

proceed by only considering results from the stepwise regression for our semi-automated realization of **Model B**. This model selected the following baseline variables: global disease activity (physician assessment), TICOPA treatment assignment (standard care/intensive management), and tender joint count.

Information used for variable selection via SMDs is presented in Table 1. The following baseline variables were selected for this realization of **Model B**: TICOPA treatment allocation, patient and physician assessments of global disease activity, enthesitis diagnosis, C-reactive protein, nail disease, dactylitis, age, tender joint count, and psoriatic arthritis quality of life score. Baseline variables for which the magnitude of the SMD is above 0.1 were all included in the SMD-based propensity score model, except the 7day assessment of pain and the Bath Ankylosing Spondylitis Disease Activity Index. We exclude these variables due to their high correlations (> 90% and > 80%, respectively) with the patient assessment of global disease activity, which is included in the model.

Figure 1 informs the DAG-based realization of **Model B**, which identified TICOPA treatment allocation, baseline patient and physician assessments of global disease activity, and baseline tender joint count as confounders of the association between treatment and outcome. For the multiple regression-based model, the same variables, except for the physician measure of global disease activity, were included.

For the SMD and DAG-based propensity score models, we also verify adequate overlap in the propensity score distributions of treatment groups in Supplemental Figure S1. Parameter estimates for propensity score models (**Model B**) and propensity score-based (**Model C**) and multiple regression-based (**Model D**) models for follow-up variables are included in Supplemental Tables S2, S3, and S4, respectively.

Model validation through comparison with an existing trial

We use treatment effect estimates from a recent trial by Mease et al. (2019) to guide selection of a model for simulating realistic trials. Mease et al. enrolled “severe” patients diagnosed with psoriatic arthritis to receive either methotrexate monotherapy, TNF- α inhibitors monotherapy, or a combination of both. At 12 weeks after treatment initiation, they found more favorable outcomes among patients in either arm involving TNF- α inhibitors

compared to those who received methotrexate alone (risk ratio = 1.5).¹¹

Risk ratio distributions for the ACR20 outcome from our simulated Trial 1 and point estimates from Mease et al. are plotted in Figure 2. Our simulated Trial 1 risk ratios for ACR20 using the stepwise regression-based propensity scores are best aligned with Mease et al.’s 12-week ACR20 endpoint, with a mean risk ratio 1.06 and 95% C.I. [0.63, 1.70]. Thus, in our subsequent assessments of simulated trial results—where enrollment criteria and endpoints differ from those used in the trial by Mease et al. (2019)—we consider only those generated using the stepwise regression model.

Intervention effects in hypothetical trials

We next consider simulated trials using five candidate endpoints (in both Trials 1 and 2) and broadened enrollment criteria (in Trial 2). Figure 3 shows risk differences for each candidate outcome across 1000 simulations of Trials 1 and 2, with estimates > 0 favoring TNF- α inhibitors. We compare treatments using the risk difference, since Trial 1 iterations generally produced no methotrexate-receiving patients who met the threshold for DAPSA low disease activity, precluding calculation of the risk ratio.

Trial 1, which enrolled a homogeneous group of “severe” patients, shows a pattern of favorable response to TNF- α inhibitors over methotrexate. Trial 2, which enrolled a more heterogeneous population of “severe” and “non-severe” patients that would be targeted in a future pragmatic trial, shows a trend favoring methotrexate over TNF- α inhibitors.

Implications for designing a future trial with a generalized patient population and novel primary endpoints

To bridge the gap between explanatory trials and clinical practice, pragmatic trials need to enroll heterogeneous groups of patients reflective of a target clinical population. Our proposed simulation-based trial design allows clinical researchers to anticipate the impact of expanding inclusion criteria and determine statistical analysis plans and primary endpoints accordingly.

Based on our simulations, the statistical analysis plan for a future pragmatic trial in psoriatic arthritis should describe estimating treatment effects after stratifying on baseline disease severity. We found that pooling all participants for

estimating treatment effects in Trial 2 precluded the assessment of differences in treatment favorability between “severe” individuals—for whom TNF- α inhibitors appear to be favored—and “non-severe” individuals who have fewer tender/swollen joints at baseline—for whom methotrexate seems more favorable. A stratified analysis would make it possible to capture these heterogeneous treatment effects.

A primary endpoint should be selected based on its relevance to the disease under study and ability to capture clinically meaningful differences between treatment groups. Simulated trials using various candidate endpoints allow researchers to rule out endpoints with more modest effect estimates, which would require larger sample sizes to estimate statistically and clinically meaningful differences between treatment groups.

In our simulations, we were able to identify endpoints that would be more likely to yield modest effect estimates (i.e., closer to 0 in Figure 3). The Psoriatic Arthritis Disease Activity Score would be an intuitive choice of a primary outcome, as our simulations show greater responsiveness to TNF- α inhibitors versus methotrexate in both Trials 1 and 2. Importantly, this psoriatic arthritis-specific score measures disease activity across multiple domains of psoriatic arthritis, such as peripheral arthritis, skin psoriasis, and nail disease, whereas ACR20 (the primary endpoint from previous studies) focuses only on peripheral joints.¹³

Discussion

In this paper, we use observational data to simulate hypothetical clinical trials with different enrollment criteria and endpoints from previous trials. We show that propensity score-adjusted models can be used to mitigate confounding in simulated hypothetical trials, which may ultimately guide aspects of decision-making in designing a future pragmatic trial.

This paper differs from earlier work in simulation-based trial design, which have focused on sample size calculation and statistical power using historical data and power priors.^{29,30} Our proposed method also differs from network meta-analyses that synthesize results from previous trials to improve sample size calculation, as we wish to address trial design in settings where there are not enough existing trials for a meta-analysis.³¹

The present study also expands on earlier work in

clinical trial generalizability.^{32,33} Our method provides a novel strategy for researchers to empirically assess and anticipate the impact of enhanced generalizability (through broadened enrollment criteria) on the direction and magnitude of treatment effect estimates for heterogeneous patient subgroups.

Our simulation-based approach to pragmatic trial design can be applied to other areas of clinical research where observational data are available and generalizable pragmatic trials have yet to be conducted. Although the observational sample considered in this paper has fewer than 200 participants, our methodology could easily be applied to large healthcare databases, such as electronic health records (EHR), insurance claims databases, or prospective cohort study data, as long as the variables measured in the observational data are similar to what would be measured in a future trial.

To implement our approach, it is necessary for the available observational data to have some representation of the group that has not yet been studied in previous trials. Researchers should stratify the observational sample according to variables that would have made patients eligible or ineligible for enrollment in previous trials, estimating parameters for models of baseline (Model A) and followup variables (Model C and Model D) separately for subgroups. In the current study, this was the role of the “severity” index: we fit models for baseline and followup variables separately for $S = 1$ (≥ 3 tender and swollen joints) and $S = 0$ (< 3 tender or swollen joints) to account for possible treatment effect heterogeneity between patients in the TICOPA observational sample who would have been eligible ($S = 1$) or ineligible ($S = 0$) to enroll in traditional trials in psoriatic arthritis.

In the absence of existing trials, a different approach to model validation could be taken, since treatment effect estimates from a previous trial (as in Figure 2) could not be used. An alternative method would be to select a model that yields qualitatively similar simulation results for different choices of endpoints. While we did not rely on this approach for the current study, we note that the stepwise regression propensity score-based model for follow-up variables yields the most consistent results across different choices of an endpoint for both Trials 1 and 2 (Supplemental Figure S2).

This study has some limitations, including the possible instability of our model selection procedure. While SuperLearner has often been used in high dimensional data settings with larger sample sizes, we used this tool to model

data with a more modest sample size.³⁴ This may have contributed to some instability of SuperLearnerbased model selection, such that we noticed changes in “best” model (according to maximizing cross-validated AUC) with different choices of a random seed. However, the variables selected for the stepwise regression model from SuperLearner do not vary with different choices of a random seed.

The modest sample size of the TICOPA data is another limitation, though we believe the strengths of these data, which were collected in a consistent manner and with a fixed schedule, outweigh sample size concerns. For future research, it is worth noting that with a larger observational sample with greater representation of disease subgroups, it would be possible to consider more variations to enrollment criteria and subgroup enrichment by selecting and fitting models for baseline and follow-up variables within more narrowly defined subgroups.

Our method can also be adapted and extended to settings where different strategies for confounding adjustment, such as inverse probability weighting or matching, may be preferred.^{21,22} Overall, this paper introduces a flexible framework for incorporating observational data in prospective trial design, providing an empirical framework to support decision-making in pragmatic trials.

Declaration of conflicting interests

Alexis Ogdie has consulted for Abbvie, Amgen, BMS, Celgene, Corrona, Janssen, Lilly, Novartis and Pfizer and has received grants from Amgen (to Forward), Novartis (to Penn), and Pfizer (to Penn). Her husband has received royalties from Novartis.

Funding

This study was funded by the Innovative Research Award from the Rheumatology Research Foundation (co-PI: Alexis Ogdie and Alisa Stephens-Shields).

Laura C. Coates is funded by a National Institute for Health Research Clinician Scientist award. The research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Sarah M. Weinstein is funded by the National Science Foundation Graduate Research Fellowship Program (NSF GRFP).

Supplemental material

Supplemental material for this article is available online.

Acknowledgements

We are grateful to Dr. Jesse Hsu and two anonymous reviewers whose constructive feedback greatly enhanced our paper.

References

1. Ford I and Norrie J. Pragmatic Trials. *The New England Journal of Medicine* 2016; 375(5): 454–463.
2. Godwin M, Ruhland L, Casson I et al. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. *BMC medical research methodology* 2003; 3(1) 28.
3. Kyriacou DN and Lewis RJ. Confounding by indication in clinical research. *Jama* 2016; 316(17): 1818–1819.
4. Hernan MA and Robins JM. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health* 2006; 60(7): 578–586.
5. Heitjan DF. Causal inference in a clinical trial: a comparative example. *Controlled clinical trials* 1999; 20(4): 309–318.
6. Jo B, Ginexi EM and Ialongo NS. Handling missing data in randomized experiments with noncompliance. *Prevention Science* 2010; 11(4): 384–396.
7. Hernan MA and Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology* 2016; 183(8): 758–764.
8. Caniglia EC, Rebecca Z, Jacobson DL et al. Emulating a target trial of antiretroviral therapy regimens started before conception and risk of adverse birth outcomes. *AIDS (London, England)* 2018; 32(1): 113. Pmid:29112066.
9. Coates LC, Moverley AR, McParland L et al. Effect of tight control of inflammation in early psoriatic arthritis (ticopa): a uk multicentre, open-label, randomised controlled trial. *The Lancet* 2015; 386(10012): 2489–2498.
10. Coates LC and Helliwell PS. Psoriatic arthritis: state of the art review. *Clinical Medicine* 2017; 17(1): 65. Pmid:28148584.
11. Mease PJ, Gladman DD, Collier DH et al. Etanercept and methotrexate as monotherapy or in combination for psoriatic arthritis: primary results from a randomized, controlled phase iii trial. *Arthritis & Rheumatology* 2019; 71(7): 1112–1124.
12. Ogdie A and Coates L. The changing face of clinical trials in psoriatic arthritis. *Current rheumatology reports* 2017; 19(4): 21.
13. Perruccio AV, Got M, Li S et al. Treating psoriatic arthritis to target: defining psoriatic arthritis disease activity score

- (pasdas) that reflects state of minimal disease activity (mda). The Journal of rheumatology 2019; Pmid:31203221.
14. Schoels MM, Aletaha D, Alasti F et al. Disease activity in psoriatic arthritis (psa): defining remission and treatment success using the dapsa score. Annals of the Rheumatic Diseases 2016; 75(5): 811–818. Pmid:26269398.
 15. Pincus T, Yazici Y and Bergman MJ. Rapid3, an index to assess and monitor patients with rheumatoid arthritis, without formal joint counts: similar results to das28 and cdaï in clinical trials and clinical care. Rheumatic Disease Clinics 2009; 35(4): 773– 778. Pmid:19962621.
 16. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70(1): 41–55.
 17. Williamson E, Morley R, Lucas A et al. Propensity scores: from naive enthusiasm to intuitive understanding. Statistical methods in medical research 2012; 21(3): 273–293.
 18. Peduzzi P, Concato J, Kemper E et al. A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology 1996; 49(12): 1373–1379. Pmid:8970487.
 19. Vittinghoff E and McCulloch CE. Relaxing the rule of ten events per variable in logistic and cox regression. American Journal of Epidemiology 2007; 165(6): 710–718.
 20. Sturmer T, Joshi M, Glynn RJ et al. A review of the application“ of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. Journal of Clinical Epidemiology 2006; 59(5): 437– e1.
 21. Vansteelandt S and Daniel RM. On regression adjustment for the propensity score. Statistics in medicine 2014; 33(23): 4053–4072.
 22. Alam S, Moodie EEM, Stephens DA et al. Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. Statistics in medicine 2019; 38(9): 1690– 1702.
 23. Polley E, LeDell E, Kennedy C et al. Package ‘superlearner’ 2019.
 24. Schuster T, Lowe WK, Platt RW Propensity score model overfitting led to inflated variance of estimated odds ratios Journal of Clinical Epidemiology 2016; 80: 97–106
 25. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensityscore matched samples Statistics in Medicine 2009; 28(5): 3083–3107
 26. Stuart EA, Lee BK, and Leacy, FP. Prognostic score-based balance measures for propensity score methods in comparative effectiveness research Journal of Clinical Epidemiology 2013; 66(8): S84–S90.
 27. Sauer BC, Brookhart MA, Roy J et al. A review of covariate selection for non-experimental comparative effectiveness research. Pharmacoepidemiology and drug safety 2013; 22(11): 1139–1145.
 28. Moodie EEM and Stephens DA. Treatment Prediction, Balance, and Propensity Score Adjustment Epidemiology 2017; 28(5): e51–e53.
 29. Shortreed SM, Rutter CM, Cook AJ et al. Improving pragmatic clinical trial design using real-world data. Clinical Trials 2019; 16(3): 273–282.
 30. De Santis, F. Using historical data for Bayesian sample size determination. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2007; 170(1): 95–113.
 31. Salanti G, Nikolakopoulou A, Sutton AJ et al. Planning a future randomized clinical trial based on a network of relevant past trials Trials 2018; 19(1): 365.
 32. Stuart EA, Bradshaw CP, and Leaf PJ. Assessing the Generalizability of Randomized Trial Results to Target Populations Prevention Science 2015; 16(3): 475-485.
 33. Stuart EA, Ackerman B, and Westreich D. Generalizability of Randomized Trial Results to Target Populations: Design and Analysis Possibilities Research on Social Work Practice 2017; 28(5): 532-537.
 34. Wyss R, Schneeweiss S, van der Laan M et al. Using super learner prediction modeling to improve high-dimensional propensity score estimation. Epidemiology 2018; 29(1): 96– 106.

Table 1. Variables measured at the time of treatment initiation, stratified by treatment assignment in TICOPA for conducting variable selection based on standardized mean differences.

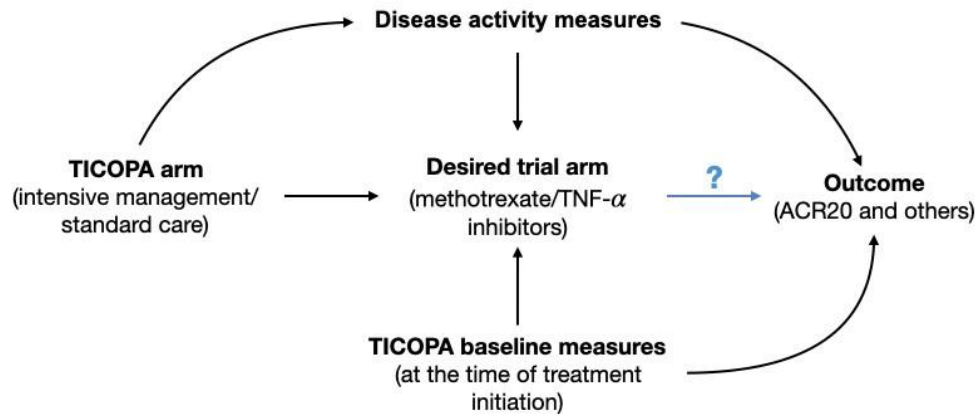
	Methotrexate (n = 179)	TNF- α inhibitors (n = 44)	Absolute value of the SMD
TICOPA Treatment Allocation (no. %)			
Standard Care	85 (47.5)	7 (15.9)	0.721

Intensive Management	94 (52.5)	37 (84.1)	
Patient 7-day pain assess (mean (SD))	51.42 (28.57)	48.23 (29.65)	0.110
Missing (no. %)	5 (2.79)	0 (0)	
Patient global disease activity assess (mean (SD))	54.65 (29.16)	50.48 (30.20)	0.140
Missing (no. %)	7 (3.91)	0 (0)	
Health assessment questionnaire score (mean (SD))	6.70 (5.28)	7.02 (5.53)	0.059
Missing (no. %)	3 (1.68)	0 (0)	
Psoriasis Area and Severity Index (mean (SD))	2.94 (4.27)	2.73 (5.58)	0.042
Missing (no. %)	0 (0)	13 (29.55)	
Enthesitis (%)			
No	40 (22.5)	13 (29.5)	0.162
Yes	138 (77.5)	31 (70.5)	
Missing (no. %)	1 (0.56)	0 (0)	
Tender Joint Count (mean (SD))	13.72 (14.67)	17.82 (17.94)	0.250
Missing (no. %)	1 (0.56)	5 (11.36)	
Swollen Joint Count (mean (SD))	7.27 (7.23)	6.66 (9.89)	0.071
Missing (no. %)	1 (0.56)	6 (13.64)	
C-reactive protein (mean (SD))	19.61 (32.39)	13.38 (23.28)	0.221
Missing (no. %)	6 (3.35)	2 (4.55)	
Physician assessment disease activity (mean (SD))	42.27 (20.32)	27.32 (20.33)	0.736
Missing (no. %)	4 (2.23)	0 (0)	
Nail Disease (no. %)			
No	70 (39.3)	20 (47.6)	0.168
Yes	108 (60.7)	22 (52.4)	
Missing (no. %)	1 (0.56)	2 (4.55)	
Dactylitis (no. %)			
No	133 (74.7)	38 (86.4)	0.297
Yes	45 (25.3)	6 (13.6)	

<i>Missing (no. %)</i>	1 (0.56)	2 (4.55)	
BASDAI (mean (SD))	5.08 (2.40)	4.72 (2.64)	0.145
<i>Missing (no. %)</i>	17 (9.50)	2 (4.55)	
Psoriatic Arthritis Quality of Life Score (mean (SD))	8.18 (6.43)	9.95 (6.44)	0.276
<i>Missing (no. %)</i>	9 (5.03)	2 (4.55)	
Sex (no. %)			
Female	87 (48.6)	23 (52.3)	0.073
Male	92 (51.4)	21 (47.7)	
Body Mass Index (mean (SD))	29.16 (6.11)	29.42 (6.92)	0.040
<i>Missing (no. %)</i>	25 (13.97)	7 (15.91)	
Age (mean (SD))	43.97 (12.71)	40.95 (12.45)	0.240

TICOPA (Tight Control of Inflammation in Early Psoriatic Arthritis⁹); SMD (standardized mean difference); TNF (tumor necrosis factor); BASDAI (Bath Ankylosing Spondylitis Disease Activity Index)

Figure 1. Directed acyclic graph-based variable selection. We identify TICOPA treatment allocation, patient and physician assessments of global disease activity, and tender joint count as confounders of the association between the desired trial arm and outcome.



TICOPA (Tight Control of Inflammation in Early Psoriatic Arthritis) study by Coates et al. (2015)⁹; TNF (tumor necrosis factor); ACR20 (American College of Rheumatology 20% Response Criteria)

Figure 2. Risk ratios for achieving the American College of Rheumatology 20% Response Criteria (ACR20) from 1000 simulations of Trial 1 compared with estimates from a trial by Mease et al. (2019).¹¹

Each box shows the distribution of risk ratios for achieving the ACR20 threshold 12 weeks after treatment initiation in our simulated trials based on four candidate models for follow-up variables. A subset of follow-up variables (X_1) are then used to derive the binary ACR20 outcome (Y). Risk ratios are calculated as $\Pr(Y | \text{TNF-}\alpha \text{ inhibitors}) / \Pr(Y | \text{methotrexate})$, where > 1 favor TNF- α inhibitors and < 1 favor methotrexate. Estimated mean and 95% confidence intervals for the ACR20 risk ratios from Trial 1 are: SuperLearner's stepwise regression 1.06 [0.63, 1.70]; directed acyclic graph 0.96 [0.56, 1.48]; standardized mean differences 0.91 [0.56, 1.38]; multiple regression with DAG-based variable selection 1.02 [0.65, 1.57]. The green line corresponds to the estimated 12-week risk ratio reported by Mease et al. (2019). We also include Mease et al.'s reported 24-week risk ratio for comparison.

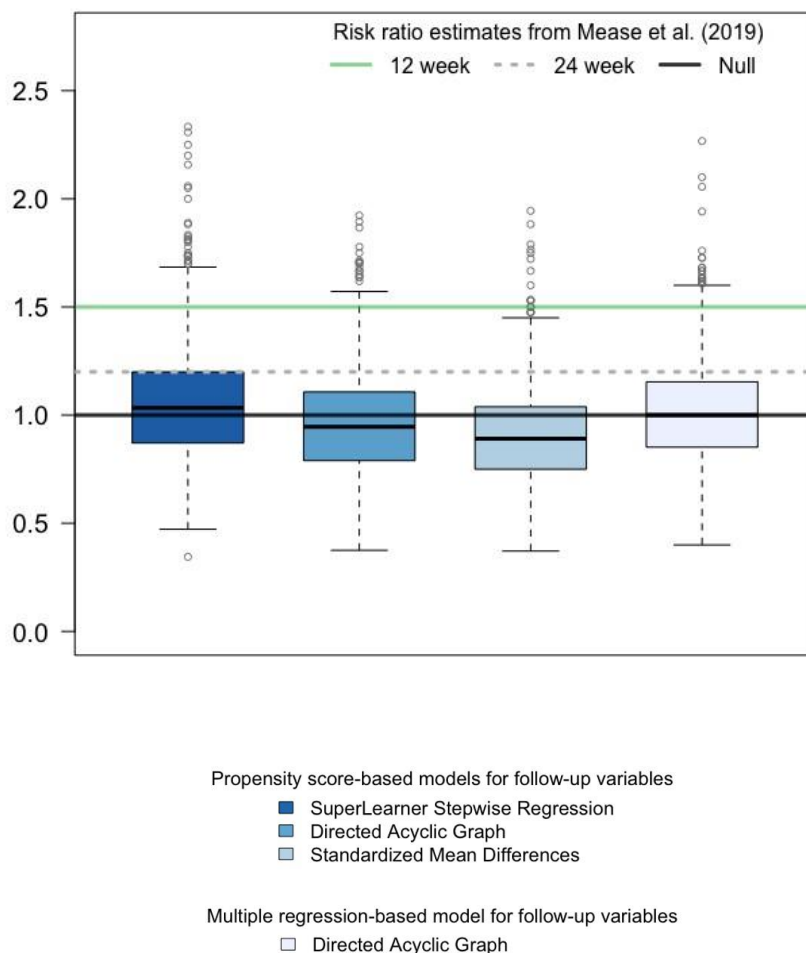


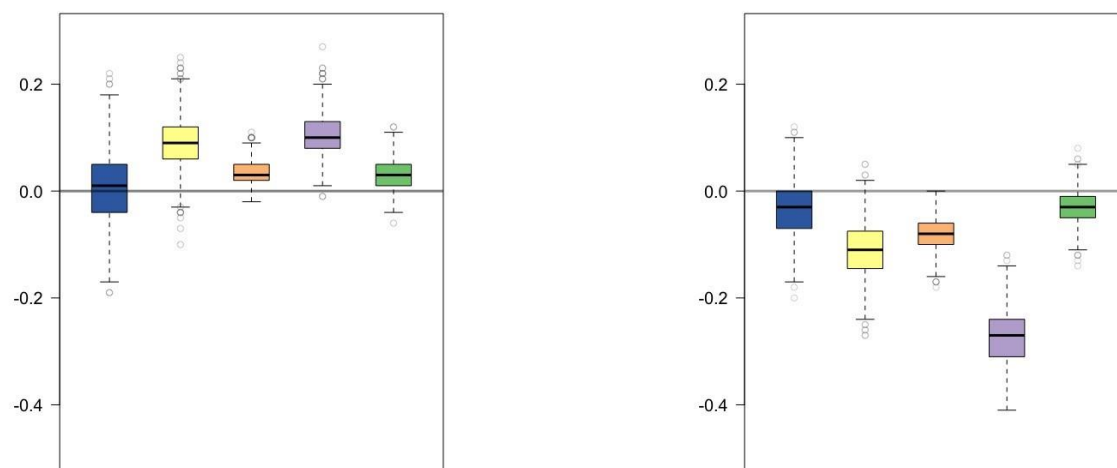
Figure 3. Risk differences for different primary endpoints based on 1000 simulations of Trials 1 and 2 using the stepwise regression propensity score-based models for follow-up variables.

A subset of follow-up variables (X_1) are used to derive different binary outcomes (Y), according to formulas provided in the Supplemental Material.

Risk differences are calculated as $RD = \Pr(Y | \text{TNF-}\alpha \text{ inhibitors}) - \Pr(Y | \text{methotrexate})$, where $RD > 0$ favor TNF- α inhibitors and $RD < 0$ favor methotrexate.

- (a) Trial 1, which enrolls “severe” patients with ≥ 3 tender joints and ≥ 3 swollen joints at baseline, tends to favor TNF- α inhibitors.

(b) Trial 2, which enrolls a heterogeneous group of patients (including both “severe” and “non-severe” individuals) tends to favor methotrexate.



- American College of Rheumatology 20% Response Criteria (ACR20)
- Psoriatic Arthritis Disease Activity Score (PASDAS)
- Disease Activity Index for Psoriatic Arthritis (DAPSA)
- Clinical DAPSA (cDAPSA)
- Routine Assessment of Patient Index Data 3 (RAPID3)