Research and Applications

An outcome model approach to transporting a randomized controlled trial results to a target population

Benjamin A. Goldstein,^{1,2} Matthew Phelan,² Neha J. Pagidipati,^{2,3} Rury R. Holman,⁴ Michael J. Pencina,^{1,2} and Elizabeth A. Stuart^{5,6}

¹Department of Biostatistics & Bioinformatics, Duke University, Durham, North Carolina, USA, ²Center for Predictive Medicine, Duke Clinical Research Institute, Durham, North Carolina, USA, ³Department of Medicine, Duke Clinical Research Institute, Center for Predictive Medicine, Duke University, Durham, North Carolina, USA, ⁴Endocrinology and Metabolism, University of Oxford, Oxford, United Kingdom, ⁵Department of Biostatistics John Hopkins University, Baltimore, Maryland, USA, and ⁶Department of Mental Health, John Hopkins University, Baltimore, Maryland, USA

Corresponding Author: Benjamin A. Goldstein, Department of Biostatistics & Bioinformatics, Duke University, 2525 Erwin Rd Suite 9023, Durham, NC 27705, USA (ben.goldstein@duke.edu).

Received 11 July 2018; Revised 12 November 2018; Editorial Decision 14 December 2018; Accepted 19 December 2018

ABSTRACT

Objective: Participants enrolled into randomized controlled trials (RCTs) often do not reflect real-world populations. Previous research in how best to transport RCT results to target populations has focused on weighting RCT data to look like the target data. Simulation work, however, has suggested that an outcome model approach may be preferable. Here, we describe such an approach using source data from the 2 × 2 factorial NAVIGATOR (Nateglinide And Valsartan in Impaired Glucose Tolerance Outcomes Research) trial, which evaluated the impact of valsartan and nateglinide on cardiovascular outcomes and new-onset diabetes in a prediabetic population.

Materials and Methods: Our target data consisted of people with prediabetes serviced at the Duke University Health System. We used random survival forests to develop separate outcome models for each of the 4 treatments, estimating the 5-year risk difference for progression to diabetes, and estimated the treatment effect in our local patient populations, as well as subpopulations, and compared the results with the traditional weighting approach.

Results: Our models suggested that the treatment effect for valsartan in our patient population was the same as in the trial, whereas for nateglinide treatment effect was stronger than observed in the original trial. Our effect estimates were more efficient than the weighting approach and we effectively estimated subgroup differences.

Conclusions: The described method represents a straightforward approach to efficiently transporting an RCT result to any target population.

Key words: electronic health records, machine learning, public health informatics, treatment heterogeneity

OBJECTIVE

Given good treatment compliance and minimal loss to follow-up, randomized controlled trials (RCTs) provide an internally valid estimate of a sample average treatment effect (SATE) for the evaluated intervention. If the RCT patient cohort is representative of the larger

patient population, one can typically extrapolate the SATE to the larger patient population, providing a population average treatment effect (PATE). Unfortunately, for both intentional and unintentional reasons, RCTs rarely fully reflect general patient populations.^{1–3} Because RCTs are designed to optimize internal, as opposed to external, validity, it is advantageous to recruit patients that are not

taking other medications, have fewer comorbidities, or are more likely to experience the primary outcome of interest. If the effectiveness of the intervention varies based on factors that differ between the RCT and general population (ie, there are effect modifications), then the PATE will not equal the SATE and the inference derived from an RCT may not be valid in different clinical populations.

While there has long been work on generalizing RCT results to overall populations,^{4–12} there has been increasing interest in transporting RCT results to smaller target populations, estimating what is referred to as the target average treatment effect (TATE).^{13,14} For both applications (generalizing and transporting), the majority of this work has aimed to account for the selection process into the trial. To do this, one estimates sample weights of the odds of a person being part of the RCT vs the target population. These weights are then used to make the RCT population look like the target population. These approaches have good theoretical properties and have been developed to incorporate a double-robust framework.¹⁴

The trial inclusion weights are estimated by “stacking” the RCT and target datasets. This means that one needs to combine the datasets together, leading to 2 potential drawbacks. First, every time results need to be transported to a new patient population, a new set of weights needs to be estimated. Second, the ability to estimate these weights accurately is partially driven by the sample size of the target population. When transporting to a single, large population (ie, generalization) these are not concerns.⁸ However, in our work we consider the target population to be patients served by the health system at a local institution. Amid changes to patient reimbursement, medical centers are becoming financially responsible for managing the health of their patient population.¹⁵ To manage a population cost effectively, health systems need to be able to reliably transport RCT results to their local, real world populations. This would provide an evidenced-based means to make clinical management decisions. Because each health system has different patient characteristics, any application to any new health system, or population subset of interest, the TATE will differ, requiring an estimation of a new set of weights. Moreover, there may be situations where the local population is small (eg, a single hospital or clinic). Accordingly, methods that are not dependent on the target sample size are desirable.

With the weighting approach limitations in mind, we consider an outcome model approach to transporting RCT results. The reasoning behind the outcome model approach, described in detail subsequently, is that prediction models are built among those receiving or not receiving the intervention and then the target population is “passed” through each model to produce potential intervention outcomes for each individual in the target population. This methodology allows individual treatment effects (ITEs) to be estimated which can be averaged to calculate the TATE. This approach resolves the 2 challenges itemized previously: the model only needs to be estimated once before applying to any target population, and that target population can be as small as desired. Kern et al¹⁰ and Dahabreh et al¹⁴ separately used simulation to compare weighting, doubly robust, and outcome model approaches and found that the outcome model approach had the best performance with respect to variance estimation and comparable results with respect to bias.

In this paper we build off this work to further develop a machine learning approach to estimate the TATE. To do so we incorporate the causal random forests (RF) framework.¹⁶ Our intent is to show how one can transport RCT results to a locally relevant target. We used source data from the 2 × 2 factorial-design NAVIGATOR (Nateglinide And Valsartan in Impaired Glucose Tolerance Out-

comes Research) international trial, which evaluated the impact of valsartan and nateglinide on cardiovascular outcomes and new-onset diabetes in 9306 prediabetic individuals. We then applied our results to prediabetic individuals found in the Duke University Health System (DUHS) electronic health record (EHR) system to estimate the TATE for this local patient population.

MATERIALS AND METHODS

We first describe the general analytic approach. Then, we describe the data used for analysis, both source and target. Last, we outline our evaluation.

Analytic approach

Figure 1 presents a schematic of the analytic approach. Our approach builds off of work by Lu et al¹⁶ on causal RF. The RF¹⁷ framework is an extension of Categorization And Regression Trees that combines multiple trees via a process called bagging (bootstrap aggregation) to create a more robust predictor. RF is a highly effective prediction model that has been used in a range of clinical studies, and is increasingly being used for causal inference.¹⁸ While we use RF for this work, we acknowledge that any machine learning method could be appropriate.

While Lu et al¹⁶ describe 7 different approaches for estimating the ITE and estimate in-sample effects, our focus is on out-of-sample effects. The approach we adopt here is referred to as counterfactual RF (cRF). In cRF, one identifies an outcome Y , treatments $T \in \{A, B\}$ and covariates W . One then splits the data between those that received the treatment A and those that received treatment B. Using the covariates W , one builds a model for the outcome separately for each treatment group. This generates 2 models:

$$\hat{m}_A \equiv E(Y|W_s, T = A); \hat{m}_B \equiv E(Y|W_s, T = B)$$

To estimate the ITE for a new observation (eg, someone from a target sample), one passes the observation through each model, generating predicted values under each condition:

$$\hat{Y}_{i,A} = m_A(W_{T,A}); \hat{Y}_{i,B} = m_B(W_{T,B})$$

We define the ITE as

$$\hat{\tau}_i = \hat{Y}_{i,A} - \hat{Y}_{i,B}$$

We then average over all $\hat{\tau}_i$ to get the TATE. We note that this is a modification of the approach by Lu et al,¹⁶ who utilize the out-of-bag sample of RF to estimate the ITE within the developmental data. The rationale behind the approach is that by generating 2 separate models, any heterogeneity is implicitly modeled, allowing the outcomes under each condition to freely differ. Moreover, as we illustrate in our application, when there are multiple potential treatments, it is straightforward to estimate the various contrasts of interest.

To generate standard errors for the TATE, we need to determine the source of variability. The previous procedure comprised 2 sets of data: an RCT source sample and a local target patient population. We use the RCT data to estimate $\hat{m}()$. We consider this sample random and consequently $\hat{m}()$ random. Conversely, we consider the target population fixed, and the TATE, once $\hat{m}()$ is estimated, fixed. Therefore, the primary source of variability comes from the estimation of $\hat{m}()$. To estimate this variability, we follow the approach of Lu et al¹⁶ and generate bootstrap samples of the data, refitting the RF models on each bootstrap and re-estimating the individual ITEs

Figure 1. Flow diagram for implementation of outcome approach. We start with randomized controlled trial (RCT) source data. Divide the data between those receiving the treatments of interest. Estimate separate outcome models. Identify the target data. Predict individual outcomes under each model. Calculate the individual treatment effect (ITE) and average to obtain the target averaged treatment effect (TATE). EHR: electronic health record.

and combined TATE. We use the estimated standard errors from the bootstrap distribution along with a normal approximation to generate a confidence interval. By generating standard errors in this way, the variability is not a function of the target sample size, only the source sample size.

Aligning the input data

An implicit component of this process is that the same W exist within both the source and target samples. Typically, RCTs have dozens of baseline covariates that are well defined and adjudicated. Conversely, EHR data typically have hundreds of covariate values that are not necessarily available for all patients. Moreover, similar measures are not necessarily equally defined. For example, within an RCT a glucose test result may be measured via fasting glucose while an EHR may contain a mixture of fasting and random glucose tests. As such, many of the same data elements may not exist in both data sources.¹⁹ Therefore, care is necessary to ensure that there is alignment between the input variables. Particularly, it may be necessary to remove W that are not present in both data sources, as we describe subsequently.

Analytic considerations

As discussed in detail by Dahabreh et al,¹⁴ there are a few considerations in this analysis. The first is that we primarily want to include variables that are related to the outcome of interest. The second is that we need to make sure the outcome model is well specified. This is where machine learning methods can provide assistance, by both “choosing” the variables to put into the model as well “discovering”

the best model form. A corollary to these assumptions—which is not verifiable—is that we have not missed any important variables. This is particularly important when we considering how best to align the input data. Another, often discussed consideration is whether the target sample was eligible for the trial, referred to as a positivity. As discussed by Dahabreh et al,¹⁴ this is primarily important for weighting-based approaches and not as much of a concern for outcome-based approaches. We assess this in our sensitivity analysis.

DATA

Source data

For our analysis we used data from the NAVIGATOR trial.²⁰ The NAVIGATOR trial was a 2×2 factorial-design trial comparing 2 medications, valsartan and nateglinide, in people with prediabetes. These medications were compared against each other, against a placebo, and in combination. In total there were 6 comparisons. Published results found that valsartan²¹ was effective in reducing the incidence of diabetes while nateglinide²² was not.

In the trial, of the 9306 participants, 2315, 2329, 2316, and 2346 were allocated to receive valsartan monotherapy, nateglinide monotherapy, valsartan-nateglinide combination therapy, and placebo, respectively. Baseline information was available across 46 clinical and demographic factors. Median follow-up was 6.5 years. We considered our primary endpoint to be the risk difference for new-onset diabetes at 5 years.

Figure 2. Consort diagram for identification of electronic health record–based target cohort. EHR: electronic health record; NAVIGATOR: Nateglinide And Valsartan in Impaired Glucose Tolerance Outcomes Research.

Target data

Our motivation for this work was to transport trial results to our institution's patient population. DUHS consists of 3 hospitals and a network of outpatient clinics. Additionally, we linked in data from a federally qualified health clinic, serving an under-served population. As the primary providers in Durham County, it is estimated that 85% of Durham County residents receive their primary care through the 2 systems.²³

To identify the local, target population we abstracted data from the DUHS EHR system (see Figure 2 for cohort diagram). The data used were abstracted from a PCORNet mapped datamart. We then further cleaned and assessed elements to see which mapped best to each other. To identify prediabetics we selected patients with a hemoglobin A1C between 5.7% and 6.4%, with the condition that they never had a previous result $\geq 6.5\%$. We considered the first encounter with an hemoglobin A1C in this range as the index date. We included patients seen between 2010 and 2016. To ensure that DUHS was a patient's medical home, we limited our analysis to individuals that lived in Durham County and had at least 2 encounters in the 2 years before the index date. We then defined a subcohort of patients who would have been eligibility for the NAVIGATOR trial. Specifically, we required individuals to have at least 1 cardiovascular risk factor (history of smoking, hypertension, left ventricular hypertrophy, microalbuminuria, reduced HDL [<40 mg/dL] or elevated LDL [>160 mg/dL]), excluded anyone currently on an angiotensin-converting enzyme inhibitor or angiotensin receptor blocker, as per the trial inclusion and exclusion criteria.

In total we were able to identify 35 (76%) (see Table 1) of the NAVIGATOR trial baseline characteristics within our EHR. This included demographic variables, vital signs, comorbidities, labs, and

medications. For variables with multiple measurements (eg, systolic blood pressure), we used the median of all measurements taken in the year before the index date. This has been shown to be a useful way to easily account for longitudinally measured clinical markers.^{24,25} Variables that we were not able to confidently identify within the EHR included: current smoking status (though smoking history was available), familial diabetes history, waist circumference, plasma glucose 2 hours after glucose load, and ratio of urinary albumin to creatinine. Because the occurrence of fasting-glucose was rare within our EHR, we used glycated hemoglobin A1C as an indicator of a prediabetic patient, a marker that has been shown to serve as a good proxy.²⁶

This was approved by our institution's institutional research board.

Analytic approach

Preliminary analyses

We compared the patient characteristics between the NAVIGATOR trial and DUHS patients using standardized mean differences. We consider an standardized mean differences >0.1 to indicate a meaningful difference between the 2 populations.²⁷

Primary analyses

We performed a series of analyses considering all 6 components of the 2×2 factorial design, but focusing on the valsartan and nateglinide comparisons vs placebo. We first estimated the SATE in the original NAVIGATOR trial, estimating the risk difference at 5 years for each of the 6 comparisons. We performed this analysis as an intent-to-treat analysis fitting a Kaplan-Meier estimator and then taking the effect estimate at 5 years. Next, we used the NAVIGATOR trial data to build a cRF for the valsartan, nateglinide, combination therapy and placebo cohorts separately. Specifically, we used random survival forests²⁸ to predict risk of diabetes at 5 years for each arm. We fit 500 trees within each forest. Using these forests we considered the 6 potential contrasts. We first used the forests to estimate the treatment effect among the NAVIGATOR trial sample. We used the out-of-bag sample from each bootstrap iteration to estimate the probability of diabetes, and then, for each contrast of interest, calculated the ITE and TATE, as described previously. We considered this analysis a test of internal validity of the method (ie, this analysis should replicate the estimate in the first analysis). Third, we applied the prediabetic individuals in the DUHS population, eligible for the NAVIGATOR trial, to the cRF to estimate the target sample risk difference. We estimated the risk difference as described previously. For each analysis, we performed 1000 bootstraps to estimate standard errors and calculate 95% confidence intervals (CIs).

As a secondary analysis, we considered the TATE across different subpopulations. Specifically, we considered differences based on: sex, race, and history of cardiovascular disease.

Sensitivity analysis

We performed 4 sensitivity analyses. First we expanded the cohort to all prediabetic individuals regardless of whether they would have met criteria for the NAVIGATOR trial. Second, we compared our analysis to the more traditional weighting approach, following the approach outlined in Westreich et al.¹³ We used the baseline characteristics to estimate selection weights. Given our sample size we used all available baseline variables, using a RF estimated probability weights. We then performed a weighted regression to estimate the risk difference and 1000 bootstraps to generate standard errors.

Table 1. Differences in baseline characteristics of target and source populations

Characteristic	Local Prediabetic Population (n = 12 132)	NAVIGATOR Trial Sample (n = 9306)	Standardized Mean Difference
Age, y	51.3 (40.8–61.6)	63.0 (58.0–69.0)	–1.025
Female	7595 (62.6)	4711 (50.6)	0.244
Race			
Asian	334 (2.8)	613 (6.6)	–0.182
Black	6259 (51.6)	236 (2.5)	1.324
Other	1633 (13.5)	723 (7.8)	0.186
White	3906 (32.2)	7734 (83.1)	–1.202
Weight, kg	89.1 (75.3–106)	82.0 (71.5–93.5)	0.422
BMI, kg/m ²	31.9 (27.5–37.9)	29.7 (26.8–33.3)	0.041
SBP, mm Hg	128 (120–139)	140 (128–150)	–0.568
DBP, mm Hg	79.0 (73.0–85.0)	84.0 (78.0–90.0)	–0.511
Cardiovascular risk factors			
Any	12 132 (100.0)	8921 (95.9)	0.294
Smoking history	5947 (49.0)	1025 (11.0)	0.911
Hypertension	7728 (63.7)	7216 (77.5)	–0.307
Left ventricular hypertrophy	513 (4.2)	268 (2.9)	0.073
Microalbuminuria	139 (1.1)	114 (1.2)	–0.007
History of cardiovascular disease			
Any	1300 (10.7)	2745 (29.5)	–0.482
Myocardial infarction	177 (1.5)	1103 (11.9)	–0.426
Angina	199 (1.6)	1561 (16.8)	–0.542
Percutaneous coronary intervention	296 (2.4)	622 (6.7)	–0.204
Coronary artery bypass grafting	11 (0.1)	521 (5.6)	–0.336
Intermittent claudication	251 (2.1)	98 (1.1)	0.082
Lower-limb angioplasty	6 (0.0)	110 (1.2)	–0.145
Nontraumatic leg or foot amputation	6 (0.0)	7 (0.1)	–0.01
Stroke	799 (6.6)	374 (4.0)	0.115
Labs at time of index date			
Fasting glucose, mmol/L	5.5 (5.0–5.9)	6.1 (5.7–6.4)	–0.591
A1C, %	5.9 (5.8–6.1)	5.8 (5.5–6.1)	0.4
Total cholesterol, mg/dL	189 (163–217)	207 (181–236)	–0.429
HDL-C, mg/dL	44 (37–54)	48 (40–57)	–0.239
LDL-C, mg/dL	116 (93–141)	124 (100–150)	–0.219
Creatinine, mg/dL	0.9 (0.7–1.0)	0.8 (0.7–1.0)	0.135
eGFR, mL/min/1.73 m ²	92 (76–108)	80 (69–91)	0.296
eGFR <60 mL/min/1.73 m ²	924 (8.8)	1025 (11.1)	0.286
Medications at time of index date			
ACE inhibitor	0 (0.0)	676 (7.3)	–0.396
Angiotensin receptor blocker	0 (0.0)	30 (0.3)	–0.08
Alpha-blocker	330 (2.7)	577 (6.2)	–0.169
Aspirin or other antiplatelet drug	2730 (22.5)	3425 (36.8)	–0.08
Beta-blocker	1622 (13.4)	3666 (39.4)	–0.618
Calcium-channel blocker	1764 (14.5)	3012 (32.4)	–0.43
Diuretic	2526 (20.8)	2960 (31.8)	–0.251
Lipid-modulating drug	3283 (27.1)	3577 (38.4)	–0.244
Any medication	6157 (50.8)	7794 (83.8)	–0.751

Values are median (interquartile range) or n (%).

ACE: angiotensin-converting enzyme; BMI: body mass index; DBP: diastolic blood pressure; eGFR: estimated glomerular filtration rate; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; NAVIGATOR: Nateglinide And Valsartan in Impaired Glucose Tolerance Outcomes Research; SBP: systolic blood pressure.

Third we assessed the impact of the variables we were unable to map from the RCT to the EHR. We built outcome models using all of the available RCT variables, and compared model fit—via the C-statistic—to the model using only the mappable variables. Finally, we considered the impact of algorithm choice. The above approach can be used by any machine learning algorithm. As a comparison, we fit the outcome models using LASSO regularized regression appropriate for time-to-event data.²⁹ We compared both the model fits as described previously as well as the inference from the LASSO-based fits.

All analyses were performed in R 3.4.2. The RF model was estimated using the package randomSurvivalForests.³⁰ This work was approved by our institution's institutional research board.

RESULTS

We identified 20 068 prediabetic patients in the DUHS EHR system, 12 132 of whom would have been eligible for the NAVIGATOR trial. Table 1 shows the comparison of the source RCT and target

EHR populations. In general, the RCT population was sicker with higher systolic blood pressure, more comorbidities, and more medication prescriptions. Of note the 2 groups had comparable Hemoglobin A1c, an important marker for prediabetes.

Assessment of valsartan and nateglinide

We first assessed the independent effect of valsartan vs placebo (Table 2). Using the original NAVIGATOR trial data we estimated a significant effect for the risk difference at 5 years (-0.056 ; 95% CI, -0.085 to -0.027), as in the originally published RCT. We were able to replicate this result when retransporting the cRF result back to the NAVIGATOR trial patient population (-0.051 ; 95% CI, -0.073 to -0.028). Finally, when transporting the effect to the Duke EHR-defined population, we get a similar, if not slightly stronger, effect estimate (-0.068 ; 95% CI, -0.119 to -0.016).

We next assessed the independent effect of nateglinide vs placebo. The original RCT reported a null effect, which we were able to confirm using the source data (0.009 ; 95% CI, -0.021 to 0.039) as well as through our internal validation of transporting back to the RCT sample (-0.004 ; 95% CI, -0.020 to 0.013). However, when transporting the effect to Duke EHR defined population we get a stronger, if not statistically significant effect (-0.030 ; 95% CI, -0.081 to 0.019).

We also assessed the additional comparisons of the 2 active drugs with each other as well as combination therapy vs monotherapy (Table 2). Of note, the original RCT found valsartan to be superior to nateglinide, which was confirmed in our reanalysis of the risk difference (-0.065 ; 95% CI, -0.094 to -0.036). However, when transporting the effect to the Duke patient population, this effect is attenuated and is no longer statistically significant (-0.038 ; 95% CI, -0.084 to 0.013), likely due to the larger effect size of nateglinide vs placebo.

Subgroup analyses

To illustrate the flexibility of the approach, we examined the treatment effect in different subpopulations of our target populations (Figure 3). In general, we found similar treatment effects among the different subpopulations. One notable exception was the treatment effect for nateglinide among Caucasians (-0.013 ; 95% CI, -0.050 to 0.024) vs African Americans (-0.042 95% CI, -0.098 to 0.012). Because there is a greater proportion of African Americans in the Duke patient population, compared with the NAVIGATOR trial sample, it is possible that this difference is what accounts for the overall different observed treatment effect for nateglinide.

Sensitivity analyses

As a sensitivity analysis we expanded the cohort to all prediabetic patients at DUHS ($n = 20\ 068$). The results were very similar to the full sample results (Table 2). We next compared our findings to the more commonly applied sample weighting approach. We found transported treatment effects of -0.028 (95% CI, -0.134 to 0.079) and 0.022 (95% CI, -0.082 to 0.122) for the valsartan vs placebo and nateglinide vs placebo comparisons, respectively (Table 2). Of particular note are large standard-errors. The average standard error for the outcome model was 0.02 vs 0.05 for the weighting approach. We next considered model fit when using all available NAVIGATOR trial variables or just those mappable to the EHR. We found the C-statistics were nearly identical—average of 0.647 across the 4 arms—suggesting that there was little lost from the dropped variables. Last, we compared the results if we used regularized regression

(LASSO) to model the outcomes. While the transported results were not identical, the inference was consistent (Supplementary Table 2). We note that the LASSO model had a slightly worse area under the curve of 0.640 .

DISCUSSION

We have illustrated an approach to transport a treatment effect from an RCT to a target population. In contrast to much of the literature in this area, which focuses on weighting the source data, we focus on developing an outcome model. This allows us to treat this as a 2-step process. Decoupling the source and target populations results in 2 key advantages. First, we are able to easily transport the result to any sample target or treatment contrast of interest, reusing the model fit for each arm of the trial. Second, the ability to transport the results is not dependent on the sample size of the target population.

Results for the NAVIGATOR sample

In comparing the RCT sample to our local patient population, we noted that the RCT population was generally sicker, with higher blood pressure, more comorbidities, and more medications. This confirms previous literature that has shown that RCT samples are generally sicker than the general population.^{3,31} Interestingly, other work has suggested that EHR-based populations are sicker than general clinical populations³² suggesting that these differences may be an underestimate, and in fact differences may be even more extreme when compared with a general patient population.

Upon transporting the valsartan vs placebo comparison to the DUHS-based population, we found an effect estimate similar to the one reported in the original RCT. Conversely, while the RCT showed no treatment benefit for the nateglinide vs placebo comparison, the TATE did suggest some treatment effect—though not statistically significant. This result followed through when we considered some of the comparisons based on combination therapy. For example, the SATE for valsartan + nateglinide vs just nateglinide showed a significant treatment effect, while the TATE did not.

General lessons for transportability

An important feature of the outcome model approach is that the estimation of the outcome model is decoupled from the transporting step. This means that an RCT can generate an outcome model without knowledge of the target population. Moreover, because we are relying on the RCT data to generate the outcome model, this approach will not be impacted by underlying biases due to informative visits within EHR data.³³ Instead, all that is required is the ability to map the covariates in the target population back to the trial data—admittedly not always an easy task. This feature is what makes transporting to additional subpopulations more efficient. When we applied our analysis to different subpopulations, we found similar treatment effects across the subgroups. The one exception was an indication of a heterogeneous racial effect among those taking nateglinide. We note that the original trial did not report a significant effect for racial heterogeneity ($P = .82$). This suggests that this effect differences may not be due solely to racial differences, but higher-order interaction effects. Detecting these higher-order effects is possible by the use of machine learning methods.

Embedded in our analysis is the ability to check the internal validity of the outcome model. By leveraging the out-of-bag samples from the RF bootstrap iterations, we reapplied the outcome model

Table 2. Risk difference at 5 years across different treatment comparisons and populations

	Valsartan vs Placebo	Nateglinide vs Placebo	Valsartan vs Nateglinide	Valsartan + Nateglinide vs Placebo	Valsartan + Nateglinide vs Valsartan	Valsartan + Nateglinide vs Nateglinide
Sample average treatment effect in NAVIGATOR trial	-0.056 (-0.085 to -0.027)	0.009 (-0.021 to 0.039)	-0.065 (-0.094 to -0.036)	-0.036 (-0.065 to -0.007)	0.020 (-0.009 to 0.048)	-0.045 (-0.075 to -0.016)
Retranslation to NAVIGATOR trial sample ^a	-0.051 (-0.073 to -0.028)	-0.004 (-0.020 to 0.013)	-0.047 (-0.064 to -0.032)	-0.047 (-0.063 to -0.031)	0.004 (-0.012 to 0.021)	-0.043 (-0.067 to -0.021)
Transporting to local prediabetic population eligible for NAVIGATOR trial ^a	-0.068 (-0.119 to -0.014)	-0.03 (-0.081 to 0.019)	-0.038 (-0.084 to 0.013)	-0.038 (-0.084 to 0.013)	0.031 (-0.023 to 0.079)	-0.007 (-0.051 to 0.040)
Transporting to full local prediabetic population ^a	-0.069 (-0.119 to -0.016)	-0.030 (-0.077 to 0.016)	-0.039 (-0.085 to 0.010)	-0.043 (-0.086 to 0.004)	0.025 (-0.025 to 0.072)	-0.013 (-0.056 to 0.030)
Transporting to local NAVIGATOR trial population using weighting to local population	-0.028 (-0.134 to 0.079)	0.022 (-0.082 to 0.122)	-0.048 (-0.152 to 0.050)	-0.030 (-0.136 to 0.069)	-0.002 (-0.099 to 0.106)	-0.051 (-0.151 to 0.043)

Values are risk differences (95% confidence intervals).

NAVIGATOR: Nateglinide And Valsartan in Impaired Glucose Tolerance Outcomes Research.

^aEach of these methods used the counterfactual random forests approach for estimation.

Figure 3. Effect estimates for different patient subgroups. CVD: cardiovascular disease.

to the RCT population. We were able to find similar effect estimates as the original trial, suggesting that our approach is relatively stable. We suggest that this should be a standard first step when applying such an analysis.

Another important consideration is how best to handle transportability when the source and target samples differ from one another. As discussed by others, weighting-based approaches can break down when there is a lack of overlap, known as positivity violations, leading to extreme weights.¹⁴ Outcome-based approaches avoid this by extrapolating the results of the outcome model. This is another area in which RF serves as an ideal machine learning algorithm. Because RF is based on trees, it is more robust to outlying values, and consequently extrapolation, from the target sample. Unlike linear models, trees do not extrapolate predictions when the observed covariates extend beyond the support of the source data, instead truncating the predictions. This is an important consideration because, as we noted, there were meaningful differences between the RCT and DUHS samples. In general the question of how to estimate the TATE when the source and target samples differ is worthy of more future research.

Finally, we note that when we compared our outcome model approach with the more typical weighting approach, we found our effect estimates to have meaningfully smaller standard errors. This empirical finding confirms the simulation results of others.^{10,14} One intuition for why this is the case is that the estimation of sample weights relies on both the target and source sample while the proposed approach only relies on the source data. Moreover, RF is a relatively stable predictor that has low predictive variance.³⁴

Limitations and future work

There are some notable limitations in our analysis and analytic approach. First, these are the results of one analysis. This approach should be tried with different source and target samples to note any additional potential complications. It would be interesting to compare the TATE in different clinical populations to assess how much potential clinical heterogeneity exists. Moreover, in our analysis we had a relatively large RCT from which to work. It is worth investi-

gating what is the minimum size under which such an approach will still provide valid and efficient estimates.

More generally, additional work in this field should consider the effects of transporting trial results outside the support of the trial data. Trees appear to be a particularly valuable target for this work. However, there are important limitations of such analytic approaches when the RCT and target populations differ. To the extent that RCTs do not represent target populations of interest, analytic methods will be limited in their ability to derive targeted treatment effects. Ultimately, it is important for RCTs to be better designed to reflect real world populations.³⁵ Finally, a key challenge in this area is what to do when variables are not directly comparable between the source and target data. This is an especially important concern with EHR data where clinical factors are measured in different ways. Hong et al⁶ have done some work in this area, showing how multiple imputation can be used for missing individual values. In our application there was minimal impact of the dropped variables, but more work is needed to understand how best to handle, observed, but unmappable, covariates.³⁶

CONCLUSIONS

As health systems become increasingly responsible for managing the health of their patient populations, it is important for them to be able make evidence-based treatment decisions. While RCTs provide gold standard evidence, due to demographic and clinical heterogeneity, the average treatment effects that they estimate are not always the best reflection of optimal treatment strategy. In this analysis, we illustrate a means to transport an RCT result to a target population. By using an outcome model approach, we are able to efficiently transport the results to target populations of different sizes and compositions.

FUNDING

This work was supported by National Institute of Diabetes and Digestive and Kidney Diseases career development award K25 DK097279 (to BAG), U.S. Department of Education Institute of Education Sciences grant

R305D150003 (to EAS). The project described was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Award Number UL1TR001117 at Duke University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTORS

BAG designed the analysis and drafted the manuscript. MP performed the analyses. NJP, RRH, MJP, and EAS edited the manuscript and provided critical feedback. All authors approve of the final version of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Conflict of interest statement. None declared.

REFERENCES

- Kalata P, Martus P, Zettl H, *et al.* Differences between clinical trial participants and patients in a population-based registry: the German Rectal Cancer Study vs. the Rostock Cancer Registry. *Dis Colon Rectum* 2009; 52 (3): 425–37.
- Mosenifar Z. Population issues in clinical trials. *Proc Am Thorac Soc* 2007; 4 (2): 185–7; discussion 187–188.
- Weng C, Li Y, Ryan P, *et al.* A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. *Appl Clin Inform* 2014; 5 (2): 463–79.
- Buchanan AL, Hudgens MG, Cole SR, *et al.* Generalizing evidence from randomized trials using inverse probability of sampling weights. *J R Stat Soc Ser A Stat Soc* 2018; 181 (4): 1193–209.
- Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am J Epidemiol* 2010; 172 (1): 107–15.
- Hong J-L, Jonsson Funk M, LoCasale R, *et al.* Generalizing randomized clinical trial results: implementation and challenges related to missing data in the target population. *Am J Epidemiol* 2018; 187 (4): 817–27.
- Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. *J R Stat Soc Ser A Stat Soc* 2001; 174 (2): 369–86.
- Stuart EA, Ackerman B, Westreich D. Generalizability of randomized trial results to target populations: design and analysis possibilities. *Res Soc Work Pract* 2018; 28 (5): 532–7. doi: 104973151772073.
- Wang W, Ma Y, Huang Y, Chen H. Generalizability analysis for clinical trials: a simulation study. *Stat Med* 2017; 36 (10): 1523–31.
- Kern HL, Stuart EA, Hill J, Green DP. Assessing methods for generalizing experimental impact estimates to target populations. *J Res Educ Eff* 2016; 9 (1): 103–27.
- Flather M, Delahunty N, Collinson J. Generalizing results of randomized trials to clinical practice: reliability and cautions. *Clin Trials* 2006; 3 (6): 508–12.
- Pressler TR, Kaizar EE. The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. *Statist Med* 2013; 32 (20): 3552–68.
- Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. *Am J Epidemiol* 2017; 186 (8): 1010–4.
- Dahabreh IJ, Robertson SE, Tchetgen EJT, Stuart EA, Hernán MA. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* 2018; doi: 10.1111/biom.13009. [Epub ahead of print] PMID: 30488513.
- Clough JD, McClellan M. Implementing MACRA: implications for physicians and for physician leadership. *JAMA* 2016; 315 (22): 2397–8.
- Lu M, Sadiq S, Feaster DJ, Ishwaran H. Estimating individual treatment effect in observational data using random forest methods. *J Comput Graph Stat* 2018; 27 (1): 209–19.
- Breiman L. Random forests. *Mach Learn* 2001; 45 (1): 5–32.
- Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc* 2018; 113: 1228–42.
- Köpcke F, Trinczek B, Majeed RW, *et al.* Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. *BMC Med Inform Decis Mak* 2013; 13: 37.
- Califf RM, Boolell M, Haffner SM, *et al.* Prevention of diabetes and cardiovascular disease in patients with impaired glucose tolerance: rationale and design of the Nateglinide and Valsartan in Impaired Glucose Tolerance Outcomes Research (NAVIGATOR) Trial. *Am Heart J* 2008; 156 (4): 623–32.
- NAVIGATOR Study Group, McMurray JJ, Holman RR, *et al.* Effect of valsartan on the incidence of diabetes and cardiovascular events. *N Engl J Med* 2010; 362 (16): 1477–90.
- NAVIGATOR Study Group, Holman RR, Haffner SM, *et al.* Effect of nateglinide on the incidence of diabetes and cardiovascular events. *N Engl J Med* 2010; 362 (16): 1463–76.
- Miranda ML, Ferranti J, Strauss B, Neelon B, Califf RM. Geographic health information systems: a platform to support the “triple aim.” *Health Aff (Millwood)* 2013; 32 (9): 1608–15.
- Goldstein BA, Pomann GM, Winkelmayer WC, Pencina MJ. A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis. *Statist Med* 2017; 36 (17): 2750–63.
- Sweeting MJ, Barrett JK, Thompson SG, Wood AM. The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study. *Statist Med* 2017; 36 (28): 4514–28.
- Vistisen D, Witte DR, Brunner EJ, *et al.* Risk of cardiovascular disease and death in individuals with prediabetes defined by different criteria: The Whitehall II Study. *Diabetes Care* 2018; 41 (4): 899–906.
- Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat* 1985; 39 (1): 33–8.
- Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat* 2008; 2 (3): 841–60.
- Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox’s proportional hazards model via coordinate descent. *J Stat Softw* 2011; 39 (5): 1–13.
- Ishwaran H, Kogalur UB. Random survival forests for R. *R News* 2007; 7 (2): 25–31.
- He Z, Wang S, Borhanian E, Weng C. Assessing the collective population representativeness of related type 2 diabetes trials by combining public data from ClinicalTrials.gov and NHANES. *Stud Health Technol Inform* 2015; 216: 569–73.
- Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annu Symp Proc* 2013; 2013: 1472–7.
- Phelan M, Bhavsar NA, Goldstein BA. Illustrating informed presence bias in electronic health records data: how patient interactions with a health system can impact inference. *EGEMS (Wash DC)* 2017; 5 (1): 22.
- Hastie T, Tibshirani R, Friedman J. *Elements of Statistical Learning*. 2nd ed. New York: Springer; 2009.
- Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 2015; 16: 495.
- Nguyen TQ, Ebnesajjad C, Cole SR, Stuart EA. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Ann Appl Stat* 2017; 11 (1): 225–47.