X Marks the Spot: Mapping Similarity Between Clinical Trial Cohorts and US Counties

Matthew C. Lenert, BA[1], Dara E. Mize, MD[2], Colin G. Walsh, MD MA (Primary Advisor) [2]

[1]Vanderbilt University, Nashville, TN; [2]Vanderbilt University Medical Center, Nashville, TN

Abstract

When patients and doctors collaborate to make healthcare decisions, they rely on clinical trial results to guide discussions. Trials are designed to recruit diverse participants. The question remains – how well do trial results apply to me or to people who live in our area? This study compared one complete clinical trial dataset (SPRINT) and one published study (ACCORD) to the Community Health Status Indicators dataset to assess the similarity of the trial populations to US county populations. Counties up to 495 miles to the closest SPRINT trial site and up to 712 miles to the closest ACCORD trial site had populations that were significantly more similar to the study cohort than counties farther away. The investigators detail a generalizable method for both assessing recruitment gaps in large multicenter trials and creating maps for clinicians to provide intuition on trial applicability in their area.

Background

The growing movement towards patient-centered care is an effort to improve the quality of health care delivery, thereby leading to a greater focus on shared decision making and evidence-based medicine[1]. Under the shared decision-making model of patient care, it is the role of the clinician to present the best applicable clinical evidence. The role of the patient is to share their preferences and values. The physician then seeks to apply their clinical judgment to determine which therapies, diagnostic tests, and/or preventative services best fit their patient[2,3]. Randomized clinical trial (RCT) data often serve as the foundation of clinical evidence on which these decisions are made[4,5].

The amount of data available for clinicians to digest continues to increase and integrating the latest study at the beside or in clinic poses a challenge for practicing providers[6,7]. It is difficult to know – in practice and in real-time – how well a trial's results apply to one's patients. This task is further complicated with pervasive difficulties in determining the validity of study findings[8]. For example, the ratio of screened patients excluded from randomization has direct implications for the significance of trial results[9]. Studies may report surrogate outcomes that do not correlate with patient-valued outcomes (e.g. reporting changes in bone density for patients with postmenopausal osteoporosis is not the same as measuring rates of long-bone fractures)[10]. Treatment effects can be nonlinear across disease states, making the average effect deceptive[11]. Providers often look to systemic reviews and meta-analyses as tools to consider these nuances.

Meta-analyses aggregate evidence by summarizing results to estimate an overall trend across many sample populations, but they have some stark limitations[12]. It is difficult to produce a high quality meta-analysis, because one has to account for studies of varying external and internal validity in their designs. This heterogeneous mixture makes it difficult to focus on the scientific scale, and many analyses instead focus on the clinically removed statistical scale[13]. These limitations are the same limitations faced by providers, so it is not surprising that they are strong confounders.

Instead of pooling data from many small trials, there is a renewed push for building infrastructure to support large pragmatic-randomized-multicenter trials comparing an experimental treatment to the standard of care[14]. These trials recruit patients across the country and are carefully crafted to control for confounders. Despite accounting for population differences through the use of many sites, selection bias persists[15]. This bias affects the generalizability of trial results to the population at large, because the number of patients that participate in clinical trials may be a fraction of those seeking routine care[16,17]. Even in trials of tens of thousands of patients, trial samples may demonstrate differences from the population at large. Differences can manifest themselves in comparative functional status, age, and the number of comorbidities[18]. Trial results should hold for patients who are similar to trial participants, but there may be complex screening criteria or differences in subgroups that complicate comparisons[19]. Even at the highest level, measuring applicability of a study to the population at large remains a grand challenge.

A method to quantify similarity would help address the applicability gap. Researchers have used multiple definitions for patient similarity. The IBM TJ Watson group developed systems that combine supervised and

unsupervised approaches[20, 21]. Their system applies decentralized physician feedback to improve upon the base measure of similarity—in this case the Euclidean distance of a vector of clinical features for a pair of patients. The feature vector was made up of clinical variables such as International Classification of Disease (ICD) codes and objective clinical readings (e.g. blood pressure). Lee et al. took a similar approach, where they selected specific clinical features and put them into vector representation[22]. This step enabled them to apply the cosine distance metric to quantify pairwise similarity. Zhang defined similarity as the Jaccard similarity over the set of ICD codes in the problem list between two patients[23]. Panahiazar et al. used k-means clustering to quantify similarity as the Mahalanobis distance between a patient's feature vector and the centroids of clustered training patients[24]. There has been some exploration of making trial characteristic comparisons directly. In a small study, Cahan and Cimino devised a model that visualized where a patient would fall within the distribution of clinical and demographic features of an RCT cohort[25].

Most of these efforts focused on comparing a single patient to another patient or cluster of patients. A pairwise comparison is not ideally fitted for the problem at hand, since clinical trials tend to summarize characteristics of the cohort as well as the treatment effect. A clinical trial cohort has a distribution of clinical characteristics, while the individual patient has discrete characteristics. For example, the prevalence of smoking in the trial cohort might be 15%, but an individual currently smokes or does not. This suggests that it may be more tractable to compare the characteristics between two groups. Abstracting to the group level may lead to insights into how similar trial patients are to the local population. The investigators hypothesized that clinical trials record sufficient demographic, comorbidity, and adverse event data to readily produce a means of comparison with U.S. county-level health statistics. A quantifiable measure could then be mapped to give physicians an expedient intuition into how an RCT's population compares to the people they see day to day. This measure may not sway a provider's decision for a specific patient, but it may inform the (prior) probability they would recommend a specific therapy in the first place.

Methods

Data Sources

This study analyzed the complete dataset from the SPRINT Group's blood pressure trial as well as the published data from the ACCORD Group's type 2 diabetes trial[26,27]. This study received IRB approval and completed a data use agreement to gain access to the SPRINT trial data. The SPRINT trial was a large pragmatic RCT that involved 9,361 patients across 96 clinical sites in the United States and 6 sites in Puerto Rico. This trial compared intensive blood pressure control to standard of care for patients with hypertension. The trial controlled for comorbidities such as chronic kidney disease, and several medication classes such as statins. The trial followed a clinically relevant primary outcome of myocardial infarction (MI), non-MI acute coronary syndrome, stroke, heart failure, or death attributable to cardiovascular disease. The SPRINT Group followed patients for a median of 3.26 years. The SPRINT trial planned demographic subgroups of: African-Americans, non-African Americans, ages less than 75, ages greater than or equal to 75, chronic kidney disease (CKD) present at the start of the trial, and CKD absent. The SPRINT trial is largely regarded as a model multi-site RCT for both its rigorous design and its methodical implementation.

The ACCORD group's trial preceded the SPRINT trial by eight years and studied the effect of intensive control of glycated hemoglobin levels in patients with type 2 diabetes. The trial was a multi-site pragmatic trial that randomized 10,251 patients across 62 clinical sites in the United States and 15 sites in Canada. The ACCORD trial limited comorbidities and controlled for relevant demographic factors for cardiovascular and diabetic complications. The researchers also controlled for several medication classes such as beta-blockers and glucophages. The ACCORD trial measured the primary outcome of nonfatal myocardial infarction, nonfatal stroke, or death from cardiovascular causes. This trial planned for the following demographic subgroup analyses: age less than 65, age greater than or equal to 65, Caucasian, non-Caucasian, males, and females. The ACCORD group had an average follow-up period of 3.5 years. The ACCORD trial is another model exemplar of a pragmatic multi-site RCT.

The county health statistics came from the Community Health Status Indicators (CHSI) Dataset published by the Centers for Disease Control and Prevention (CDC). The CHSI dataset contains county-level social, environmental, and prevalence health data[28]. This dataset is composed of a variety of CDC and Census Bureau studies from 2002 through 2014. It covers a wide array of disorders, diseases, and social behaviors. Its wide coverage of demographic and disease information at the county-level made it an ideal choice for this application. The investigators looked to map demographic and outcome data from the SPRINT and ACCORD trials to demographic and prevalence statistics from the counties based on the hypothesis that patients from a similar population should display similar features.

Mapping Between Data Sources

Mapping variables from one data set to another requires input from clinical experts to ensure that nuances from the trial variables are accounted for and that the comparisons make clinical sense. The investigators established that a CHSI variable was comparable to a trial variable if and only if both physician authors agreed the comparison was reasonable. Demographic features such as obesity, smoking prevalence, and African-American race, mapped easily between the CHSI data and the SPRINT trial. The investigators assumed correlation between incidence rates of adverse events in the SPRINT control group and age-adjusted death rates in the CHSI county data. Outcomes of control patients were averaged to compare the prevalence of cardiac event and death, as well as stroke and death from the SPRINT control group to Coronary Heart Disease (CHD) death rates and stroke death rates in the CHSI data.

For the ACCORD trial the authors mapped the demographic features of African-American race, Hispanic race, and smoking status. The authors again used CHD death rates from CHSI and cardiac related deaths from the published data. The authors did not have access to the full ACCORD trial data; which prevented the inclusion of more features, such as the prevalence of obesity.

After selecting the variables for comparison the investigators needed to perform additional processing of the CHSI data set to ensure as fair a comparison was made as possible. Direct comparison between CHSI and trial variables is not advised, as trial participants are generally recruited conditional on the presence of some disease, disorder, or clinical state. For example, the SPRINT trial only recruited patients who had hypertension and the ACCORD trial only recruited patients who had type 2 diabetes. Patients with type 2 diabetes and patients with hypertension are subgroups of the general population, which may have markedly different characteristics.

Missing Data

The investigators did not find missing values in the CHSI data to be predictably missing (missing not at random). Therefore, the authors assumed that values in the data were missing at random. Predictably missing variables suggest bias in how the data were collected or how participants responded. Mean imputation was used to fill in values for variables with less than 3% (60 of 3,141 counties) of values missing. The authors used this threshold because the cost of computation time and complexity outweighed inferential benefits. The investigators used the mean of the state for mean imputation of missing county values.

Multiple imputation was used to fill in missing county values for variables with more than 3% of the feature missing. In multiple imputation, one uses a model to predict what the missing value would have been based on data one does have for that case. The investigators built one model per variable that had missing values. A bootstrapped elastic net (mixture of L1 and L2 penalty) penalized regression was used to select the features for each imputation model. Each model used all fully recorded demographic variables, features that were completed with mean imputation, quadratic, cubic, and log transforms of all features, and all interaction terms of first order features. The hyper-parameters (penalization for small regression coefficients and the ratio of L1 regularization to L2 regularization) of the elastic net regression were re-tuned for each bootstrapped sample over 200 bootstraps. The investigators optimized the hyper-parameters by minimizing the log loss of the model on the "out of bag" data. Once the most significant features were selected through the bootstrapped elastic net, the investigators needed to determine the effect or weights of those features before being able to predict the missing values. This required fitting the models over another 200 bootstraps where hyper-parameters were again retuned on each bootstrap. The resulting effects (coefficients) are the average of the coefficients for the 200 bootstrapped models. The variance of the imputed coefficients is the calculated variance from those 200 bootstrapped models.

Conditioning on Disorders

All CHSI variables required probabilistic conditioning on hypertension for the SPRINT trial and type 2 diabetes for the ACCORD trial. This step is crucial because all SPRINT trial participants have hypertension and all ACCORD trial participants have type 2 diabetes. The CHSI dataset is made up of multiple county averages over time along with the variance of that measure. The investigators used these moments to define a normal distribution for the county average of the various mapped variables (e.g. stroke deaths). The investigators needed to assume that each mapped variable was normally and identically distributed within a county for the years 2003-2014. This assumption means statistics over different years for the same county are independent and identically distributed. This assumption was necessary to leverage a correlated bivariate normal distribution to inform the conditional prevalence of all variables given hypertension or type 2 diabetes using the statistical identity $f_{X|Y} = f_{XY} / f_Y$. The investigators used the stated distributional assumptions to transform the general population statistics into condition/disease specific statistics. The correlated bivariate normal distribution has a correlation parameter. The investigators estimated the correlation parameter using a national average correlation of each variable and the conditioned variable (e.g. the correlation between obesity and hypertension).

Measuring Similarity

Similar to previous work, the investigators used feature vector representation to enable comparisons between a county and the trial population. Similarity was defined as the cosine similarity between a county feature vector and the average of the trial controls. For the SPRINT study, the feature vector was made up of the average rate of CHD death, stroke death, obesity, active smoking status, and African-American race. The investigators found the cosine similarity of each county vector to the SPRINT control group average. The same procedure was applied for the ACCORD trial control group. The cosine similarity measures were normalized and scaled to fall between 0-100, where larger values represented greater similarities.

Geographic Information System (GIS) Methods

CHSI data are keyed by state and county Federal Information Processing Standard (FIPS) codes. For example, Davidson County (county FIPS 037) in Tennessee (state FIPS 47) can then be mapped to other public data sources such as census data. The FIPS codes can be linked to GeoJSON data from the U.S. Census Bureau. The U.S. Census Bureau mappings allowed the conversion of FIPS codes to county border coordinates on GIS maps of the U.S. The investigators used SciPy and Folium (an open sourced package) for this process. The package allowed the investigators to illustrate the similarity of SPRINT/ACCORD participants to the people with hypertension or type 2 diabetes living in each county.

Statistical Analysis

The investigators sought to quantify if counties near trial sites were in fact more similar than those farther away. The investigators used latitude and longitude distances from the centroid of each county to the nearest trial site. The definition of "near" a trial site was increased one mile at a time starting at a radius of five miles. Figure 1 provides an illustration of the process of defining "near" and "far" sites. At each increment, a Wilcoxon Rank Sum test was used to examine if there was a difference in medians between the "near" sites and the "far" sites. The investigators repeated this procedure until no statistically significant difference, factoring in a Bonferroni-Holm correction, existed between counties "near" a trial site and those "far" away. The thought behind this procedure was that trial sites would mostly recruit patients within their geographic sphere of influence. However, as one expands the definition of "near", differences should wash out as the average of the "near" group of counties becomes more like that of the "far" group of counties.

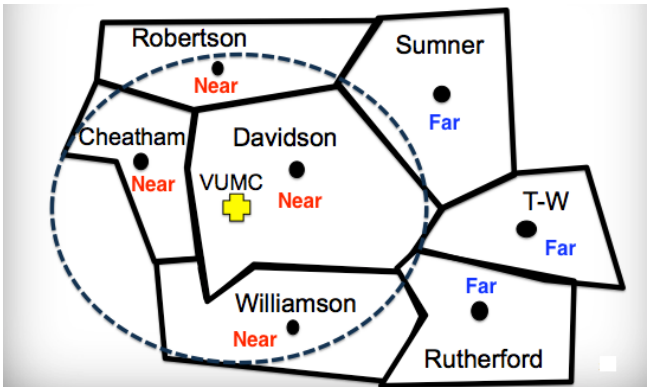

Figure 1. Example of Near/Far County Definition From Vanderbilt University Medical Center (VUMC) Trial Site

The investigators also analyzed if similarity had any correlation with differences in the rate of primary outcomes for the intervention group, essentially whether similarity was clinically meaningful. To begin this analysis, the investigators grouped the SPRINT participants by trial site. Next, the investigators calculated the similarity of the site-specific control group to the entire trial control group average, using the same feature vector of CHD death, stroke death, obesity, active smoking status, and African-American race. The investigators defined clinical significance as the absolute value of the site-specific difference in the primary outcome rate for the intervention group with the trial average primary outcome rate for the intervention group. Basically, does site similarity to the trial average (as measured by the control group) explain divergence in effectiveness of the intervention in the intervention group? Robust regression was applied to the measure of clinical significance with the following explanatory variables: similarity, number of participants at the site, and an interaction term. The investigators used the number of participants and the interaction term to address issues with discreteness. For example, a site with 3 participants can only observe rates of 0%, 33%, 67%, and 100%; such quintiles are far from a continuous measure.

Robust regression makes a Huber-White correction to the model variance, which often widens the confidence intervals for the model coefficients.

Results

The characteristics of the CHSI dataset are summarized in Table 1. All of the variables except for African-American race and Hispanic race appear to be symmetrically distributed. Symmetry is important for the rapid convergence of the sampling distribution (county measures over different years) to the normal distribution. The CHD variable was the only variable with missing values subjected to mean imputation (less than 3% missing). All other variables with missing values went through the multiple imputation pipeline. Table 1 and Table 2 are summarized to the national level for readability, but all the analysis was done at the county level.

Table 1. CHSI Data Characteristics

	Mean	Median	Number Missing
CHSI Hypertension	32436 ± 7104	32200	551/3141 (18%)
CHSI % African-American	9.0%	2.1%	0/3141 (0%)
CHSI % Hispanic	7.0%	2.3%	0/3141 (0%)
CHSI Diabetes Type 2	10094 ± 3224	10000	682/3141 (22%)
CHSI Obese per 100,000	30779 ± 6146	30660	236/3141 (8%)
CHSI Smoker per 100,000	21281 ± 6314	20800	430/3141 (14%)
CHSI Stroke Death per 100,000	54±19	53	474/3141 (15%)
CHSI CHD Death per 100,000	249 ± 80	245	48/3141 (1.9%)

After imputation, the CHSI variables were scaled to probabilities and transformed into conditional probabilities, e.g. the probability of smoking given hypertension (SPRINT) or type 2 diabetes (ACCORD). Table 2 reports the conditioned national averages at the 95% confidence level. Again, the analyses of this study took place at the county level, not the national level, making Table 2 a summarized view of the data.

Table 2. CHSI Conditioned Variable National Averages

	95% Confidence Interval
CHSI % African-Americans with Hypertension	[8.48%, 9.50%]
CHSI % Obese with Hypertension	[34.95%, 36.02%]
CHSI % Smoker with Hypertension	[22.94%, 23.67%]
CHSI % Stroke Death with Hypertension	[0.10%, 0.11%]
CHSI % CHD Death with Hypertension	[0.48%, 0.53%]
CHSI % African-American with Type 2 Diabetes	[8.48%, 9.50%]
CHSI % Hispanic with Type 2 Diabetes	[6.58%, 7.45%]
CHSI % Smoker with Type 2 Diabetes	[23.89%, 25.13%]
CHSI % CHD Death with Type 2 Diabetes	[0.46%, 0.52%]

The investigators compared each county data point to the average mapped characteristics of the trial control groups. The investigators used the trial control groups, because some of the features used were part of the primary outcome measured (CHD death and stroke death). These variables would likely be confounded by the intervention. Table 3 visualizes what the trial populations looked like for both the SPRINT trial and the ACCORD trial.

Table 3. Trial Control Group Data Characteristics

	African American	Current Smoker	Cardiac Event Death	Hispanic	Obesity	Stroke and Death
SPRINT	1493/4673 (31%)	590/4673 (13%)	67/4673 (1.4%)	NA	1952/4673 (42%)	17/4673 (0.3%)
ACCORD	968/5123 (19%)	702/5123 (14%)	124/5213 (2.4%)	379/5123 (7.4%)	NA	NA

Figure 2 visualizes the similarity of all US counties to the SPRINT trial. The pins represent SPRINT Trial sites and were placed by latitude and longitude independent of the shading. The darker shading in the figure corresponds to greater similarity between that county and the SPRINT control group. One can see that the South Eastern United States has the greatest concentration of both trial sites and darkly shaded counties.

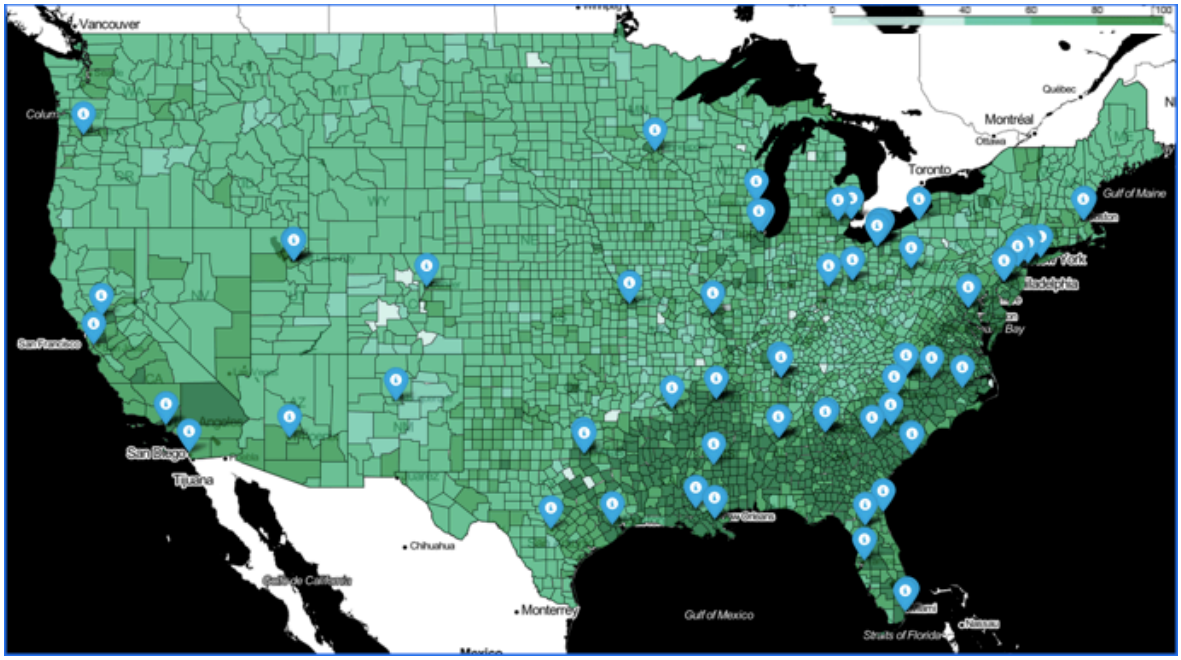

Figure 2. Applicability Map of SPRINT Trial to US Counties

The distribution of similarity scores is not symmetric across the similarity range. Similarity tends to cluster in two places, around a similarity of 60 and around a similarity of 95. The histogram does not describe an obvious distribution. Similarity also appears to have more significance in relative terms than in absolute terms.

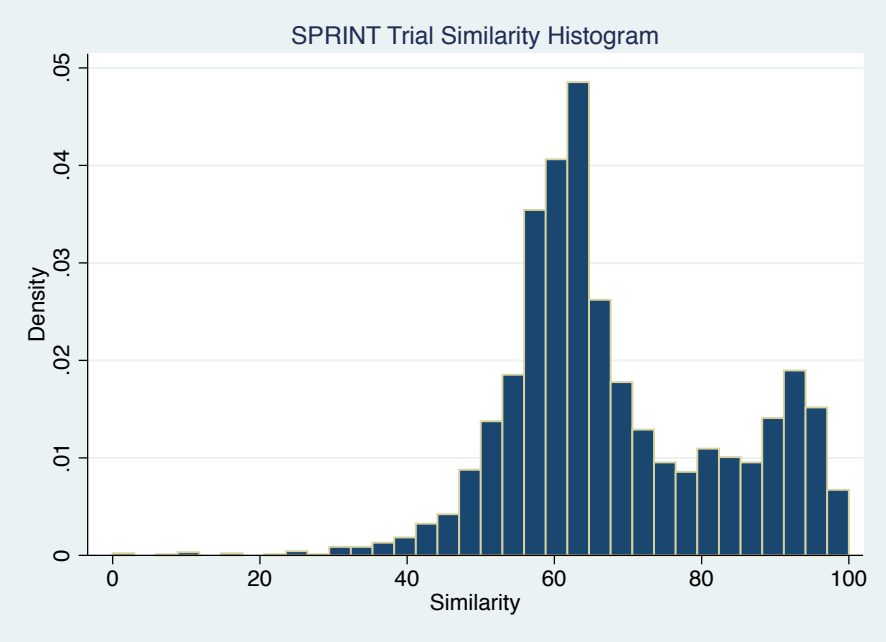

Figure 3. Distribution of SPRINT Similarity Across U.S. Counties

Figure 4 shows the applicability scores of the ACCORD trial to each US county. One can see again that the South East and Eastern seaboard appear fairly well represented by the trial population. The western and central counties appear tend to be less similar.

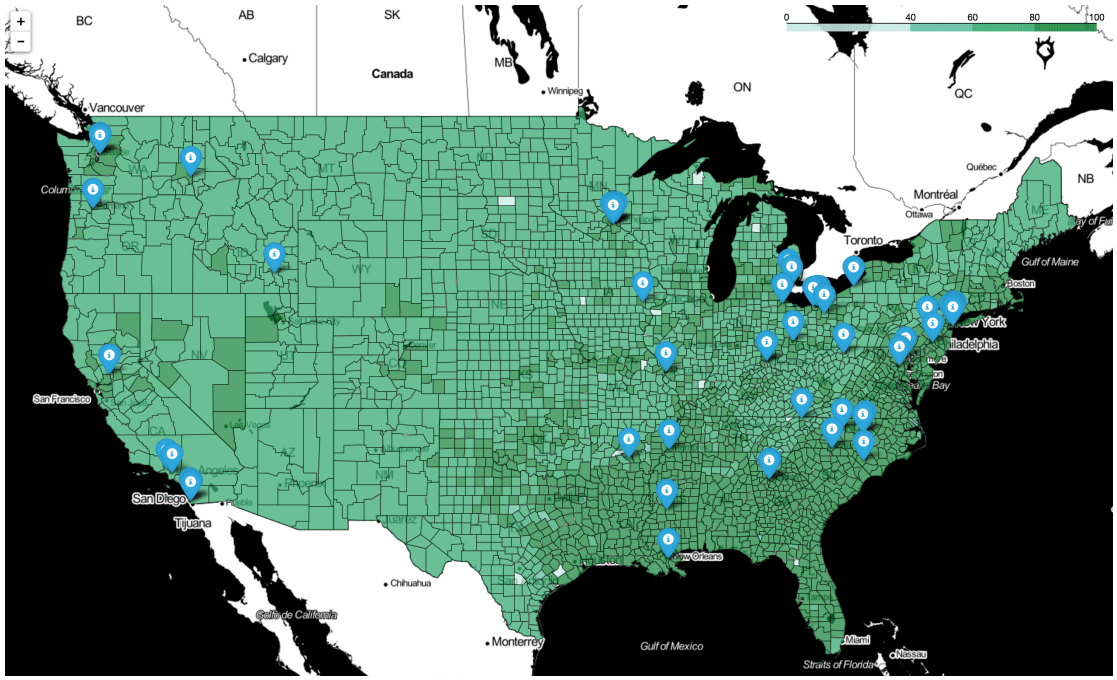

Figure 4. Applicability Map of ACCORD Trial to US Counties

The distribution of similarity for the ACCORD trial, shown in Figure 5, also appears to be bimodal and asymmetric. In this histogram, similarity is concentrated around 75 and around 98. Similarity also appears to be more concentrated and less differentiated than in the SPRINT trial.

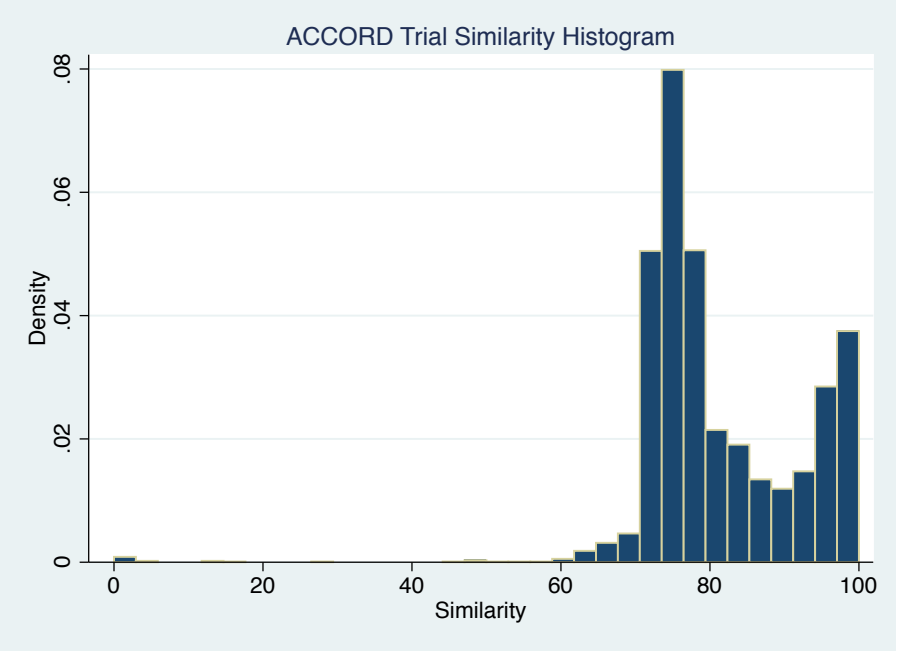

Figure 5. Distribution of ACCORD Similarity Across U.S. Counties

The investigators tested if the similarity of counties near trial sites was greater than counties farther away from any one trial site. For the SPRINT trial the investigators observed a statistical difference in similarity between near counties and far counties up until the inflection point of 495 miles from the nearest trial site. Past this distance, similarities no longer differed between the two groups. In the ACCORD data, differences in control group similarity between counties near trial sites and counties far from trial sites were significant up until 712 miles. Note that these distances are measured from a trial site to the geometric center of a county.

When assessing if similarity had any correlation with differences in intervention outcomes the investigators wanted to get a sense about the distribution of differences across sites. Figure 6 presents a box plot of differences in the rate of primary outcomes (in the intervention group) by SPRINT trial site. Note that in the SPRINT trial, the primary outcome was a negative (MI, non-MI acute coronary syndrome, stroke, heart failure, or death attributable to cardiovascular disease). The distribution is centered at zero, but deviances are skewed toward higher rates of the primary outcome than the average.

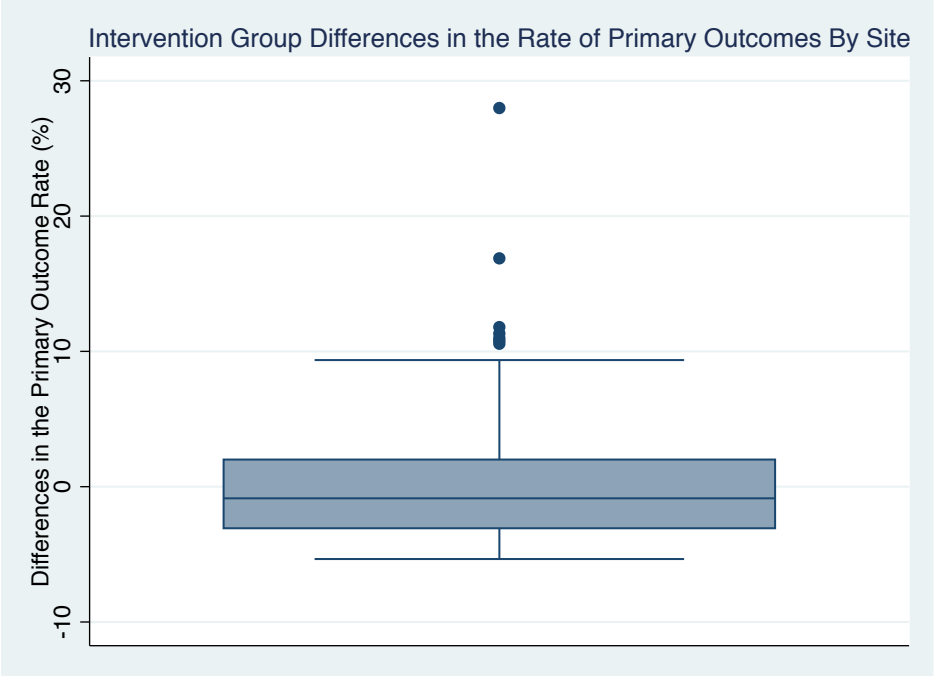

Figure 6. SPRINT Intervention Group Difference in Primary Outcome Rates Grouped by Site

The next part of the analysis took the absolute value of all the site intervention group differences and regressed them on site control group similarity, total number of patients recruited at the site, and an interaction term. Table 4 shows the confidence intervals of all the regressors for this model. Both similarity and the number of site participants were statistically significant predictors. The interaction term was not significant, but was left in the model because it had a significant contribution to the model R^2. This simplistic robust regression was able to explain 36% of the variance of site-specific intervention group differences in the rate of primary outcomes. The similarity coefficient is negative meaning that greater similarity leads to smaller differences in intervention group outcomes. More site participants also corresponded to smaller intervention group differences in outcomes.

Table 4. Robust Regression Coefficients of SPRINT Trial Sites

Regressor	**95% Confidence Interval**
Control Group Similarity	[-0.55, -0.03]*
Number of Participants at Site	[-1.27, -0.01]*
Interaction Term	[0.00, 0.01]
Intercept	[8.00, 58.27]*

Discussion

The primary findings of this work demonstrate that the geographic similarity of trial cohorts to population data can be estimated using publicly available health indicators. Proximity to study sites correlated with increased similarity from trial controls to county data. This result is intuitive and face valid, yet had not been demonstrated using this kind of approach, to the investigators' knowledge. In the SPRINT trial most of the country, geographically speaking, is lightly shaded, meaning the trial cohort does not represent them as well. A possible rationale for this result is that the SPRINT trial population generally has more African Americans, fewer smokers, and more people with high Body Mass Indices (BMI) than what the authors found the national hypertensive population to be. Both trials performed various subgroup analyses on different demographic and clinical criteria. While neither publication mentioned oversampling explicitly, both research groups had predefined subgroups in mind during study design. The predefined subgroup analyses may have influenced the demographics of recruiting with regard to racial features, nevertheless the overall results the researchers reported reflect the mixture of who they recruited.

This study demonstrated weak correlation between cohort similarity and outcome similarity. This result offers further face validity to the method, because one might expect results to hold in a population that looks similar to the population used for testing. The regression results also demonstrate the dangers of over generalizing results; intervention outcomes can differ significantly from the trial reported average. Further study is needed to better establish the link between population similarity metrics and differences in intervention outcomes.

The investigators conjecture that comparing variables that are more contextually similar will increase the power of this method. For example, many studies report the mean BMI and its standard deviation; this cannot be mapped to the CHSI data set easily. Instead, it would be of greater value for this method to know the percentage of trial participants considered obese. The investigators were able to use the prevalence of obesity in the case of SPRINT trial because they had received the data as part of a New England Journal of Medicine (NEJM) Challenge. Generally, the full trial data set is kept privately as the hard-earned spoils of the trial authors' large and taxing effort to design, fund, and execute the trial. The decision to share SPRINT trial data in rapid fashion by the NEJM was not without controversy, but reflects a cultural trend toward open data[29].

The strengths of this work include the application of publicly available data to generate a generalizable method for assessing similarity of trial controls to county population health data. The CHSI data cover many of the most impactful and prevalent diseases and demographic behaviors in the population. The investigators' method is scalable to other trials, because very little trial specific data are needed. The data needed by this method can be aggregated, lessening the risk of health privacy issues. Mapping fields between CHSI and a trial can largely be done with control population summary statistics and tables of adverse outcomes. These two sources of information are often published as part of the trial results. Open source software was used to generate all results to encourage generalizability. While this method does require some clinical expertise for variable mapping and some work is required for each trial, the results can be served across the country, hinting that this approach has economy of scale returns. Another strength of the approach is its use of the normal distribution to condition general health statistics onto sub populations. The properties of the normal distribution allow one to condition on an infinite number of conditions using similar methods.

A major limitation of this approach is the lack of joint probabilities in CHSI data with respect to outcomes such as stroke death and risk factors such as hypertension. Having joint and marginal probabilities would allow for flexible conditioning and would not require the assumptions made here. This approach may benefit from better assumptions on distribution (e.g. Gamma or Poisson)[30]. These bivariate distributions are challenging to find conditional expectations for analytically. Altering the distributional assumptions represents future work. This method may also benefit from using a likelihood-based approach to compare the likelihood function of one population to another. Another assumption was the use of equal weights for all population features. It is likely that some population characteristics affect study applicability more than others. Future work should include replicative efforts with other trial data and assessment of whether the applicability differences actually manifest.

Conclusions

Large public data sets with a broad coverage of demographics, lifestyle variables, and clinical states empower researchers to think at a national level. The rapid assessment of an interactive map comparing trial cohorts to county populations, such as the one presented, may help inform shared decision-making between providers and patients. Specifically, this work would help providers set a prior probability of study applicability to their local population. This may also help providers quickly prioritize which trial results to integrate into their practice. Furthermore, trial administrators could use this method to assess geographic recruitment gaps in rapid fashion.

References

1. Institute of Medicine. Crossing the quality chasm: a new health system for the 21th century. *Inst Med*. 2001:1-8.
2. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ*. 1996;312(7023):71-72.
3. Greenhalgh T, Howick J, Maskrey N. Evidence based medicine: a movement in crisis? *BMJ*. 2014;348:3725.
4. Sackett DL. Evidence-based medicine. *Semin Perinatol*. 1997;21(1):3-5.
5. Haynes B, Haines A. Barriers and bridges to evidence based clinical practice. *BMJ*. 1998;317:273-276.
6. Straus SE, Mcalister FA. Evidence-based medicine: a commentary on common criticisms. *Can Med Assoc J*. 2000;163(7):3-7.
7. Alper BS, Hand JA, Elliott SG, et al. How much effort is needed to keep up with the literature relevant for primary care? *J Med Libr Assoc*. 2004;92(4):429-437.
8. Rothwell PM. External validity of randomised controlled trials : "To whom do the results of this trial apply?" *Lancet*. 2005;365:82-93.
9. Charlson ME, Horwitz RI. Applying results of randomised trials to clinical practice : impact of losses before randomisation. *BMJ*. 1984;289:1281-1284.
10. Bucher HC, Mcalister FA. Users' guides to the medical literature XIX: how to use an article measuring the effect of an intervention on surrogate end points. *JAMA*. 1999;282(8):771-778.
11. Kent DM, Hayward RA. Limitations of applying summary results of clinical tials to individual patients. *JAMA*. 2007;298(10):1209-1211.
12. Stroup DF, Berlin JA, Morton SC. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *JAMA*. 2000;283(15):2008-2012.
13. Feinstein AR. Meta-analysis: statistical alchemy for the 21st century. *J Clin Epidemiol*. 1995;48(1):71-79.
14. Detsky AS, Naylor CD, O'Rourke K, et al. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol*. 1992;45(3):255-265.
15. Natanson C, Esposito CJ, Banks SM. The sirens' songs of confirmatory sepsis trials: selection bias and sampling error. *Crit Care Med*. 1998;26(12).
16. Peppercorn JM, Weeks JC, Cook EF, Joffe S. Comparison of outcomes in cancer patients treated within and outside clinical trials: conceptual framework and structured review. *Lancet*. 2004;363:263-270.
17. Vangeneugden T, Laenen A, Geys H, Renard D, Molenberghs G. Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics*. 2005;61(1):295-304.
18. Elting LS, Cooksley C, Bekele BN, et al. Generalizability of cancer clinical trial results: prognostic differences between participants and nonparticipants. *Cancer*. 2006;106(11):2452-2458.
19. Dans AL, Dans LF, Guyatt GH, Richardson S. Users' guides to the medical literature XIV: how to decide on the applicability of clinical trial results to your patient. *JAMA*. 1998;279(7):545-549.
20. Wang F, Sun J. PSF: a unified patient similarity evaluation framework through metric learning with weak supervision. *IEEE J Biomed Heal Informatics*. 2015;19(3):1053-1060.
21. Sun J, Wang F, Hu J, Edabollahi S. Supervised patient similarity measure of heterogeneous patient records. *ACM SIGKDD Explor Newsl*. 2012;14(1):16-24.
22. Lee J, Maslove DM, Dubin JA. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PLoS One*. 2015;10(5):1-13.
23. Zhang P, Wang F, Hu J, Sorrentino R. Towards personalized medicine: leveraging patient similarity and drug similarity analytics. *AMIA Jt Summits Transl Sci Proc*. 2014:132-136.
24. Panahiazar M, Taslimitehrani V, Pereira NL, Pathak J. Using ehrs for heart failure therapy recommendation using multidimensional patient similarity analytics. *Stud Heal Technol Inf*. 2015;210:369-373.
25. Cahan A, Cimino JJ. Visual assessment of the similarity between a patient and trial population: is this clinical trial applicable to my patient? *Appl Clin Inform*. 2016;7(2):477-488.
26. The SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. *N Engl J Med*. 2015;373(22):2103-2116.
27. Gerstein HC, Miller ME, Byington RP, et al. Effects of intensive glucose lowering in type 2 diabetes. *N Engl J Med*. 2008;358(24):2545-2559.
28. Metzler M. Data sources, definitions, and notes. *Community Health Status Indicators: CDC*; 2009.
29. Ledford H. Open-data contest unearths scientific gems — and controversy. *Nature News*. 9 March, 2017.
30. Anderson RN, Rosenberg HM. Age standardization of death rates: implementation of the year 2000 standard. *Natl Vital Stat Rep*. 1998;47(3):1-16, 20.