Replicate Engineered Virtual Patient Populations as Surrogates for Real Patient-Level Data

Francis J. Alenghat

Section of Cardiology, Department of Medicine, University of Chicago, Chicago, Illinois, USA

Corresponding Author: Francis J. Alenghat, MD, PhD; University of Chicago; 5841 S. Maryland Ave, MC 6080; Chicago, IL 60637; USA; Email: falenghat@bsd.uchicago.edu

Short Title: Replicate Engineered Virtual Patient Populations (RE-ViPPs)

Keywords: virtual patients, simulation, open data, clinical trials

Abstract

Objectives: To demonstrate a new method for generating virtual, individual-level data by testing it on a known clinical trial population.

Design: Virtualization of aggregate data from a clinical trial.

Setting: Virtual

Participants: 936,100 virtual patients

Interventions: None

Main Outcomes Measures: Odds ratios for adverse outcomes in virtual patient populations compared to clinical trial participants.

Methods: The replicate engineered virtual patient populations (RE-ViPPs) method, based on aggregate cross-tabulated categorical population data, does not require access to individual-level data. Using sequential regression combined with randomization, it generates virtual individual patients to comprise populations that, on average, closely resemble the real population in question. The method is validated by applying it to aggregated data from the seminal SPRINT trial, which compared intensive versus standard blood pressure treatment goals on major adverse cardiovascular events.

Results: The method yields virtual populations, each with 9361 patients, faithfully mimicking the real SPRINT participants. Multiple logistic regression on 100 such populations shows that factors with the highest odds ratios for the primary event are, in descending order, past clinical

cardiovascular disease, age ≥ 75, chronic kidney disease, high non-HDL, and smoking history. Intensive blood pressure treatment, the trial's intervention, had an odds ratio of 0.74 [0.63-0.87]. On all these measures, the 100 RE-ViPPs mirrored the real SPRINT participants, including the intensive therapy result (actual SPRINT odds ratio: 0.74 [0.62-0.88]).

Conclusions: Clinical data dissemination has limitations. The most coveted data is descriptive at the individual level but comes with significant cost, effort, and time. There is potential for privacy breaches, and the open-data movement has progressed slowly due to data-ownership concerns. RE-ViPPs closely matched the true SPRINT population. Applied to trials, registries, and databases, RE-ViPPs could reduce open-data burdens by encouraging dissemination of aggregate cross-tabulated real data that allow investigators to generate and measure virtual patients.

Introduction

Patient-level data from clinical trials, registries, and the electronic health record (EHR) is the basis for most clinical research. Obtaining and sharing such data can be challenging, and concerns have been raised about cost, effort, patient privacy, and data ownership [1]. Even journals with strong policies to promote data sharing of published papers have suboptimal adherence by authors [2]. Paying the right amount of attention to who requests individual level data, the purpose of the request, the structure and format of the data, and its de-identification to an extent sufficient to eliminate any chance of deducing identity, can be administratively taxing on the source institution, and often those costs are passed on to the requesting investigator [3-5].

Several entities have taken the initiative to open their data for specific trials and registries, which is a positive step forward. These include a consortium of clinical study sponsors initiating clinicalstudydatarequest.com [6,7], the Yale University Open Data Access (YODA) project [8], Duke's Supporting Open Access Research (SOAR) initiative [9], ImmPort [10], and Project Data Sphere [11]. Some have associated costs in the thousands of dollars whereas others are free. Biolincc is the NHLBI's free source for trial data and has reasonable access [12,13]. Besides clinical trials, sources of data can also include institutional, local, national, and international registries as well as clinical data warehouses driven by EHR data from one or more institutions. Typical time from request to obtaining individual-level de-identified data from one's institutional EHR ranges from weeks to months and costs from hundreds to thousands of dollars [4,5].

However, there are sources of *aggregated* data, based on real clinical populations, which are currently underutilized for their potential. Public systems such as Center for Disease Control's WONDER [14], the Healthcare Cost and Utilization Project (HCUPnet) [15], and the Behavioral Risk Factor Surveillance System Web Enabled Analysis Tool [16] are examples of this sort of aggregate open data for vital statistics, registries, and survey results, respectively. As for EHR data, the Informatics for Integrating Biology & the Bedside platform (i2b2) and similar platforms are deployed by multiple hospital systems to query data in an aggregate fashion at those institutions; they are mostly promoted as tools for cohort identification for planning future requests for individual-level data, but they have potential to be more widely used for research purposes [17-19]. Beyond informing requests for individual level data, these sources of aggregate data, as long as they can be queried in a richly cross-tabulated fashion, could contain enough information to approximate real populations. If this can be demonstrated, it would allow investigators using open data to analyze clinical populations without infringing upon patient privacy or overly taxing current systems in place, and at the same time, allow investigators running trials or maintaining registries to continue to have ownership of their individual-level data.

Here a method is described for creating populations of virtual patients from cross-tabulated aggregate data. It is applied to the SPRINT trial, a study of 9361 patients randomized to intensive versus standard blood pressure control, for which the individual-level data summarized in the original publication [20] was recently 'opened' through NHLBI's Biolincc and is thus known [21]. SPRINT was chosen because it is a landmark trial with a well-defined population, and most importantly, provides verified individual-level data that allows for

87 comparison of the replicate engineered virtual patient populations (RE-ViPPs) generated from

88 aggregate data with the actual SPRINT population.

Methods

Study data

This is a secondary analysis of SPRINT data, which is provided publicly, in de-identified fashion, by the National Heart, Lung, and Blood Institute (NHLBI) via the Biolincc data repository (https://biolincc.nhlbi.nih.gov/home/). The study was deemed exempt by the University of Chicago IRB and the data was obtained after a signed data use agreement.

In SPRINT, participants were > 50 years old with a systolic blood pressure (SBP) of 130 to 180 mm Hg and had an increased risk of cardiovascular events for at least one of the following four reasons: clinical or subclinical cardiovascular disease, chronic kidney disease (CKD), a 10-year risk of 15% or greater on the basis of the Framingham risk score, or ≥ 75 years old. Patients with diabetes or prior stroke were excluded. Participants were randomized to an SBP target of either less than 140 mm Hg (the standard-treatment group) or less than 120 mm Hg (the intensive-treatment group). The primary composite outcome was myocardial infarction, other acute coronary syndromes, stroke, heart failure, or death from cardiovascular causes, measured over an average of 3.3 years [20].

The SPRINT individual-level data were aggregated by first choosing 9 independent categorical factors, all of which are well-established as influencing cardiovascular outcomes. Of these, eight were categorical in the original data [sex, senior (age≥75), black race, current or former smoker (grouped together in the present study), SBP above the highest tertile at the start of the trial, CKD (eGFR < 60 ml/min/1.73m^2), history of clinical cardiovascular disease, and intensive SBP treatment]. The last (high non-HDL) was determined by calculating the non-HDL

110 for each patient and identifying those with non-HDL > 160 mg/dL. The dependent variable was
111 the primary composite outcome, also categorical.

112

113 **Construction of Replicate Engineered Virtual Patient Population (RE-ViPP)**

114 One hundred RE-ViPPs were created. The starting material for RE-ViPP construction consists of
115 aggregate data cross-tabulated for every pair of factors to be analyzed. The cross-tabulation is
116 created by partitioning the entire population into subgroups by the status of one independent
117 factor and recording, for each subgroup, the prevalence of each of the remaining factors, along
118 with the rate of the dependent variable (**Supplementary Figure S1**, Steps 1-2). For SPRINT, this
119 is straightforward, since the individual-level data is known. For EHR, registry, and large trial
120 databases, this can be done, without giving investigators access to the individual level data, as
121 long as the database can be queried for two factors at a time. The results of all this partitioning
122 will yield a cross-tabulation of factor prevalences and outcome rates for each subgroup (as in
123 **Figure 1** for SPRINT).

124 With this cross-tabulation in hand, to start a RE-ViPP, the exact number of patients with
125 and without the first independent variable is assigned to match the real population
126 (**Supplementary Figure S1**, Step 3). In the case of the SPRINT analysis, this was, arbitrarily,
127 female sex. Then, in those virtual patients with the first factor, the second factor (in SPRINT,
128 this was senior status) was assigned to the exact percentage measured in the real population
129 and the same was done for those virtual patients without the first factor. Thus the status of the
130 first and second chosen factors will match the real population exactly.

Starting with the third factor (**Supplementary Figure S1**, Step 4), linear regression was performed on the aggregate data using all the previously assigned factors to estimate the chances of a patient having the new factor as a function of the previously assigned factors. The β coefficients and y-intercept from this linear regression comprise a formula to calculate a probability of having the new factor for each virtual patient (Supplementary Materials, Appendix 1). Significant interactions between pairs of factors are checked in the linear regression and included if the interaction β is within a factor of 1000 from the other β coefficients (in SPRINT, no such interactions were found between any pair of the 9 factors).

Once the probability (between 0 and 1) of having the new factor is calculated for each virtual patient, a random number (between 0 and 1) is generated for each patient (**Supplementary Figure S1**, Step 5). If the random number is less than the patient's probability of having the new factor, the patient is assigned as having the new factor. Otherwise, the virtual patient does not. This repeated for each patient. Once all patients are assigned, the process repeats for the next factor to be assigned. After sequentially assigning the status for all the factors for all the virtual patients in the above manner, the outcomes are assigned in the same manner.

Once all factors and outcomes have been assigned to all virtual patients, the overall rate of each factor and outcome is measured against the actual population, and the virtual population is only accepted if within a specified tolerance, for all factors, of the actual population. If the rate of any factor is outside this tolerance, the random number generation is repeated for all patients and factors in order to create a new virtual population until a

population is found that is within tolerance for all factors (**Supplementary Figure S1**, Step 6). For this study of SPRINT, the tolerance was set at ±0.25%. One hundred such acceptable RE-ViPPs are generated in the above manner.

Analysis of RE-ViPPs

Each of the 100 RE-ViPPs is checked for correlation between the independent variables. It is then subjected to multiple logistic regression using the nine categorical independent variables and the primary composite outcome as the dependent variable. Random number generation, assignment of virtual patient factors, and tolerance testing were conducted in Microsoft Excel with Macros. The linear and logistic regression analyses were all conducted using R. Figures were created with ggplot2 in R.

Patient and public involvement

Patients were integral to the SPRINT trial, and patient representatives were involved in the decision to make the trial data open through the National Heart, Lung, and Blood Institute. Patients were not involved in the design or implementation of the RE-ViPPs method or in the interpretation of the results, and results of this study cannot be disseminated directly to study participants due to the virtual nature of the data. The findings will be disseminated to the general population through national and international press, with the hope that it will encourage aggregate open data sharing.

Results

The process for creating RE-ViPPs is presented in the methods section and summarized in **Supplementary Figure S1**. The process starts with curating cross-tabulated aggregate data. For SPRINT, this was obtained by analyzing the individual-level data. However, in its application for populations where individual-level data is not desired or easily available, the creation of the cross-tabulated data would come from querying an EHR database (such as via i2b2) or open aggregate/summary data from registries or clinical trials. **Figure 1** depicts the cross-tabulated aggregate data from SPRINT.

In the case of RE-ViPPs based on SPRINT, the populations were created based on 9 independent factors and the combined primary outcome event. Each RE-ViPP is composed of 9361 patients, utilizes eight equations (Supplementary Materials, Appendix 1), derived from linear regression analysis, for estimating the chance of a patient having a factor or outcome, and requires a total of 74,888 random numbers (9361 virtual patients x 8 random numbers per virtual patient). The tolerance was set at 0.25%, meaning that the prevalence of each factor and the primary event rate all had to be within this absolute percentage from the true rate in SPRINT. If this was not true for a given virtual population, all 9361 virtual patients were regenerated until the criterion was met. The "SPRINT RE-ViPP Generator" is available for download (Supplementary Materials).

One hundred RE-ViPPs were created in this randomized fashion, constrained by the chosen tolerance (each of these RE-ViPPs is available for download in Supplementary Materials). The rate of each included demographic factor (sex, age≥75, black race), clinical

193 factor (prior/current smoking, SBP above the highest tertile, non-HDL>160 mg/dL, CKD defined
194 by eGFR>60, past history of clinical CVD), trial factor (intensive vs standard BP therapy), and the
195 primary outcome (composite myocardial infarction, other acute coronary syndromes, stroke,
196 heart failure, or cardiovascular death), as required by the design, all adhere very closely to the
197 SPRINT rates (**Figure 2**).

198 Also across all subgroups, the prevalence of each demographic and clinical factor in each
199 of the 100 virtual patient populations clusters around the actual rates in SPRINT (**Figure 3**). The
200 virtual patient population is randomized well to intensive and standard BP therapy, just as the
201 real SPRINT patients, and primary event rates also match across all subgroups (**Figure 3**). It is
202 worth noting that the SPRINT population has higher female and lower senior prevalence
203 amongst black participants, lower rates of smoking and higher non-HDL amongst women, more
204 frequent CKD amongst seniors, and higher event rates in seniors, those with CKD, those with
205 past clinical CVD, and those randomized to non-intensive treatment. All these features are
206 preserved in all of the virtual patient populations.

207 Beyond subgroup characteristics, the RE-ViPPs can be analyzed on the individual level
208 similar to a multivariate analysis performed on real populations. Correlation matrices show no
209 strong correlations between the analyzed factors in the RE-ViPPs, just as in the real SPRINT
210 population (**Supplementary Figure S2**).

211 Multiple logistic regression performed on each of the 100 virtual populations show that
212 the odds ratios (ORs) for each factor closely approximates the ORs in SPRINT (**Figure 4**). The
213 confidence intervals of the RE-ViPP ORs are also similar to SPRINT. A volcano plot shows the OR

214 plotted against p values for each factor in each RE-ViPP, again demonstrating close adherence
215 to SPRINT in terms of both magnitude and significance of the findings (**Figure 5**). The risk factors
216 with the highest odds ratio for the primary event are, in descending order, past clinical
217 cardiovascular disease, age ≥ 75, chronic kidney disease, high non-HDL, and smoking history.
218 Factors associated with fewer events are female sex and intensive blood pressure treatment.
219 Just as with the true SPRINT data, neither black race nor high initial SBP had confidence
220 intervals that clear the line of neutrality.

Discussion

The approach described here was chosen to demonstrate that meaningful information can be gleaned from certain aggregated data to simulate individual patients. The method extends beyond what is traditionally done with summary information and it can approximate the results from the true individual-level data. It is also distinct from prior methods that rely on individual-level data to create virtual patients for the purposes of physiological experiment simulation or clinical trial simulation [22,23]. It is very important to emphasize that the method cannot be done with any form of aggregate data—the source data must be cross-tabulated by the studied factors and outcomes, and, as such, all the factors should be categorical, either nominal or ordinal. In the current example with SPRINT, the variables were binary, but this process could be extended to incorporate a limited number of non-binary categorical variables, in which case the cross-tabulation should account for all values of those variables. Another important consideration is the quality of the source data. In a clinical trial such as SPRINT, it is likely very reliable, but in EHR databases, variables such as race and smoking status can be variably and inconsistently reported, and diagnoses and outcomes may rely on ICD codes which primarily serve a billing/documentation function and may not be subject to rigorous independent verification [24,25]. Another limitation is that although the method checks and includes any pairwise interaction between variables that is large and significant, it does not address 3-way or more complex interactions. There were no important interactions between the variables chosen for this SPRINT study, and for other studies this limitation could be addressed by checking for such complex interactions during the linear regression analyses and/or by choosing, when possible, sufficiently distinct variables to analyze.

The method is reminiscent of imputation, the process by which missing data for a variable is replaced by a reasonable prediction, often using regression models, based on the extant values for that variable as a function of the other variables in the dataset [26,27]. This has perhaps taken its most robust form in multiple imputation of chained equations [28,29]. The method here is distinct from multiple imputation in several ways, but an analogy would be that every patient's status for all the variables beyond the first two are treated as missing, and they are imputed sequentially based on logistic regression analysis of the percentage distributions in the aggregate data.

The current study is meant to serve as a practical proof-of-principle example. Data known at the individual level was used to highlight the method and show that the RE-ViPPs result matches that of the real population. In practice, this method would be most useful in situations when the individual-level data is either impossible or difficult to obtain. These are common situations. In many cases, there are concerns about data ownership or patient privacy [30]. In others, it can also involve excess time, effort, and cost [3,31]. Still in others, the issue may be one of data standardization, where a study would ideally be done across datasets or platforms, but the data exist in incongruent formats in each platform [32]. Also, the sheer bulk of individual-level data may be prohibitive, particularly when there is a desire to combine data from multiple EHRs across the world. With these applications in mind, the "SPRINT RE-ViPP Generator" is provided in the Supplementary Materials and can be customized to other sets of aggregate data and variables.

263 There are many recent high profile examples, not only in healthcare, of bulk individual-
264 level data being hacked or leaked with major fallout [33,34]. In all these situations, keeping the
265 individual level data undisclosed while providing the requisite summary statistics could allow
266 other investigators to study the data while abrogating these concerns. One way this could be
267 done is to keep the individual level data behind a firewall but allow conditional queries to
268 extract summary data – this is what the i2b2 platform and others do for EHR data warehouses,
269 what Center for Disease Control's WONDER provides for vital statistics and other limited clinical
270 information, and what the Behavioral Risk Factor Surveillance System Web Enabled Analysis
271 Tool does for its large national health survey results [14,16,17,19]. Another approach would be
272 to have the "data owners" provide the richly cross-tabulated results across as many categorical
273 variables as possible as a small static file (like **Figure 1**) and make it open to interested
274 investigators; this approach would probably be more feasible and require less digital
275 infrastructure for individual clinical trial results in which full data releases are not planned. This
276 could also allay concerns about the provenance of the source data in secondary analysis, as the
277 shared information would be more succinct and verifiable. In all the above scenarios, it is worth
278 emphasizing that none of the analyzed individual patients derived from this method are real.
279 Patients in the real population are well represented by the virtual patients, but as digital
280 versions of creation, the virtual patients are nameless composites only [35], with no
281 opportunity to deduce identity.

282 This method should allow us to analyze trial, registry, and routinely collected clinical
283 information over large populations to find potential connections between heterogeneous
284 clinical and demographic factors in a novel way that does not directly access patient-level data.

The method could reduce barriers that currently impede access and use of open data. It would thereby stimulate more meaningful use of extant information, ultimately to generate new hypotheses for further investigation in real individuals and to identify previously unrecognized clinical relationships.

Acknowledgements

Author Contributions: Dr. Alenghat had full access to the aggregate data, preformed the analysis, and takes responsibility for its accuracy

Competing Interests: none

Funding: This work was supported, in part, by the National Heart, Lung, and Blood Institute, grant number HL116600.

Data sharing statement: Access to SPRINT data is available from the National Heart, Lung and Blood Institute via Biolinnc. Access to the 100 RE-ViPPs is provided in the Supplementary Materials.

298 ## References

299 1. Rockhold F, Nisen P, Freeman A. Data Sharing at a Crossroads. N Engl J Med. 2016;375(12):1115-
300 7. PMID: 27653563.
301 2. Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, Ioannidis JPA. Data sharing and
302 reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing
303 policy: survey of studies published in The BMJ and PLOS Medicine. BMJ. 2018;360:k400. PMID:
304 29440066; PMCID: PMC5809812.
305 3. Sebaa A, Chikh F, Nouicer A, Tari A. Medical Big Data Warehouse: Architecture and System
306 Design, a Case Study: Improving Healthcare Resources Distribution. J Med Syst. 2018;42(4):59.
307 PMID: 29460090.
308 4. University of Chicago Clinical Research Data Warehouse (CRDW) [3/28/2018]. Available from:
309 http://cri.uchicago.edu/crdw/.
310 5. UCSF Request Clinical Data for Research [3/28/2018]. Available from:
311 https://data.ucsf.edu/data-assets/clinical-research.
312 6. Vaduganathan M, Nagarur A, Qamar A, Patel RB, Navar AM, Peterson ED, Bhatt DL, Fonarow GC,
313 Yancy CW, Butler J. Availability and Use of Shared Data From Cardiometabolic Clinical Trials.
314 Circulation. 2018;137(9):938-47. PMID: 29133600; PMCID: PMC5828960.
315 7. Nisen P, Rockhold F. Access to patient-level data from GlaxoSmithKline clinical trials. N Engl J
316 Med. 2013;369(5):475-8. PMID: 23902490.
317 8. Krumholz HM, Waldstreicher J. The Yale Open Data Access (YODA) Project--A Mechanism for
318 Data Sharing. N Engl J Med. 2016;375(5):403-5. PMID: 27518657.
319 9. Pencina MJ, Louzao DM, McCourt BJ, Adams MR, Tayyabkhan RH, Ronco P, Peterson ED.
320 Supporting open access to clinical trial data for researchers: The Duke Clinical Research
321 Institute-Bristol-Myers Squibb Supporting Open Access to Researchers Initiative. Am Heart J.
322 2016;172:64-9. PMID: 26856217.
323 10. Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, Berger P, Desborough V,
324 Smith T, Campbell J, Thomson E, Monteiro R, Guimaraes P, Walters B, Wiser J, Butte AJ.
325 ImmPort: disseminating data to the public for the future of immunology. Immunol Res.
326 2014;58(2-3):234-9. PMID: 24791905.

327 11. Green AK, Reeder-Hayes KE, Corty RW, Basch E, Milowsky MI, Dusetzina SB, Bennett AV, Wood
328 WA. The project data sphere initiative: accelerating cancer research by sharing data. Oncologist.
329 2015;20(5):464-e20. PMID: 25876994; PMCID: PMC4425388.
330 12. Ross JS, Ritchie JD, Finn E, Desai NR, Lehman RL, Krumholz HM, Gross CP. Data sharing through
331 an NIH central database repository: a cross-sectional survey of BioLINCC users. BMJ Open.
332 2016;6(9):e012769. PMID: 27670522; PMCID: PMC5051517.
333 13. Coady SA, Wagner E. Sharing individual level data from observational studies and clinical trials: a
334 perspective from NHLBI. Trials. 2013;14:201. PMID: 23837497; PMCID: PMC3750470.
335 14. Friede A, Rosen DH, Reid JA. CDC WONDER: a cooperative processing architecture for public
336 health. J Am Med Inform Assoc. 1994;1(4):303-12. PMID: 7719813; PMCID: PMC116209.
337 15. Weiss AJ, Elixhauser A, Steiner C. Readmissions to U.S. Hospitals by Procedure, 2010: Statistical
338 Brief #154. Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville (MD)2006.
339 16. Fisher MA, Ma ZQ. Multiple chronic conditions: diabetes associated with comorbidity and
340 shared risk factors using CDC WEAT and SAS analytic tools. J Prim Care Community Health.
341 2014;5(2):112-21. PMID: 24327591.
342 17. Klann JG, Buck MD, Brown J, Hadley M, Elmore R, Weber GM, Murphy SN. Query Health:
343 standards-based, cross-platform population health surveillance. J Am Med Inform Assoc.
344 2014;21(4):650-6. PMID: 24699371; PMCID: 4078284.
345 18. Murphy SN, Gainer V, Mendis M, Churchill S, Kohane I. Strategies for maintaining patient privacy
346 in i2b2. J Am Med Inform Assoc. 2011;18 Suppl 1:i103-8. PMID: 21984588; PMCID:
347 PMC3241166.
348 19. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the
349 enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med
350 Inform Assoc. 2010;17(2):124-30. PMID: 20190053; PMCID: 3000779.
351 20. Group SR, Wright JT, Jr., Williamson JD, Whelton PK, Snyder JK, Sink KM, Rocco MV, Reboussin
352 DM, Rahman M, Oparil S, Lewis CE, Kimmel PL, Johnson KC, Goff DC, Jr., Fine LJ, Cutler JA,
353 Cushman WC, Cheung AK, Ambrosius WT. A Randomized Trial of Intensive versus Standard
354 Blood-Pressure Control. N Engl J Med. 2015;373(22):2103-16. PMID: 26551272; PMCID:
355 PMC4689591.
356 21. Burns NS, Miller PW. Learning What We Didn't Know - The SPRINT Data Analysis Challenge. N
357 Engl J Med. 2017;376(23):2205-7. PMID: 28445656.

358 22. Teutonico D, Musuamba F, Maas HJ, Facius A, Yang S, Danhof M, Della Pasqua O. Generating
359 Virtual Patients by Multivariate and Discrete Re-Sampling Techniques. Pharm Res.
360 2015;32(10):3228-37. PMID: 25994981; PMCID: PMC4577546.
361 23. Resalat N, El Youssef J, Tyler N, Castle J, Jacobs PG. A statistical virtual patient population for the
362 glucoregulatory system in type 1 diabetes with integrated exercise model. PLoS One.
363 2019;14(7):e0217301. PMID: 31344037; PMCID: PMC6657828 following competing interests:
364 JRC and PGJ have a financial interest in Pacific Diabetes Technologies Inc., a company that may
365 have a commercial interest in the results of this research and technology.
366 24. Williams DJ, Shah SS, Myers A, Hall M, Auger K, Queen MA, Jerardi KE, McClain L, Wiggleton C,
367 Tieder JS. Identifying pediatric community-acquired pneumonia hospitalizations: Accuracy of
368 administrative billing codes. JAMA Pediatr. 2013;167(9):851-8. PMID: 23896966; PMCID:
369 PMC3907952.
370 25. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD
371 code accuracy. Health Serv Res. 2005;40(5 Pt 2):1620-39. PMID: 16178999; PMCID:
372 PMC1361216.
373 26. Rubin DB. Basic Ideas of Multiple Imputation for Nonresponse. Survey Methodology.
374 1986;12(1):37-47.
375 27. Rubin DB, Schenker N. Multiple Imputation for Interval Estimation from Simple Random Samples
376 with Ignorable Nonresponse. J Am Stat Assoc. 1986;81(394):366-74. PMID:
377 WOS:A1986C648000012.
378 28. Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing
379 predictor values in a risk model with a binary outcome. Stat Methods Med Res. 2007;16(3):277-
380 98. PMID: 17621472.
381 29. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it
382 and how does it work? Int J Methods Psychiatr Res. 2011;20(1):40-9. PMID: 21499542; PMCID:
383 PMC3074241.
384 30. Kostkova P, Brewer H, de Lusignan S, Fottrell E, Goldacre B, Hart G, Koczan P, Knight P, Marsolier
385 C, McKendry RA, Ross E, Sasse A, Sullivan R, Chaytor S, Stevenson O, Velho R, Tooke J. Who
386 Owns the Data? Open Data for Healthcare. Front Public Health. 2016;4:7. PMID: 26925395;
387 PMCID: PMC4756607.
388 31. Ishikawa KB. Medical Big Data for Research Use: Current Status and Related Issues. Japan Med
389 Assoc J. 2016;59(2-3):110-24. PMID: 28299245; PMCID: PMC5333614.

390 32. Hauser RG, Quine DB, Ryder A. LabRS: A Rosetta stone for retrospective standardization of
391 clinical laboratory test results. J Am Med Inform Assoc. 2018;25(2):121-6. PMID: 28505339.
392 33. The Equifax Data Breach: What to Do [3/29/2018]. Available from:
393 https://www.consumer.ftc.gov/blog/2017/09/equifax-data-breach-what-do.
394 34. Patient Data Landed Online After a Series of Missteps 2011 [3/30/2018]. Available from:
395 https://www.nytimes.com/2011/10/06/us/stanford-hospital-patient-data-breach-is-
396 detailed.html.
397 35. Shelley MW. Frankenstein, or, The modern Prometheus : the 1818 text: Oxford ; New York :
398 Oxford University Press, 1998.; 1998.

399

400 ## Figure Legends

401

402 **Figure 1. Cross-tabulated aggregate data from SPRINT.** This represents the results from Steps
403 1-2 outlined in the methods and in Figure S1. The primary event rate and the prevalence of 9
404 independent factors are shown for subgroups defined by presence or absence of those same
405 factors. The verbal equivalent of the top two rows, for instance, is: "Of the 3332 female
406 participants of SPRINT, 100% were female, 30.01% were senior, 40.82% were black, [etc..], and
407 4.98% suffered the primary event. Of the 6029 non-female participants of SPRINT, 0% were
408 female, 27.14% were senior, 26.32% were black, [etc..], and 6.57% suffered the primary event."
409 The data shown here is the only starting material required to generate RE-ViPPs.

410

411 **Figure 2. Patient counts in each RE-ViPP subgroup by demographic factor, clinical factor, trial**
412 **factor, and primary outcome all adhere closely to SPRINT.** One hundred replicate populations,
413 each with 9361 virtual patients, were generated. The rates were, by design, within an absolute
414 0.25% of the actual SPRINT trial for every factor in every RE-ViPP. The full range of RE-ViPP
415 counts are shown, as are p values from one-sample t-tests comparing the 100 RE-ViPPs to the
416 actual SPRINT participants.

417

418 **Figure 3. RE-ViPPs match the real SPRINT population.** In each plot, a dot represents a single
419 virtual patient population subgroup and horizontal bars are the real SPRINT rates. Red and gray
420 respectively depict the subgroups with and without the corresponding x-axis factor. The
421 prevalence of demographic factors [female sex (A), age≥75 (B), and black race (C)] and baseline
422 clinical factors [prior/current smoking (D), highest tertile SBP (E), non-HDL>160 mg/dL (F), CKD
423 defined by eGFR>60 (G), and past history of clinical CVD (H)] in each of the 100 virtual patient
424 populations clusters around the actual rate in SPRINT across all subgroups. The virtual patient
425 population is randomized well to intensive vs. standard BP therapy (I) across subgroups just as
426 the real SPRINT patients, and primary event rates also match across all subgroups (J). Notable

427 features of the SPRINT population include higher female and lower senior rates amongst black
428 participants (A-C), lower rates of smoking and higher non-HDL amongst women (D,F), higher
429 CKD amongst seniors (G), and higher event rates in seniors, those with CKD, those with past
430 clinical CVD, and those randomized to non-intensive treatment (J). All these features are
431 preserved in all of the virtual patient populations.

432

433 **Figure 4. Multiple logistic regression of RE-ViPPs.** Multiple logistic regression performed on
434 the 100 virtual populations (9361 patients each) show that the Odds Ratio (OR) for each factor
435 approximate those of SPRINT. Each colored dot is the calculated OR for a virtual patient
436 population. The OR calculated from the mean beta coefficient for each factor along with the
437 central 95% of these ORs is shown in black. For comparison, the actual SPRINT OR and 95%
438 confidence interval for each factor is shown in red.

439

440 **Figure 5. Multiple logistic regression of RE-ViPPs match SPRINT.** Volcano plot shows the OR
441 plotted against the -log of the p value for each factor (color-coded) in each RE-ViPP and in
442 SPRINT. Lower panel is a magnified area from upper panel. The magnitude and significance of
443 the each factor's impact in SPRINT is recapitulated the by RE-ViPPS, as the SPRINT value falls in
444 the middle of the values for the RE-ViPPs for each respective factor.

Figure 1

		Count (%)	% Female	% Senior	% Black	% Smokers	% Hi-SBP	% Hi-Non-HDL	% CKD	% Clinical CVD	% Intensive	Event Rate (%)
Female	+	3332 (35.6)	100	30.01	40.82	44.87	39.05	31.90	31.75	11.43	50.54	4.98
	-	6029 (64.4)	0	27.14	26.32	61.67	31.28	20.42	26.34	19.59	49.66	6.57
Senior (≥75 years)	+	2636 (28.2)	37.94	100	17.75	52.85	39.23	14.95	44.04	21.85	49.96	9.29
	-	6725 (71.8)	34.68	0	36.86	56.80	32.01	28.25	22.08	14.66	49.98	4.71
Black / African-American	+	2947 (31.5)	46.15	15.88	100	55.55	34.98	27.28	21.85	10.52	49.34	4.99
	-	6414 (68.5)	30.75	33.80	0	55.75	33.61	23.23	31.21	19.52	50.27	6.47
Smoker (Current/Former)	+	5213 (55.7)	28.68	26.72	31.40	100	33.72	23.13	27.58	19.85	50.18	6.85
	-	4148 (44.3)	44.29	29.97	31.58	0	34.45	26.23	29.12	12.70	49.71	4.94
SBP in highest tertile	+	3187 (34.0)	40.82	32.44	32.35	55.16	100	27.11	27.55	15.25	50.39	6.59
	-	6174 (66.0)	32.90	25.95	31.03	55.96	0	23.16	28.64	17.43	49.76	5.70
Non-HDL > 160 mg/dL	+	2294 (24.5)	46.34	17.18	35.05	52.57	37.66	100	23.67	9.85	50.22	6.63
	-	7067 (75.5)	32.11	31.72	30.32	56.70	32.87	0	29.76	18.90	49.89	5.80
CKD (eGFR <60)	+	2646 (28.3)	39.98	43.88	24.34	54.35	33.18	20.52	100	21.01	50.26	8.84
	-	6715 (71.7)	33.86	21.97	34.30	56.22	34.39	26.08	0	14.98	49.86	4.88
Clinical CVD	+	1562 (16.7)	24.39	36.88	19.85	66.26	31.11	14.47	35.60	100	49.87	12.29
	-	7799 (83.3)	37.84	26.41	33.81	53.57	34.63	26.52	26.80	0	49.99	4.74
Intensive Arm	+	4678 (50.0)	36.00	28.15	31.08	55.92	34.33	24.63	28.43	16.65	100	5.19
	-	4683 (50.0)	35.19	28.17	31.88	55.46	33.76	24.39	28.10	16.72	0	6.81

Figure 2

		N: SPRINT	N: 100 RE-ViPPs				p
			Mean	SD	Minimum	Maximum	
	Total	9361	9361	0.0	9361	9361	n/a
	Subgroups						
Demographic Factors	**Female**	3332	3332.0	0.0	3332	3332	n/a
	Senior (≥75 years)	2636	2636.0	0.0	2636	2636	n/a
	Black / African-American	2947	2945.9	13.3	2924	2970	0.39
Clinical Factors	**Smoker (Current/Former)**	5213	5210.9	12.0	5190	5235	0.08
	SBP in highest tertile	3187	3187.7	13.6	3164	3210	0.59
	Non-HDL > 160 mg/dL	2294	2292.0	13.4	2271	2317	0.14
	CKD (eGFR <60)	2646	2646.6	12.5	2624	2669	0.64
	Clinical CVD	1562	1561.9	13.1	1539	1584	0.92
Trial Factor	**Intensive Arm**	4678	4679.4	12.9	4656	4701	0.29
	Primary Outcome	562	560.1	11.8	539	583	0.11

Figure 3

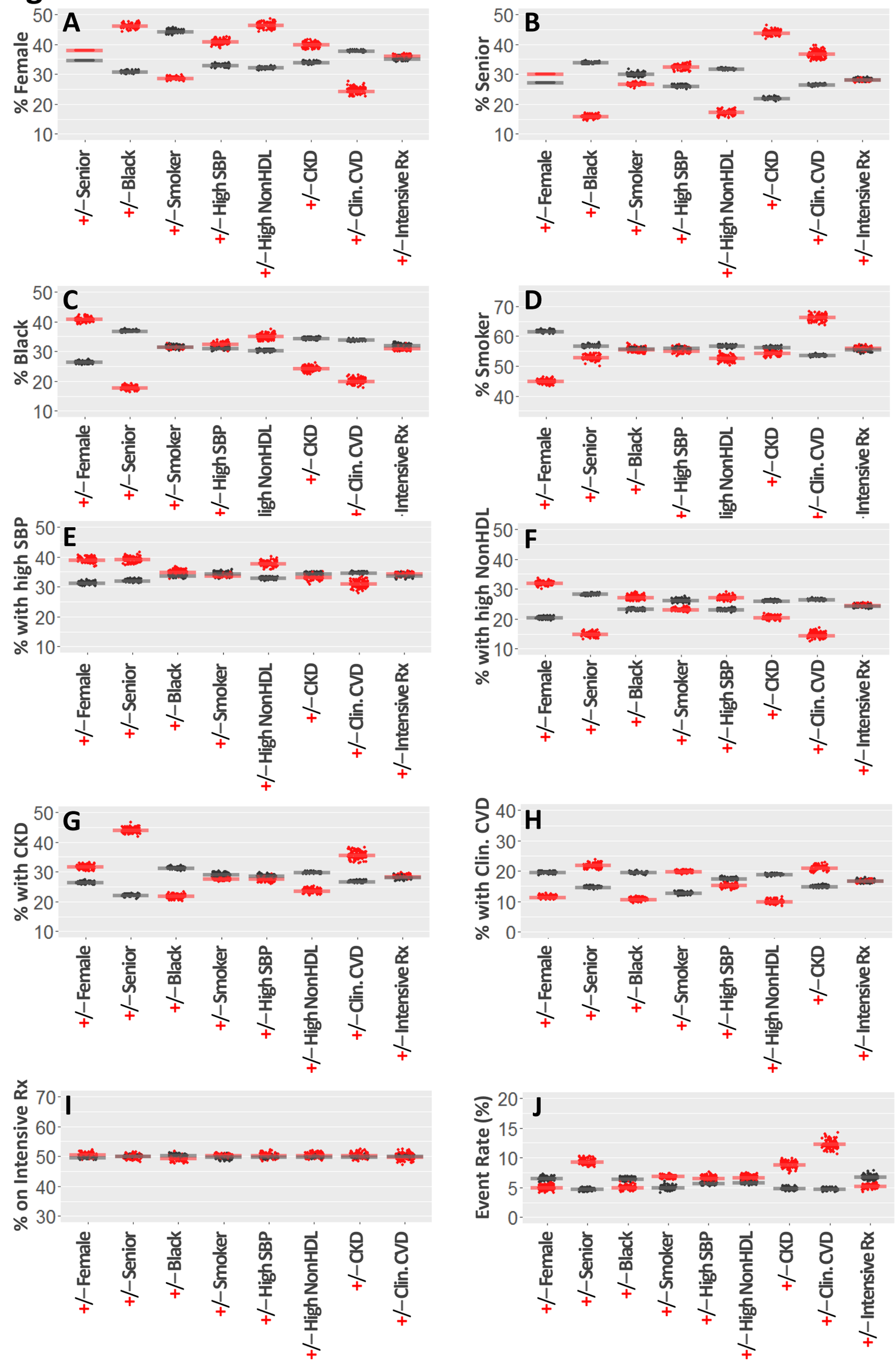

A
% Female
+/−Senior
+/−Black
+/−Smoker
+/−High SBP
+/−High NonHDL
+/−CKD
+/−Clin. CVD
+/−Intensive Rx
B
% Senior
+/−Female
+/−Black
+/−Smoker
+/−High SBP
+/−High NonHDL
+/−CKD
+/−Clin. CVD
+/−Intensive Rx
C
% Black
+/−Female
+/−Senior
+/−Smoker
+/−High SBP
ligh NonHDL
+/−CKD
+/−Clin. CVD
·Intensive Rx
D
% Smoker
+/−Female
+/−Senior
+/−Black
+/−High SBP
ligh NonHDL
+/−CKD
+/−Clin. CVD
·Intensive Rx
E
% with high SBP
+/−Female
+/−Senior
+/−Black
+/−Smoker
+/−High NonHDL
+/−CKD
+/−Clin. CVD
+/−Intensive Rx
F
% with high NonHDL
+/−Female
+/−Senior
+/−Black
+/−Smoker
+/−High SBP
+/−CKD
+/−Clin. CVD
+/−Intensive Rx
G
% with CKD
+/−Female
+/−Senior
+/−Black
+/−Smoker
+/−High SBP
+/−High NonHDL
+/−Clin. CVD
+/−Intensive Rx
H
% with Clin. CVD
+/−Female
+/−Senior
+/−Black
+/−Smoker
+/−High SBP
+/−High NonHDL
+/−CKD
+/−Intensive Rx
I
% on Intensive Rx
+/−Female
+/−Senior
+/−Black
+/−Smoker
+/−High SBP
+/−High NonHDL
+/−CKD
+/−Clin. CVD
J
Event Rate (%)
+/−Female
+/−Senior
+/−Black
+/−Smoker
+/−High SBP
+/−High NonHDL
+/−CKD
+/−Clin. CVD
+/−Intensive Rx

Figure 4

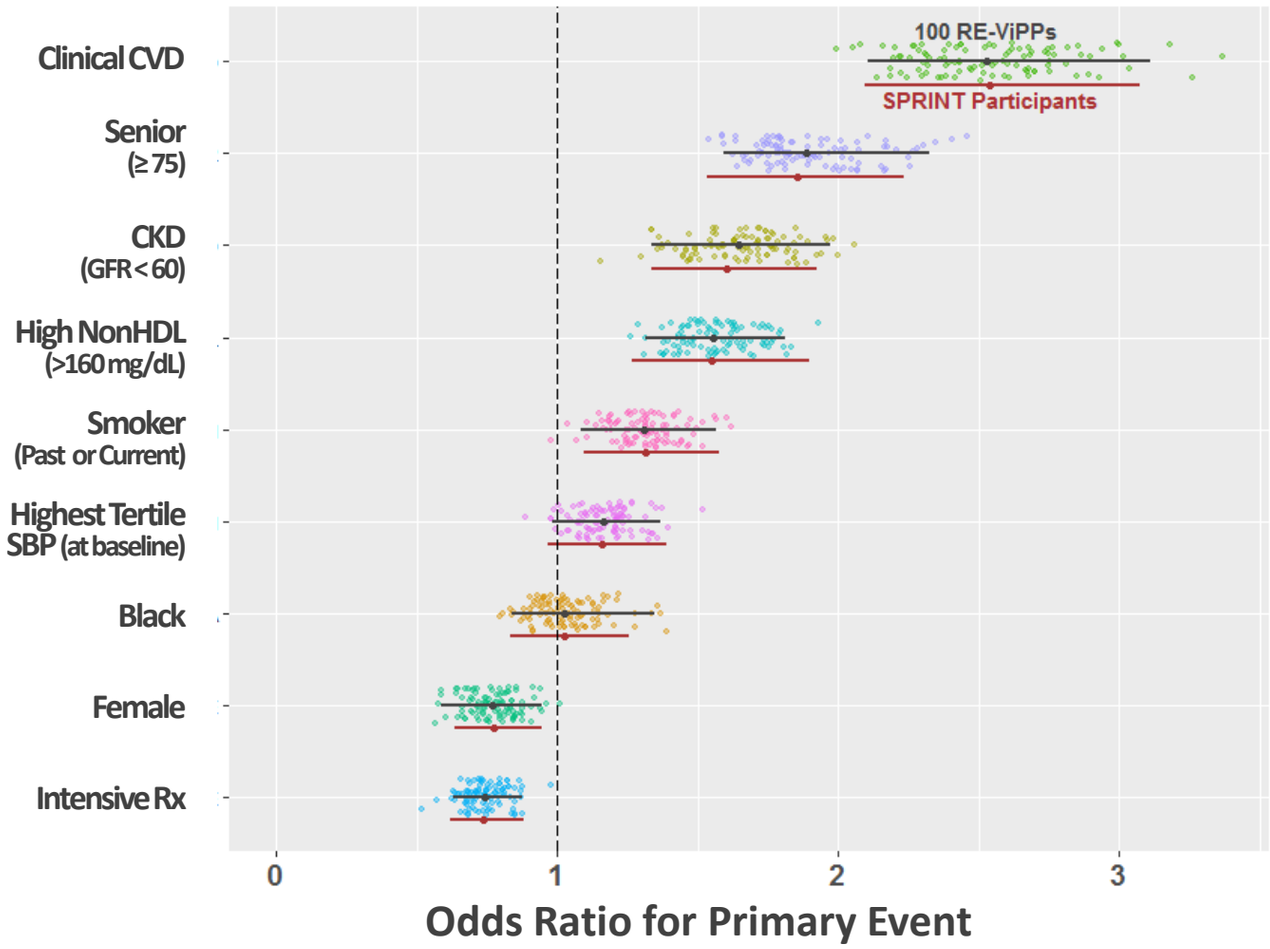
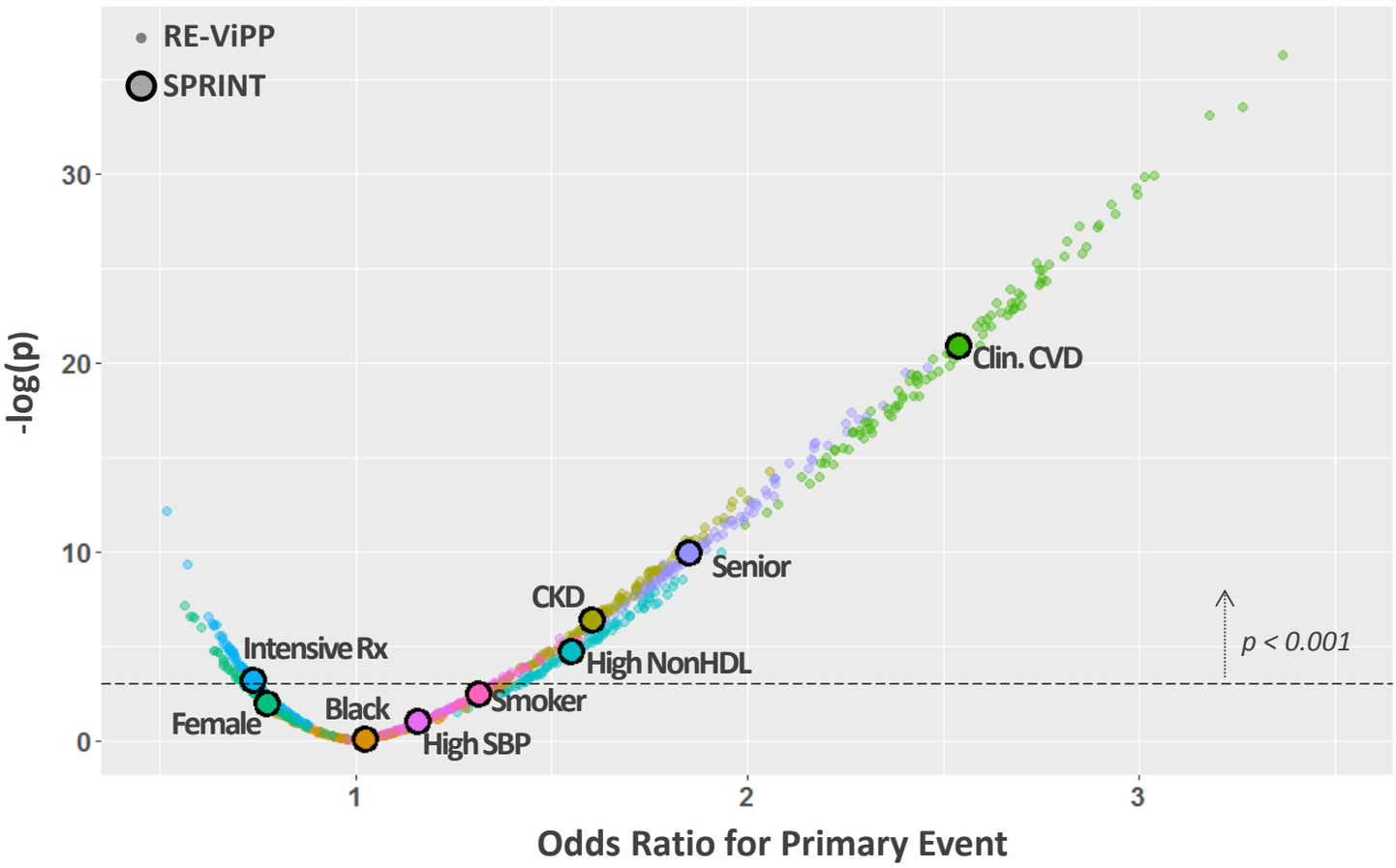
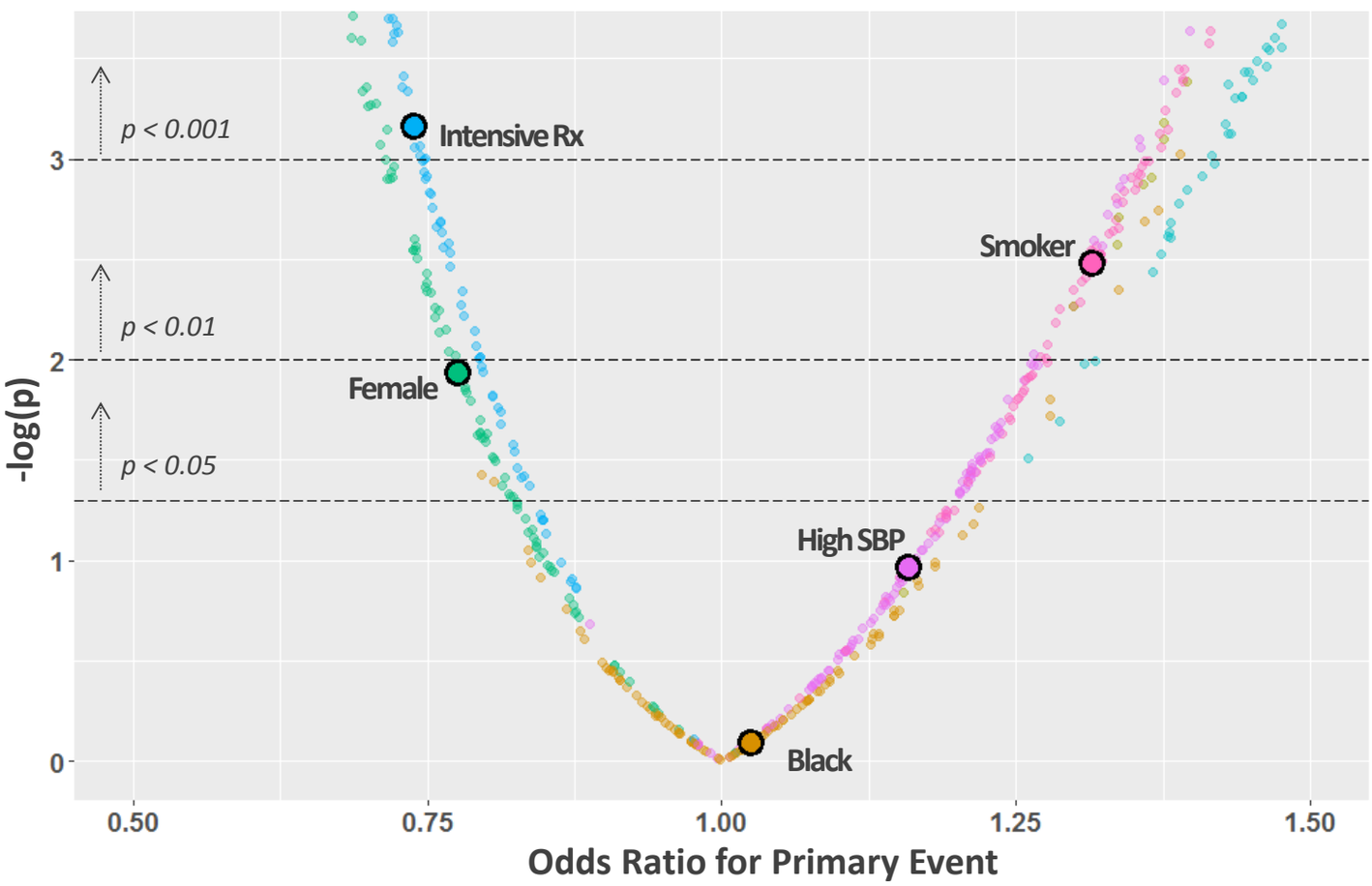

100 RE-ViPPs
SPRINT Participants
Clinical CVD
Senior
(≥75)
CKD
(GFR<60)
High NonHDL
(>160 mg/dL)
Smoker
(Past or Current)
Highest Tertile
SBP (at baseline)
Black
Female
Intensive Rx
0
1
2
3
Odds Ratio for Primary Event

Figure 5


RE-ViPP
SPRINT
-log(p)
30
20
10
0
Clin. CVD
Senior
CKD
High NonHDL
Intensive Rx
Smoker
Black
Female
High SBP
p < 0.001
1
2
3
Odds Ratio for Primary Event


p < 0.001
p < 0.01
p < 0.05
Intensive Rx
Smoker
Female
High SBP
Black
-log(p)
3
2
1
0
0.50
0.75
1.00
1.25
1.50
Odds Ratio for Primary Event