Reverse 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

Email: alenghat@uchicago.edu (FJA)

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

Abstract

Dissemination of clinical data for research has limitations. The most coveted data is richly descriptive at the individual level, but acquiring such granularity comes with significant cost, effort, or time. De-identification of individual records is not foolproof, with potential for privacy breaches, especially for “real-world” data derived from electronic health records. Also, the open data movement has progressed slowly for clinical trials, partly due to concerns about data ownership. Here, reverse engineered virtual patient populations (RE-ViPPs) are described, based on aggregate cross-tabulated categorical data from populations. The method does not require end-user access to individual-level data. Rather, using sequential linear regressions and random number generation, 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 derived from the seminal SPRINT trial, for which the individual-level data is known. The method yields virtual populations, each with 9361 patients, faithfully mimicking the 9361 real SPRINT participants. Multiple logistic regression on 100 such populations shows that, just as in SPRINT, risk factors with the highest odds ratio for the primary event are, in descending order, past clinical cardiovascular disease, age ≥ 75, chronic kidney disease, high non-HDL, and smoking history. Factors associated with fewer events are female sex and intensive blood pressure treatment (the trial’s intervention). Application of RE-ViPPs to trials, registries, and health record databases could reduce the cost, time, ownership, and de-identification burdens hindering open data by encouraging dissemination of aggregate, richly cross-tabulated real data that investigators can use to construct virtual patients and make meaningful conclusions.

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 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, when provided 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 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 reverse engineering 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]. This allows for comparison of the reverse engineered virtual patient populations (RE-ViPPs) generated from aggregate SPRINT data with the actual SPRINT population.

77 # Methods

78 **Study data**

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

83 In SPRINT, participants were > 50 years old with a systolic blood pressure (SBP) of 130 to
84 180 mm Hg and had an increased risk of cardiovascular events for at least one of the following
85 four reasons: clinical or subclinical cardiovascular disease, chronic kidney disease (CKD), a 10-
86 year risk of 15% or greater on the basis of the Framingham risk score, or ≥ 75 years old. Patients
87 with diabetes or prior stroke were excluded. Participants were randomized to an SBP target of
88 either less than 140 mm Hg (the standard-treatment group) or less than 120 mm Hg (the
89 intensive-treatment group). The primary composite outcome was myocardial infarction, other
90 acute coronary syndromes, stroke, heart failure, or death from cardiovascular causes,
91 measured over an average of 3.3 years [20].

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

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

99

100 **Construction of a Reverse Engineered Virtual Patient Population (RE-ViPP)**

101 The starting material for RE-ViPP construction consists of aggregate data cross-tabulated for
102 every pair of factors to be analyzed. The cross-tabulation is created by partitioning the entire
103 population into subgroups by the status of one independent factor and recording, for each
104 subgroup, the prevalence of each of the remaining factors, along with the rate of the
105 dependent variable (**Fig 1**, Steps 1-2). For SPRINT, this is straightforward, since the individual-
106 level data is known. For EHR, registry, and large trial databases, this can be done, without giving
107 investigators access to the individual level data, as long as the database can be queried for
108 multiple factors at a time.

109 To start a RE-ViPP, the exact number of patients with and without the first independent
110 variable is assigned to match the real population (**Fig 1**, Step 3). In the case of the SPRINT
111 analysis, this was female sex. Then, in those virtual patients with the first factor, the second
112 factor (in SPRINT, this was senior status) was assigned to the exact percentage measured in the
113 real population and the same was done for those virtual patients without the first factor. Thus
114 the status of the first and second chosen factors will match the real population exactly.

115 Starting with the third factor (**Fig 1**, Step 4), linear regression was performed on the
116 aggregate data using all the previously assigned factors to estimate the chances of a patient

117 having the new factor as a function of the previously assigned factors. The β coefficients and y-
118 intercept from this linear regression comprise a formula to calculate a probability of having the
119 new factor for each virtual patient (Supplemental Information, Appendix 1). Significant
120 interactions between pairs of factors are checked in the linear regression and included if the
121 interaction β is within a factor of 1000 from the other β coefficients (in SPRINT, no such
122 interactions were found between any pair of the 9 factors).

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

130 Once all factors and outcomes have been assigned to all virtual patients, the overall rate
131 of each factor and outcome is measured against the actual population, and the virtual
132 population is only accepted if within a specified tolerance, for all factors, of the actual
133 population. If the rate of any factor is outside this tolerance, the random number generation is
134 repeated for all patients and factors in order to create a new virtual population until a
135 population is found that is within tolerance for all factors (**Fig 1**, Step 6). For this study of
136 SPRINT, the tolerance was set at ±0.25%. One hundred such acceptable populations are
137 generated in this manner.

138

139 **Analysis of RE-ViPPs**

140 Each RE-ViPP is checked for correlation between the independent variables. It is then subjected
141 to multiple logistic regression using the nine categorical independent variables and the primary
142 composite outcome as the dependent variable.

143 Random number generation, assignment of virtual patient factors, and tolerance testing
144 were conducted in Microsoft Excel with Macros. The linear and logistic regression analyses
145 were all conducted using R. Figures were created with ggplot2 in R.

146 ## Results

147 The process for creating reverse engineered virtual patient populations (RE-ViPPs) is presented
148 in the methods section and summarized in **Fig 1**. The process starts with curating cross-
149 tabulated aggregate data. For SPRINT, this was obtained by analyzing the individual-level data.
150 However, in its application for populations where individual-level data is not desired or easily
151 available, the creation of the cross-tabulated data would come from querying an EHR database
152 (such as via i2b2) or open aggregate/summary data from registries or clinical trials. **Fig 2** depicts
153 the cross-tabulated aggregate data from SPRINT.

154 In the case of RE-ViPPs based on SPRINT, the populations were created based on 9
155 independent factors and the combined primary outcome event. Each RE-ViPP is composed of
156 9361 patients, utilizes eight equations (Supporting Information, Appendix 1), derived from
157 linear regression analysis, for estimating the chance of a patient having a factor or outcome,
158 and requires a total of 74,888 random numbers. The tolerance was set at 0.25%, meaning that
159 the prevalence of each factor and the primary event rate all had to be within this absolute
160 percentage from the true rate in SPRINT. If this was not true for a given virtual population, all
161 9361 patients were regenerated until the criterion was met. The "SPRINT RE-ViPP Generator" is
162 available for download (Supporting Information, upon publication).

163 One hundred RE-ViPPs were created in this randomized fashion, constrained by the
164 chosen tolerance (each of these RE-ViPPs is available for download in Supporting Information,
165 upon publication). The rate of each included demographic factor (sex, age≥75, black race),
166 clinical factor (prior/current smoking, highest tertile SBP, non-HDL>160 mg/dL, CKD defined by

167 eGFR>60, past history of clinical CVD), trial factor (intensive vs standard BP therapy), and the
168 primary outcome (composite myocardial infarction, other acute coronary syndromes, stroke,
169 heart failure, or cardiovascular death), as required by the design, all adhere very closely to the
170 SPRINT rates (**Fig 3**).

171 Also across all subgroups, the prevalence of each demographic and clinical factor in each
172 of the 100 virtual patient populations clusters around the actual rates in SPRINT (**Fig 4**). The
173 virtual patient population is randomized well to intensive and standard BP therapy, just as the
174 real SPRINT patients, and primary event rates also match across all subgroups (**Fig 4**). It is worth
175 noting that the SPRINT population has higher female and lower senior rates amongst black
176 participants, lower rates of smoking and higher non-HDL amongst women, more frequent CKD
177 amongst seniors, and higher event rates in seniors, those with CKD, those with past clinical
178 CVD, and those randomized to non-intensive treatment. All these features are preserved in all
179 of the virtual patient populations.

180 Beyond subgroup characteristics, however, the RE-ViPPs can be analyzed on the
181 individual level similar to a multivariate analysis performed on real populations. Correlation
182 matrices show no strong correlations between the analyzed factors in the RE-ViPPs, just as in
183 the real SPRINT population (**Fig 5**)**.**

184 Multiple logistic regression performed on each of the 100 virtual populations show that
185 the odds ratios (ORs) for each factor approximates the ORs in SPRINT (**Fig 6**). The confidence
186 intervals of the RE-ViPP ORs are also similar to SPRINT. A volcano plot shows the OR plotted
187 against p values for each factor in each RE-ViPP, again showing close adherence to SPRINT (**Fig**

188 **7**). The risk factors with the highest odds ratio for the primary event are, in descending order,
189 past clinical cardiovascular disease, age ≥ 75, chronic kidney disease, high non-HDL, and
190 smoking history. Factors associated with fewer events are female sex and intensive blood
191 pressure treatment. Just as with the true SPRINT data, neither black race nor starting in the
192 highest tertile of SBP had confidence 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. The method cannot be done with any form of aggregate data. The 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. Thus, there are important limitations to this method. Another important limit 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 [22,23]. 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 described here 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 [24,25]. This has perhaps taken its most robust form in multiple imputation of chained equations [26,27]. 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 [28]. In others, it can also involve excess time, effort, and cost [3,29]. 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 [30]. 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 Supporting Information (upon publication) and can be customized to other sets of aggregate data and variables.

233 There are many recent high profile examples, not only in healthcare, of bulk individual-
234 level data being hacked or leaked with major fallout [31,32]. In all these situations, keeping the
235 individual level data undisclosed while providing the requisite summary statistics could allow
236 other investigators to study the data while abrogating all these concerns. One way this could be
237 done is to keep the individual level data behind a firewall but allow conditional queries to
238 extract summary data – this is what the i2b2 platform and others do for EHR data warehouses,
239 what Center for Disease Control's WONDER provides for vital statistics and other limited clinical
240 information, and what the Behavioral Risk Factor Surveillance System Web Enabled Analysis
241 Tool does for its large national health survey results [14,16,17,19]. Another approach would be
242 to have the "data owners" provide the richly cross-tabulated results across as many categorical
243 variables as possible as a small static file (like **Fig 2**) and make it open to interested
244 investigators; this approach would probably be more feasible and require less digital
245 infrastructure for individual clinical trial results in which full data releases are not planned. This
246 could also allay concerns about the provenance of the source data in secondary analysis, as the
247 shared information would be more succinct and verifiable. In all the above scenarios, it is worth
248 emphasizing that none of the analyzed individual patients derived from this method are real.
249 Patients in the real population are well represented by the virtual patients, but as digital
250 versions of creation, the virtual patients are nameless composites only [33], with no
251 opportunity to deduce identity.

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

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

Acknowledgements

Dr. Alenghat had full access to the aggregate data and takes responsibility for the accuracy of the data analysis. Data were provided by the National Heart, Lung, and Blood Institute via Biolincc.

263 ## References

264 1. Rockhold F, Nisen P, Freeman A. Data Sharing at a Crossroads. N Engl J Med. 2016;375(12):1115-
265 7. PMID: 27653563.

266 2. Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, Ioannidis JPA. Data sharing and
267 reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing
268 policy: survey of studies published in The BMJ and PLOS Medicine. BMJ. 2018;360:k400. PMID:
269 29440066; PMCID: PMC5809812.

270 3. Sebaa A, Chikh F, Nouicer A, Tari A. Medical Big Data Warehouse: Architecture and System
271 Design, a Case Study: Improving Healthcare Resources Distribution. J Med Syst. 2018;42(4):59.
272 PMID: 29460090.

273 4. University of Chicago Clinical Research Data Warehouse (CRDW) [accessed 3/28/2018]. Available
274 from: http://cri.uchicago.edu/crdw/.

275 5. UCSF Request Clinical Data for Research [accessed 3/28/2018]. Available from:
276 https://data.ucsf.edu/data-assets/clinical-research.

277 6. Vaduganathan M, Nagarur A, Qamar A, Patel RB, Navar AM, Peterson ED, Bhatt DL, Fonarow GC,
278 Yancy CW, Butler J. Availability and Use of Shared Data From Cardiometabolic Clinical Trials.
279 Circulation. 2018;137(9):938-47. PMID: 29133600; PMCID: PMC5828960.

280 7. Nisen P, Rockhold F. Access to patient-level data from GlaxoSmithKline clinical trials. N Engl J
281 Med. 2013;369(5):475-8. PMID: 23902490.

282 8. Krumholz HM, Waldstreicher J. The Yale Open Data Access (YODA) Project--A Mechanism for
283 Data Sharing. N Engl J Med. 2016;375(5):403-5. PMID: 27518657.

284 9. Pencina MJ, Louzao DM, McCourt BJ, Adams MR, Tayyabkhan RH, Ronco P, Peterson ED.
285 Supporting open access to clinical trial data for researchers: The Duke Clinical Research

286 Institute-Bristol-Myers Squibb Supporting Open Access to Researchers Initiative. Am Heart J.
287 2016;172:64-9. PMID: 26856217.
288 10. Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, Berger P, Desborough V,
289 Smith T, Campbell J, Thomson E, Monteiro R, Guimaraes P, Walters B, Wiser J, Butte AJ.
290 ImmPort: disseminating data to the public for the future of immunology. Immunol Res.
291 2014;58(2-3):234-9. PMID: 24791905.
292 11. Green AK, Reeder-Hayes KE, Corty RW, Basch E, Milowsky MI, Dusetzina SB, Bennett AV, Wood
293 WA. The project data sphere initiative: accelerating cancer research by sharing data. Oncologist.
294 2015;20(5):464-e20. PMID: 25876994; PMCID: PMC4425388.
295 12. Ross JS, Ritchie JD, Finn E, Desai NR, Lehman RL, Krumholz HM, Gross CP. Data sharing through
296 an NIH central database repository: a cross-sectional survey of BioLINCC users. BMJ Open.
297 2016;6(9):e012769. PMID: 27670522; PMCID: PMC5051517.
298 13. Coady SA, Wagner E. Sharing individual level data from observational studies and clinical trials: a
299 perspective from NHLBI. Trials. 2013;14:201. PMID: 23837497; PMCID: PMC3750470.
300 14. Friede A, Rosen DH, Reid JA. CDC WONDER: a cooperative processing architecture for public
301 health. J Am Med Inform Assoc. 1994;1(4):303-12. PMID: 7719813; PMCID: PMC116209.
302 15. Weiss AJ, Elixhauser A, Steiner C. Readmissions to U.S. Hospitals by Procedure, 2010: Statistical
303 Brief #154. Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville (MD)2006.
304 16. Fisher MA, Ma ZQ. Multiple chronic conditions: diabetes associated with comorbidity and
305 shared risk factors using CDC WEAT and SAS analytic tools. J Prim Care Community Health.
306 2014;5(2):112-21. PMID: 24327591.
307 17. Klann JG, Buck MD, Brown J, Hadley M, Elmore R, Weber GM, Murphy SN. Query Health:
308 standards-based, cross-platform population health surveillance. J Am Med Inform Assoc.
309 2014;21(4):650-6. PMID: 24699371; PMCID: 4078284.

310 18. Murphy SN, Gainer V, Mendis M, Churchill S, Kohane I. Strategies for maintaining patient privacy
311 in i2b2. J Am Med Inform Assoc. 2011;18 Suppl 1:i103-8. PMID: 21984588; PMCID:
312 PMC3241166.
313 19. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the
314 enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med
315 Inform Assoc. 2010;17(2):124-30. PMID: 20190053; PMCID: 3000779.
316 20. Group SR, Wright JT, Jr., Williamson JD, Whelton PK, Snyder JK, Sink KM, Rocco MV, Reboussin
317 DM, Rahman M, Oparil S, Lewis CE, Kimmel PL, Johnson KC, Goff DC, Jr., Fine LJ, Cutler JA,
318 Cushman WC, Cheung AK, Ambrosius WT. A Randomized Trial of Intensive versus Standard
319 Blood-Pressure Control. N Engl J Med. 2015;373(22):2103-16. PMID: 26551272; PMCID:
320 PMC4689591.
321 21. Burns NS, Miller PW. Learning What We Didn't Know - The SPRINT Data Analysis Challenge. N
322 Engl J Med. 2017;376(23):2205-7. PMID: 28445656.
323 22. Williams DJ, Shah SS, Myers A, Hall M, Auger K, Queen MA, Jerardi KE, McClain L, Wiggleton C,
324 Tieder JS. Identifying pediatric community-acquired pneumonia hospitalizations: Accuracy of
325 administrative billing codes. JAMA Pediatr. 2013;167(9):851-8. PMID: 23896966; PMCID:
326 PMC3907952.
327 23. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD
328 code accuracy. Health Serv Res. 2005;40(5 Pt 2):1620-39. PMID: 16178999; PMCID:
329 PMC1361216.
330 24. Rubin DB. Basic Ideas of Multiple Imputation for Nonresponse. Survey Methodology.
331 1986;12(1):37-47.

332 25. Rubin DB, Schenker N. Multiple Imputation for Interval Estimation from Simple Random Samples
333 with Ignorable Nonresponse. J Am Stat Assoc. 1986;81(394):366-74. PMID:
334 WOS:A1986C648000012.
335 26. Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing
336 predictor values in a risk model with a binary outcome. Stat Methods Med Res. 2007;16(3):277-
337 98. PMID: 17621472.
338 27. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it
339 and how does it work? Int J Methods Psychiatr Res. 2011;20(1):40-9. PMID: 21499542; PMCID:
340 PMC3074241.
341 28. Kostkova P, Brewer H, de Lusignan S, Fottrell E, Goldacre B, Hart G, Koczan P, Knight P, Marsolier
342 C, McKendry RA, Ross E, Sasse A, Sullivan R, Chaytor S, Stevenson O, Velho R, Tooke J. Who
343 Owns the Data? Open Data for Healthcare. Front Public Health. 2016;4:7. PMID: 26925395;
344 PMCID: PMC4756607.
345 29. Ishikawa KB. Medical Big Data for Research Use: Current Status and Related Issues. Japan Med
346 Assoc J. 2016;59(2-3):110-24. PMID: 28299245; PMCID: PMC5333614.
347 30. Hauser RG, Quine DB, Ryder A. LabRS: A Rosetta stone for retrospective standardization of
348 clinical laboratory test results. J Am Med Inform Assoc. 2018;25(2):121-6. PMID: 28505339.
349 31. The Equifax Data Breach: What to Do [accessed 3/29/2018]. Available from:
350 https://www.consumer.ftc.gov/blog/2017/09/equifax-data-breach-what-do.
351 32. Patient Data Landed Online After a Series of Missteps 2011 [accessed 3/30/2018]. Available
352 from: https://www.nytimes.com/2011/10/06/us/stanford-hospital-patient-data-breach-is-
353 detailed.html.
354 33. Shelley MW. Frankenstein, or, The modern Prometheus : the 1818 text: Oxford ; New York :
355 Oxford University Press, 1998.; 1998.

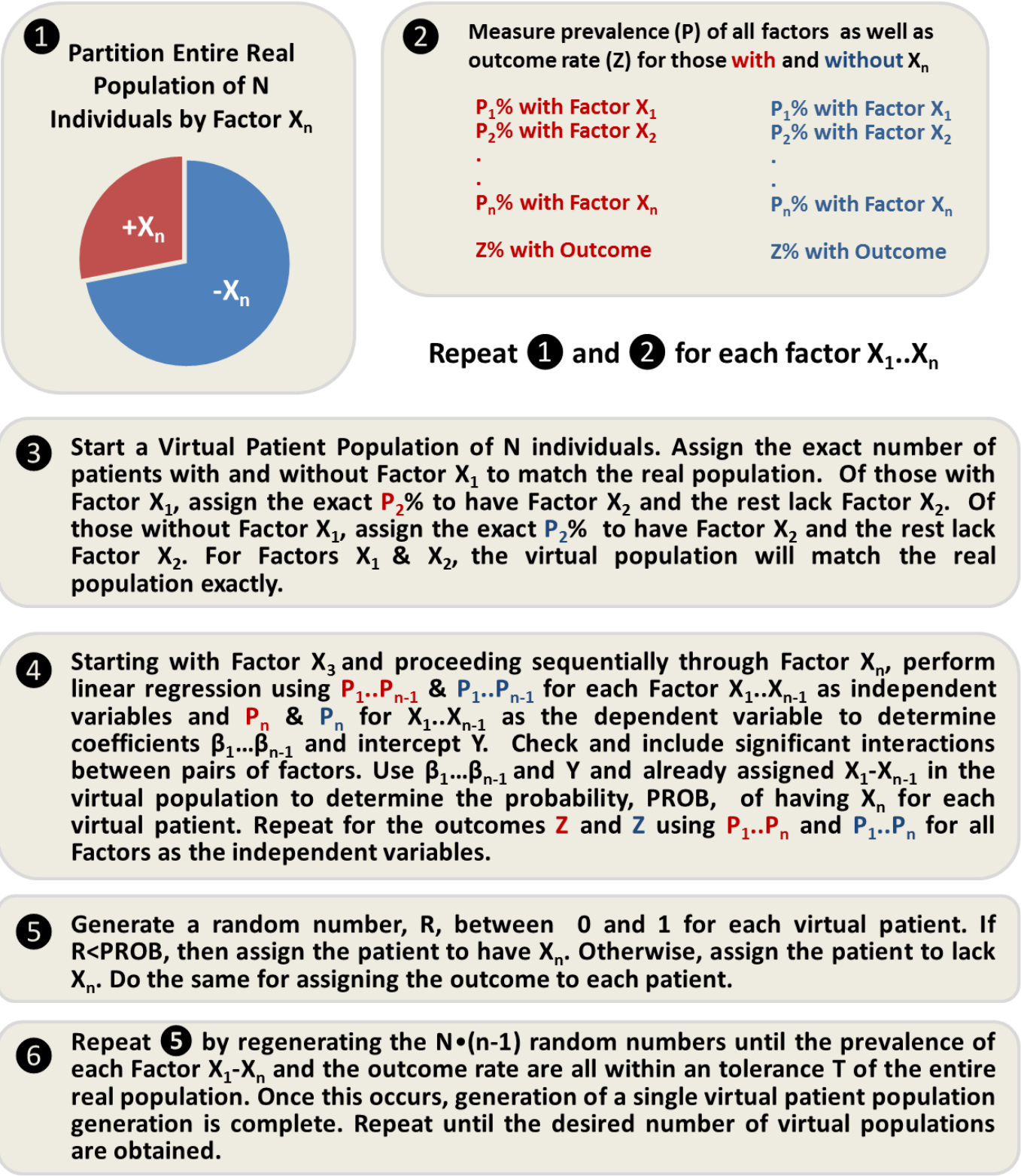


Fig 1. Reverse Engineered Virtual Patient Populations (RE-ViPPs): A step-by-step method. The source material can be created by querying a database of accessible clinical data (Steps 1-2) or be provided by the data owners. Generation of RE-ViPPs (Steps 3-6) relies on assigning variable values to each virtual patient in a sequential manner based on linear regression models, driven by random number generation, and subjected to tolerance limits.

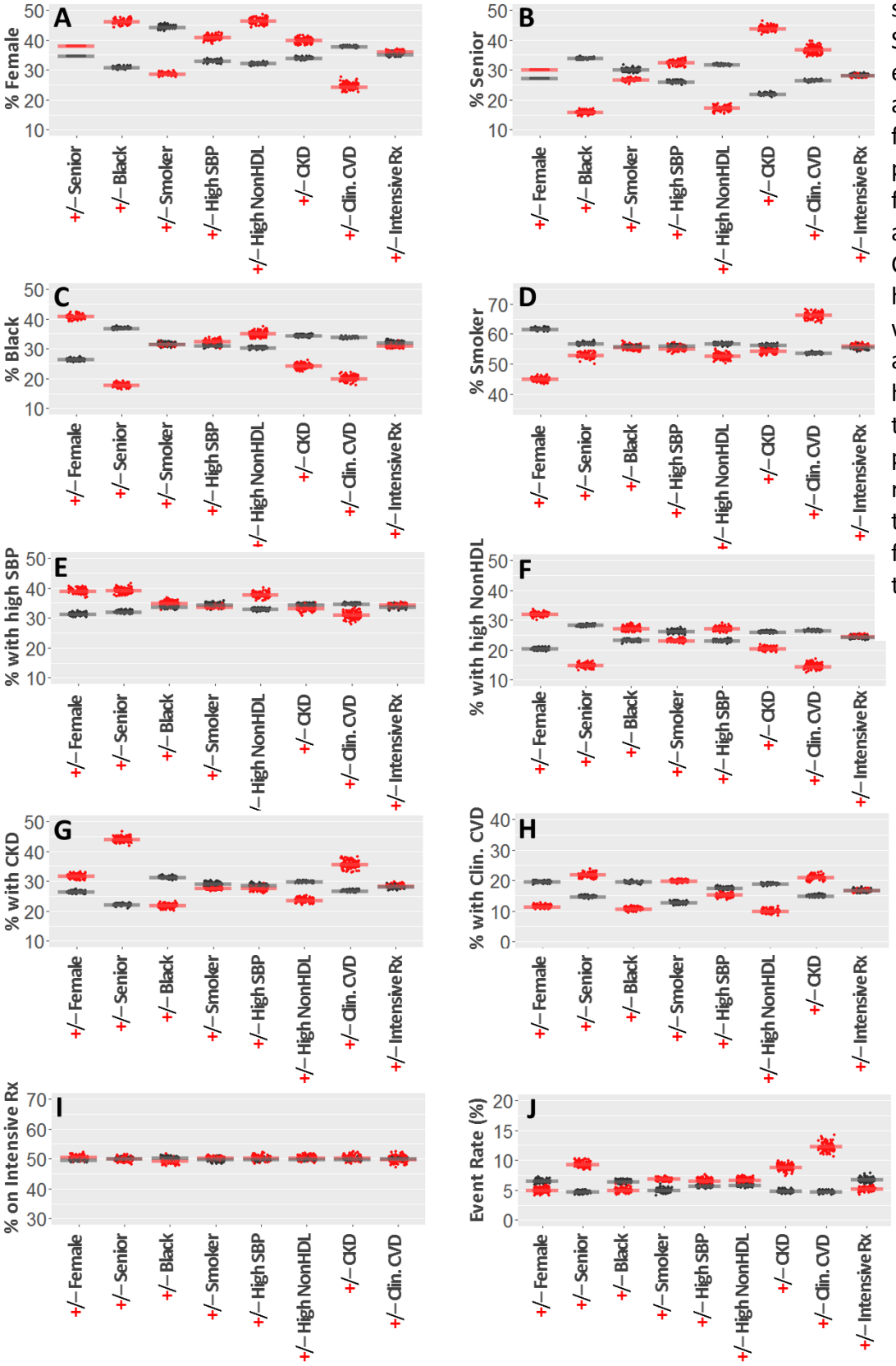
		Count (%)	% Female	% Senior	% Black	% Smokers	% Hi-SBP	% Hi-Non-HDL	% CKD	% Clinical CVD	% Intensive	Event Rate (%)
Female	+	3332 (35.6)	100	30.01	40.82	44.87	39.05	31.90	31.75	11.43	50.54	4.98
	-	6029 (64.4)	0	27.14	26.32	61.67	31.28	20.42	26.34	19.59	49.66	6.57
Senior (≥75 years)	+	2636 (28.2)	37.94	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

Fig 2. Cross-tabulated aggregate data from SPRINT. The data here represent the results from Fig 1, Steps 1-2. The primary event rate and the prevalence of nine independent factors are shown for subgroups defined by presence or absence of those same factors. This is the starting material necessary to generate RE-ViPPs.

		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

Fig 3. Patient counts in each RE-ViPP subgroup by demographic factor, clinical factor, trial factor, and primary outcome all adhere closely to SPRINT. One hundred populations, each with 9361 virtual patients, were generated. The rates were, by design, within an absolute 0.25% of the actual SPRINT trial for every factor in every RE-ViPP. The full range of RE-ViPP counts are shown, as are p values from one-sample t-tests comparing the 100 RE-ViPPs to SPRINT.

Fig 4. Reverse engineered virtual patient populations match the real SPRINT population. In each plot, a dot represents a single virtual patient population subgroup and horizontal bars are the real SPRINT rates. Red and gray respectively depict the subgroups with and without the corresponding x-axis factor. The prevalence of demographic factors [female sex (**A**), age≥75 (**B**), and black race (**C**)] and baseline clinical factors [prior/current smoking (**D**), highest tertile SBP (**E**), non-HDL>160 mg/dL (**F**), CKD defined by eGFR>60 (**G**), and past history of clinical CVD (**H**)] in each of the 100 virtual patient populations clusters around the actual rate in SPRINT across all subgroups. The virtual patient population is randomized well to intensive vs. standard BP therapy (**I**) across subgroups just as the real SPRINT patients, and primary event rates also match across all subgroups (**J**). Notable features of the SPRINT population include higher female and lower senior rates amongst black participants (A-C), lower rates of smoking and higher non-HDL amongst women (D,F), higher CKD amongst seniors (G), and higher event rates in seniors, those with CKD, those with past clinical CVD, and those randomized to non-intensive treatment (J). All these features are preserved in all of the virtual patient populations.



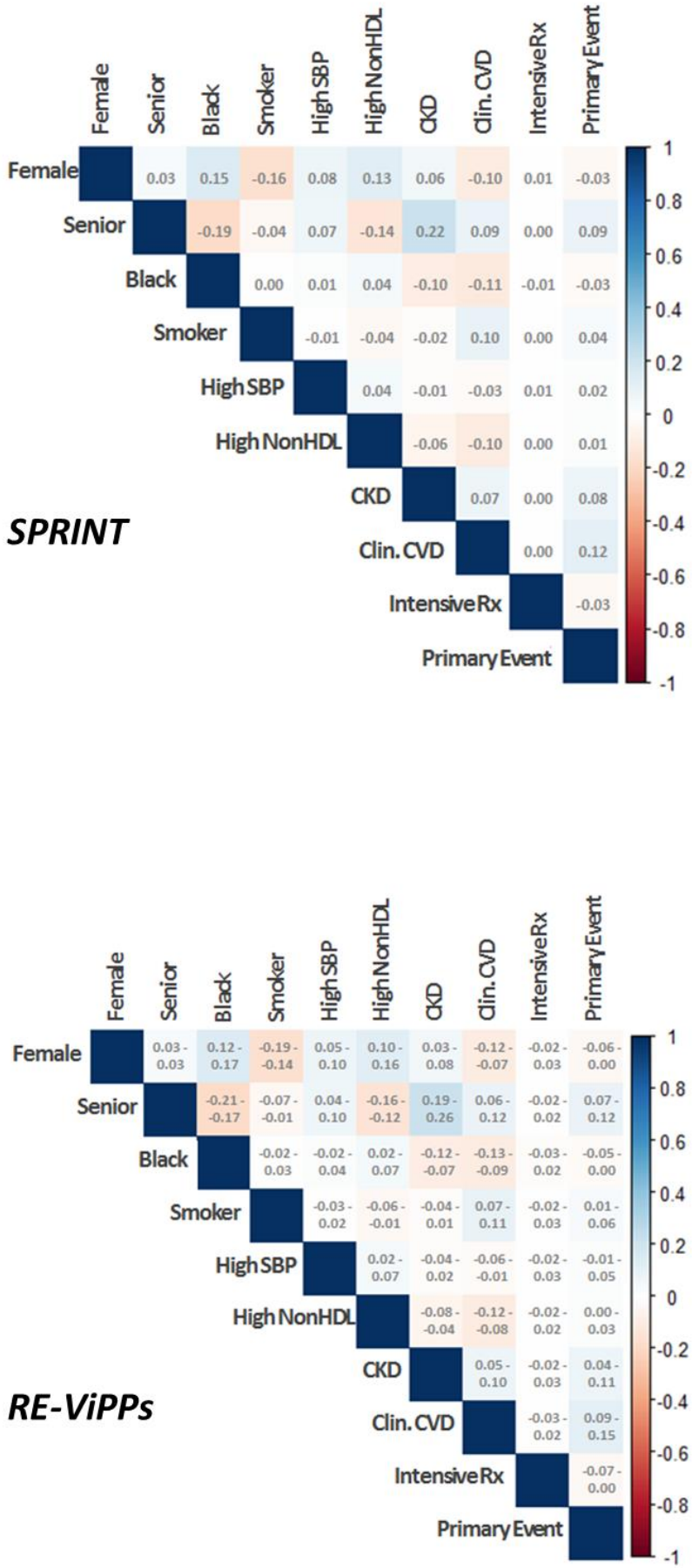


Fig 5. Correlation coefficients of SPRINT and RE-ViPPs. Correlation coefficients are shown for each pair of factors in SPRINT (upper panel). For the RE-ViPPS (lower panel), the average correlations for each pair of factors is reflected by the color scale, and the full range of correlations is provided. The pattern is similar to SPRINT. No coefficient exceeds 0.3 between any pair of factors in any RE-ViPP.

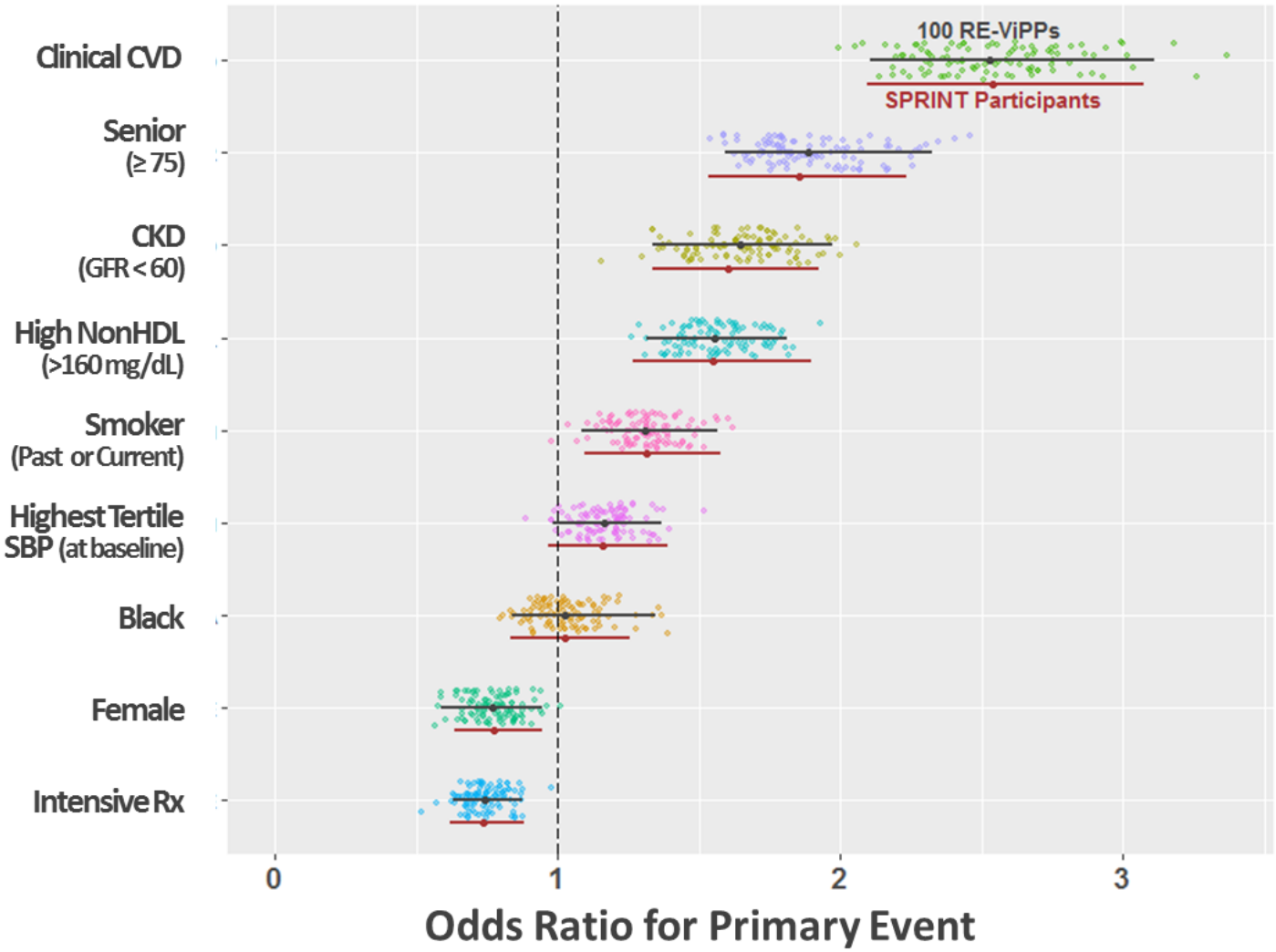


Fig 6. Multiple logistic regression results of RE-ViPPs. Multiple logistic regression performed on the 100 virtual populations (9361 patients each) show that the Odds Ratio (OR) for each factor approximates those of SPRINT. Each colored dot is the calculated OR for a virtual patient population. The OR calculated from the mean β coefficient for each factor along with the central 95% of these ORs is shown in black. For comparison, the actual SPRINT OR and 95% confidence interval for each factor is shown in red.

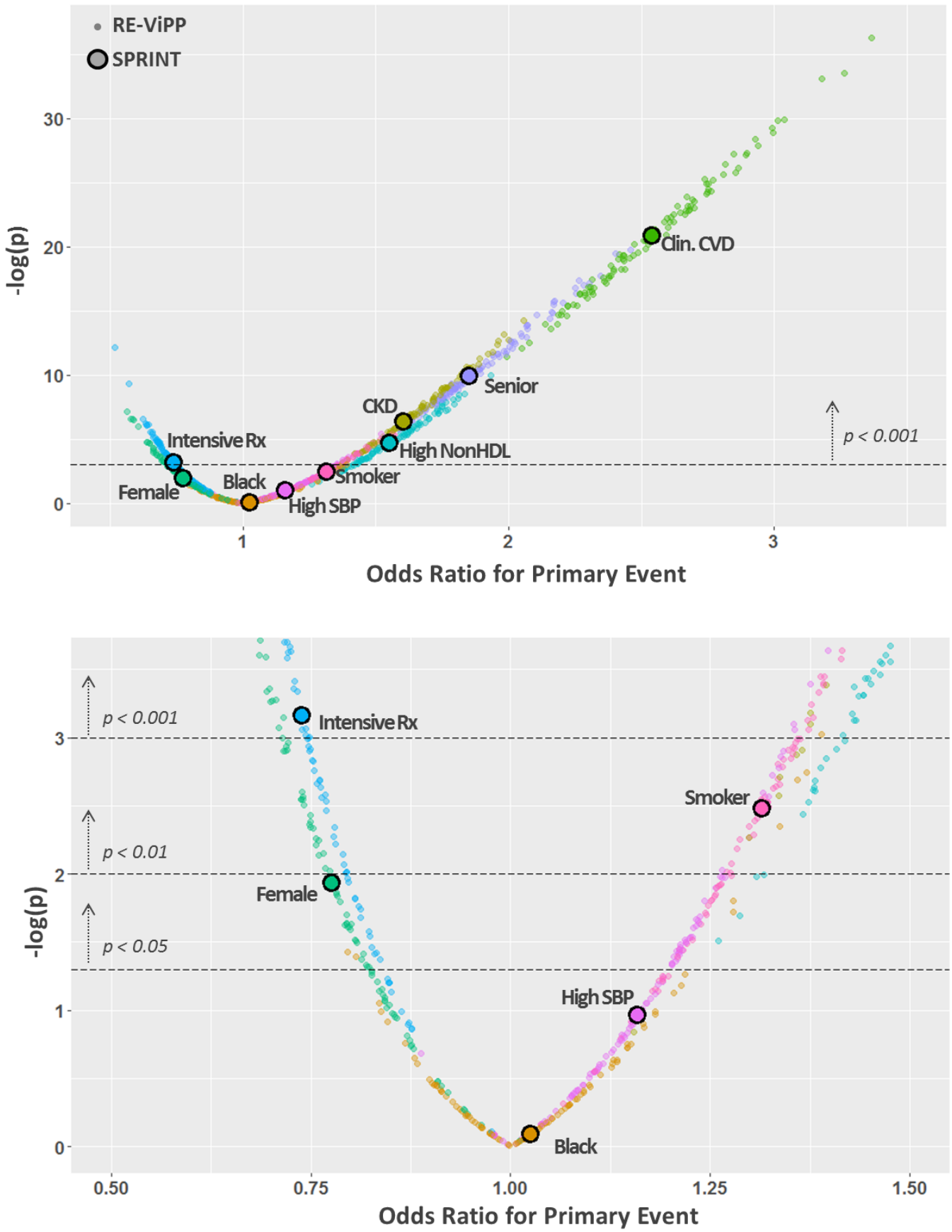


Fig 7. Multiple logistic regression results of RE-ViPPs. Volcano plot shows the OR plotted against the -log of the p value for each factor (color-coded) in each RE-ViPP and in SPRINT. Lower panel is a magnified area from upper panel. For all factors, the SPRINT value falls in the middle of the values for the RE-ViPPs.