Replicability, Reproducibility, and Agent-based Simulation of Interventions

R. Stanley Hum, MD, MA[1], Samantha Kleinberg, PhD[2]
[1] Columbia University, New York, NY; [2] Stevens Institute of Technology, Hoboken, NJ

Abstract

Secondary use of medical data and use of observational data for causal inference has been growing. Yet these data bring many challenges such as confounding due to unobserved variables and variation in medical processes across settings. Further, while methods exist to handle some of these problems, researchers lack ground truth to evaluate these methods. When a finding is not replicated across multiple sites, it is unknown whether this is a failure of an algorithm, a genuine difference between populations, or an artifact of structural differences between the sites. We show how agent-based simulation of medical interventions can be used to explore how bias, error, and variation across settings affect inference. Our approach enables users to model not only interventions and outcomes, but also the complex interaction between patients with different risks of mortality and providers with different observed and latent treatment effects. Ultimately we propose that such simulations can be used to better evaluate the behavior of new methods with known ground truth and better calculate sample size for EHR-based studies.

Introduction

Electronic health records (EHRs) have made it possible to study large populations over longer timescales than has previously been feasible, yet the observational nature of these data lead to a number of challenges, including difficulties in both reproducing studies and understanding when we should expect results to be replicated in a new setting. While confounding due to unobserved variables is a known challenge for causal inference (and methods for inferring latent variables have been developed to address this), the effect of variation in medical processes and unobserved treatments on false positives and negatives is less appreciated and can lead to the same outcomes as selection bias – even without hidden common causes. The problem is further compounded when we repeat studies in multiple settings, whether as part of a multi-site trial or as an attempt at reproducing a particular finding. Standards of care, documentation, and variation in both treatments and patients may differ, and may all contribute to a failure to reproduce a genuine finding. Studies may also fail to show an effect due to being underpowered, and in some cases effective power may be lower than expected due to latent effects that are not accounted for.

For example, a study may aim to test the role of ventilators in treating pneumonia, but ventilators are combined with antibiotics in some cases, and this choice may depend on the population, a clinician's background, and the patient's treatment preferences. While ideally we will have data on all interventions so as to control for such differences, documentation practices and the nature of EHR data mean that such information may be systematically missing. That is, if only text-based data are used, and that is where interventions such as ventilation are documented, while medications are prescribed electronically and stored in a structured database, then work focusing solely on free text will not be able to capture the role of antibiotics.

However, it is difficult to tease apart the effect of each possible source of error and bias from EHR data directly, especially given the lack of ground truth for evaluating results. At the same time, replicability, or the lack thereof, is receiving increasing attention across many fields. We propose that simulations can be used to test algorithms for robustness in the face of these challenges, and to develop solutions to address the unique difficulties of reusing medical data. Here we present the results of an agent-based simulation of medical interventions, showing the individual and compounded impact of hidden interventions and process features. In particular, we examine how combinations of latent and observed interventions may lead to reduced study power, and show how agent-based modeling can help researchers explore these issues.

Background

Reproducibility and causal inference

A key part of the scientific method is that hypotheses should be falsifiable and results should be consistent. A protocol developed by one person should be able to be followed by another and should produce the same results under the same

circumstances. Yet recent articles have been raising concerns that many scientific results cannot be replicated, and that the problem may be getting worse[1]. Attempts by pharmaceutical companies to replicate academic findings led to only 20-25% of findings being repeated[2] and a review of 53 major findings in cancer biology found only 6 were replicable[3]. Recent works such as the Many Labs project[4] have investigated this in social psychology research by attempting large-scale replications. The Open Science Collaboration[5] conducted detailed replications of 100 psychology studies, with only about half being replicable.

While much of the focus has been on the lack of reproducibility in experimental studies, much of the work in biomedical informatics relies on secondary use of observational data. Instead of prospectively collecting data to test a specific hypothesis, this research is primarily data driven, using massive EHR datasets and other sources to generate new hypotheses. Reproducing work is both more important (given the increased possibility for confounding and error in these observational data that were not primarily generated for research purposes) and more difficult in this case. One study looked at 52 claims from observational studies and found none were reproduced in later randomized controlled trials (RCTs)[6]. However, one of the challenges of reproducing claims, particularly in biomedical informatics, is the lack of information about experimental protocols. Ioannidis[7] looked at a sample of 441 biomedical articles, finding only one made a full protocol available and none shared their raw data. Raw data often cannot be shared due to privacy and ethics reasons, though some benchmark datasets in related areas such as pharmacovigilance have been shared[8,9].

Reproducibility in biomedical informatics has primarily been studied in terms of how portable algorithms for identifying phenotypes are between institutions[10,11]. However, few studies have been done to evaluate the reproducibility of machine learning results across settings. The difficulty in reproducing studies is further compounded when we aim to infer causal relationships and not simply correlations. Here we must be sure that any effects we observe are not the result of hidden common causes, such as a latent variable causing both a treatment to be selected and outcome to be observed.

One work attempted to replicate a data-driven study of risk factors for congestive heart failure in two EHR systems[12]. However, when using existing, observational, data, there are often many differences between two settings – data formats, population characteristics, medical practice – and it can be impossible to disentangle which of these are responsible for differences in results. As a result, it was both logistically difficult to reproduce the main approach for identifying cases and controls (as different data were collected) and difficult to evaluate whether results were reproduced. Many of the same problems plague multi-site randomized controlled trials (MRCTs), where standardization across sites is critical to finding the true impact of an intervention[13]. Chesworth et al.[14] assessed treatment fidelity for a specific trial, finding that in many cases actual procedures differed substantially from the protocol. Spirito et al.[15] examined the role of differences in sample population and study protocol in a 6-site MRCT, and found that among other factors attrition varied by site and accounted for some outcome differences.

Medical simulations

The studies described in the previous section have shown difficulties in reproducing results between settings, and have retrospectively analyzed reasons for lack of reproducibility for individual experiments. However, the same incorrect finding can be made in two places, and when we are testing a new method, a failure to reproduce does not mean the method is incorrect (as the difference may be due to differences in data or the population). To systematically test how factors such as error in documentation or variation in protocol affect inference algorithms, we need 1) ground truth (so we know what should be found) and 2) variation (systematically changing features of the problem). Simulations allow us to replicate real-world scenarios such as a hospital stay, and enable one to learn about how outcomes would differ under various parameters (e.g. how mortality rate would differ if doctors had less variation in their processes). Importantly, all variables are observed and we have perfect ground truth in the simulation, and can test what happens in other cases (e.g. latent treatment) by removing some variables from the output analyzed.

Simulations have primarily focused on modeling spread of infectious diseases such as influenza[16], with less work on simulating medical data itself. The Observational Medical Dataset Simulator 2 (OSIM2)[17,18] created simulated longitudinal medical data, using real data as input to determine realistic population characteristics. The approach generates populations where individuals each have a set of diseases and treatment assignments, then simulates the effect of treatment on outcomes. This approach is similar to the one we take, using probabilistic modeling of treatments

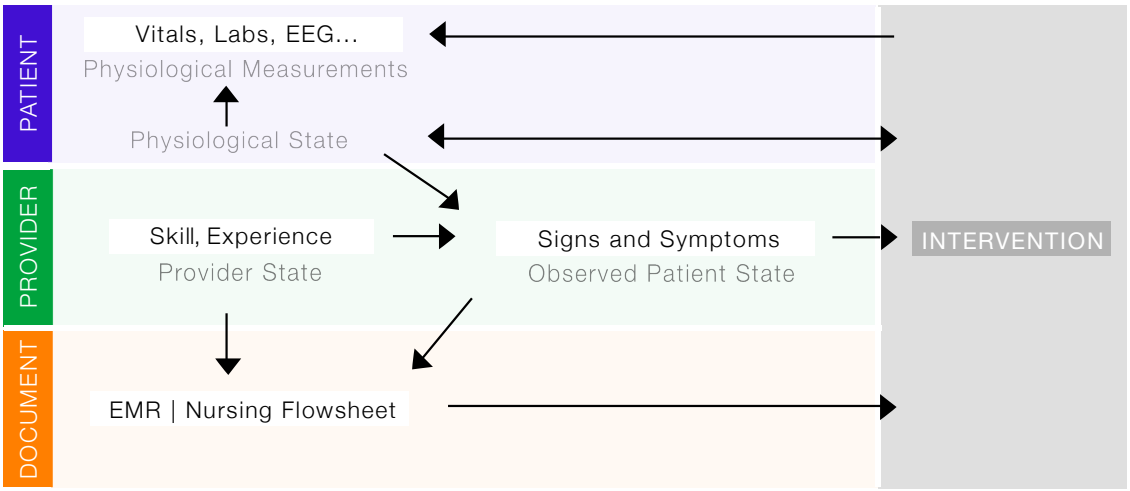


Figure 1: Overview of system. A patient has an underlying physiological state, measured with some error. A provider makes qualitative observations, observes the noisy measurements, and chooses interventions based on these. Interventions affect both underlying physiology and measurements. Finally, documentation is a noisy and incomplete recording of a patient's status, and affects intervention choice.

and outcomes and aiming to capture the confounding that makes inference so difficult. However, OSIM2 models only patients and treatments, and not the bias and latent variables that make causal inference challenging[19]. On the other hand, we aim to capture the heterogeneity and error introduced in the medical process through providers and documentation, and directly model unobserved factors that may affect outcomes.

Methods

In this section we describe the architecture of our simulation, data generation approach, and analysis methodology.

Simulation

We propose an agent-based approach to simulating medical interventions, which can capture the complex interactions between a patient's state, a provider's translation of their observations into interventions, and the documentation of these processes (that become input to other cognitive processes). One of the advantages of an agent-based approach is being able to simulate individual characteristics of providers, versus having the same parameters for all. This approach will ultimately allow us to model the interaction between providers (nurses, doctors) over time. An overview of the proposed approach is shown in figure 1. The idea is that we can clearly separate the different sources of error and bias so they can be systematically varied in combination and alone. A patient (purple segment at top) has a true underlying physiological state, which is distinct from measurements of this state (e.g., laboratory tests), which may be error-prone. In the clinician layer (green, middle), a clinician's assessment of a patient is dependent on their skill and experience, and these plus noisy measurements and documentation guide intervention choices. Thus we can simulate how an error in a note could lead to errors in intervention (e.g., incorrect medication dosage). Finally, EHRs and other documentation are not simply a faithful recording of all events, but depend on skill and experience, and this documentation (orange, bottom layer) feeds back into intervention choice.

In this work we focus on the middle layer in the diagram: modeling provider features. We are interested primarily in features of the medical process that can lead to erroneous inferences, rather than the specifics of any particular illness. Thus we develop an agent-based simulation to model interventions at a high-level, using Repast Simphony, an open-source toolkit for agent-based simulations.[1] To simplify the problem, we focus on the case where each patient has a single medical encounter with a single provider, and can receive (or not) an observable treatment and can also receive (or not) a latent treatment. A latent treatment is one that is not documented (unobserved). The effect of both the latent

[1] https://repast.github.io/index.html

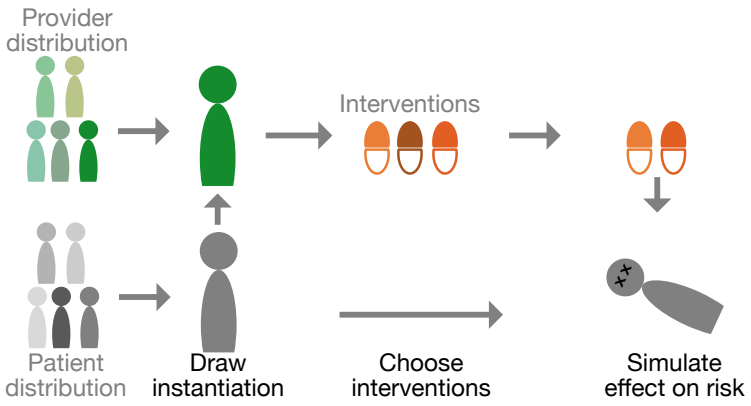


Figure 2: Detail of simulation process. Individual patients and providers are drawn randomly from the relevant defined distributions. For patients, these are overall risk of death and for providers they are efficacy of interventions. Then, at each time step instantiations of treatment are drawn from the distributions. Patient, provider, and treatment characteristics all modify overall risk rate.

and observed treatments can vary in a provider or site dependent way, enabling simulation of some of the challenges of multi-site RCTs. For example, there may be systematic differences in standard of care or adherence to protocols, which can confound results, such as finding no difference between intervention and control if the intervention protocol has significant variation. While at a large scale such differences are assumed to average out, cluster randomized trials for example have lower effective sample size, as conditions within a site are not independent[20].

The flow of the simulation is shown in figure 2. There is a set of patient and provider agents, each of whom has their own features drawn from a distribution of distributions. Instead of describing a treatment in terms of only its mean effect and standard deviation (s.d.), we draw the mean and s.d. from distributions. For each encounter between a patient and provider, the intervention effects for latent and observed interventions are chosen from the intervention distributions. Finally, each patient's outcome is simulated by combining the factors that modify their relative risk (personal risk rate, latent treatment, observed treatment). In this work, the outcome for each patient at the end of the simulation is survival, so each factor modifies risk of mortality.

The key parameter distributions in the simulation are as follows:

- Risk of death (mean and standard deviation)
- Treatment effect mean (mean and standard deviation)
- Treatment effect standard deviation (mean and standard deviation)
- Latent effect mean (mean and standard deviation)
- Latent effect standard deviation (mean and standard deviation)

Again, rather than a single distribution for each treatment parameter, the key parameters themselves vary according to user-defined distributions. This allows groups of patients (risk of death) and groups of providers (treatment and latent effects) to be modeled automatically. For example, if there is one group of providers being studied, such as in a single center study, the treatment effect would draw upon a single mean and a single standard deviation. If there are several groups of providers being studied, such as a multi-center study, then each center would have a treatment mean with an associated standard deviation. We can choose the four variables that affect treatment mean and treatment standard deviation to describe the overall distributions that inform each center's treatment effect parameters.

We can thus simulate effects which are uniform or highly variable (normally centered around a single mean average effect) and which are evenly or unevenly applied (also normally centered around a single mean standard deviation).

Parameter	No Tx effect, vary latent	Low/Med/High Tx effect, vary latent	Med Tx, vary latent s. d.
Risk of death mean	0.05 (0.0005)	0.05 (0.0005)	0.05
Risk of death s. d.	0.003	0.003	0.003
Tx effect mean (mean)	1	0.9, 0.5, 0.3	0.5
Tx effect mean (s. d.)	0	0	0
Tx effect s.d. (mean)	0.03	0.03	0.03
Tx effect s.d. (s. d.)	0	0	0
Latent effect mean (mean)	[0.1:0.1:1]	[0.1:0.1:1]	0.7
Latent effect mean (s. d.)	0	0	0
Latent effect s.d. (mean)	0.01	0.01	[0.5:0.5:2.5]
Latent effect s.d. (s. d.)	0	0	0

Table 1: Parameters across experiments. The two risk of death mean values shown as high mortality (low mortality in parentheses) are constant across experiments. Brackets indicate iteration with a given step size.

For example, we can simulate an intervention that has minimal variation in its standard deviation at each site, but large variations in its mean effect. The treatment effects and risk of death are expressed as relative risk (RR), which is the probability of death under exposure divided by the probability in a non-exposed group. An RR of 1 leaves risk unchanged, while effects < 1 reduce risk and those > 1 increase it.

The latent treatment effect can be used to express a number of unrecorded or unobserved qualities, such as actual unobserved treatments or factors affecting outcome such as provider skill. For example, a latent effect mean of 1 and standard deviation of 0 mean the provider outcome is exactly the same as that of the treatment. However, one can also imagine cases with mean < 1 but wide standard deviation, meaning that on average the provider's outcomes improve upon the treatment effect, but that there is high variance.

In reality, many important factors appear as latent for varying reasons. In some cases data may not be recorded or available, such as when using only claims data rather than the full EHR. In other cases, data on interventions may be recorded with a delay, so we do not know when a treatment happened relative to a change in a patient's condition and cannot use such data to identify a causal relationship. With this simulation we aim to be able to systematically vary such factors, to test how well algorithms can handle these cases and develop new approaches that are more robust.

Data simulated

To demonstrate our approach, we create three types of datasets that illustrate features of our simulation and the difficulties posed for even simple comparisons. In each experiment we simulate 1000 patients, varying the risk of death from the illness itself (high vs. low mortality rate) as well as the efficacy of both the latent and observed treatments. Due to the randomness inherent in the simulation, we run 10 simulations for each parameter setting and average the results. Parameters for the three datasets are shown in table 1. In all cases, risk of death is given as a relative risk. Risk of death is that at each time step, so that for a simulation with t time steps, the actual risk is $(1 - \text{risk})^T$. In this work we use $t = 10$, so the high mortality settings have an overall mortality rate of 40%, across the 10 time steps. The primary outcome of each simulation is the mortality rate, which we compare across simulations, aiming to replicate what happens in a multisite trial. We aim to determine under what conditions we can find statistically significant differences between groups, and how much unmeasured latent effect is required to draw erroneous conclusions.

Varying latent treatment effect: We begin with the common case of determining whether there is a difference in mortality rate between two conditions. In this example we fix the treatment as having no effect on risk of death, and compare two population risk of death means (0.05 for high mortality, 0.005 for low mortality) while varying how effective the latent treatment is (iterating from 0.1 to 1.0 by 0.1). This case can occur when there are unknown risk factors for a disease, such as environmental exposure, that differ between sites in a multi-site trial.

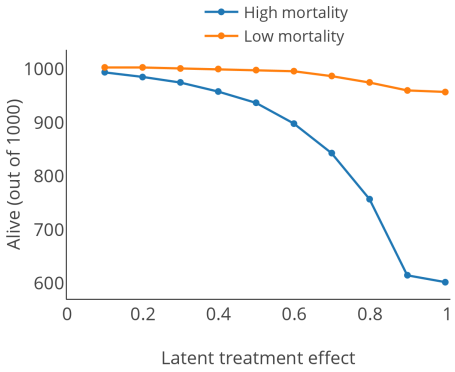


Figure 3: Varying latent treatment effect with two mortality rates and ineffective observed treatment.

Mixed latent and observed treatment effects: Once again we use a sample size of 1000 patients, but now include an observed treatment effect. All settings are the same as for the previous set-up, except now there are high (RR=0.3), medium (RR=0.5), and low (RR=0.9) observed treatment effects simulated in each of the high and low mortality conditions. As before, we iterate over latent treatment effects from 0.1 to 1.0 by increments of 0.1 and average the results of 10 runs with each setting.

Constant mean effect, varying standard deviation: Finally, we examine the effect of variance in latent treatment effects. Here we fix mortality as high (0.05) and treatment as moderately effective (RR=0.5), and use a latent effect of 0.7, based on results of the first two simulations. These parameters are now fixed as we iterate over latent effect standard deviation mean from 0.5 to 2.5 by increments of 0.5. Once again we use 1000 patients in each run and 10 runs for each parameter setting.

Analysis

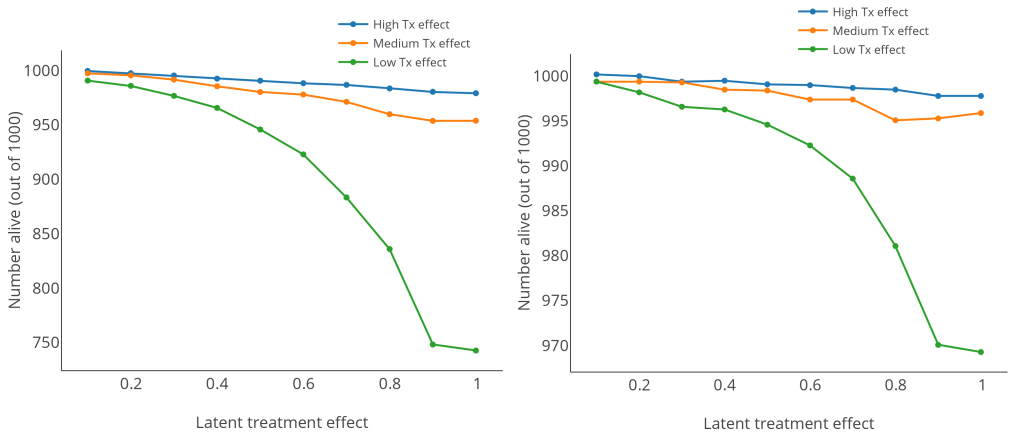
The primary outcome from each dataset generated is the mortality rate. We first aim to compare groups to determine if we can distinguish between more and less effective treatments. We test statistical significance using a Fisher exact test and threshold of $p < 0.05$ for statistical significance. We also assess observed study power (using actual mortality rate of the simulations and other simulation parameters), as it can differ from expected power.

Results

We now examine results for each of the three simulated scenarios, where the primary outcome is mortality and we aim to determine whether there is a statistically significant difference in mortality rate between treatment groups.

Effect of latent treatment

In the first case, we simulated a totally ineffective treatment, but varied the latent treatment and base risk of mortality. In this case we ask, at what point will there be a statistically significant difference between groups? Figure 3 shows the total number of survivors (out of 1000 patients) as we vary how effective the latent treatment is, holding constant an ineffective observed treatment under the two mortality conditions (high and low). When base mortality rate is high, the latent treatment has a substantial impact on total number of deaths, and a statistically significant effect on mortality once RR=0.8. This is because with a higher mortality rate, there is more opportunity to show a difference. On the other hand when mortality rate is low, it takes a much stronger latent treatment effect to show a difference in mortality (compared to a treatment effect of 1, meaning no impact). It is only once the relative risk reaches 0.7, that the difference between that and the condition where the latent treatment has no effect (RR=1) reaches statistical significance using a Fisher exact test, though the effect size remains small. This means that if we are comparing two groups with a high mortality condition (e.g. such as stroke), even if the treatment we are interested in has no effect on outcome, a relatively modest latent effect may confound our results.



(a) High mortality **(b)** Low mortality

Figure 4: Effect of mixed latent and observed treatment effects, under two different population mortality rates. Note that y-axes differ so that the different outcomes in the low mortality condition will be visible.

Mixed latent and observed treatment effects

Now we combine observed and latent treatments, studying the impact of latent treatment effects with varying efficacy of observed treatments. Results are shown in figure 4. Different y-axis scales are used to better highlight subtle differences within the low mortality condition. First, within the high mortality condition, when the latent treatment effect is at least 0.6, the high and medium treatment conditions are no longer statistically distinguishable ($P > 0.14$). That is, even though there is an actual difference between these two treatments, if these were two groups in an RCT, we would no longer be able to find a difference between them because of the impact of the unobserved latent treatment. Further, in the low treatment effect condition, the latent effect makes the intervention seem to have a bigger effect on mortality than it actually does, which could lead to erroneous conclusions about the efficacy of an intervention.

In the low mortality condition (note that the range for the Y-axis is 970-1000), none of the differences between high and medium treatment effect are significant (P-values > 0.6), as the number of deaths ranges from 0 to 4. Thus, even though the observed treatments have a different effect on mortality in the ground truth of the simulation, we are unable to distinguish between them. Further, once the latent effect reaches 0.6, the medium and low treatment effects are no longer distinguishable (P-value > 0.2). The fact that the effects are indistinguishable does not represent the fact that the effects are somehow the same (the ground truth is that they are different). It simply means that the experiments are underpowered to detect a difference among them. If we were to increase the total number of patients in the experiment we would eventually reach a number where we could detect these differences.

Using a standard power calculation, the power for a study with 1000 patients in each group (high and medium efficacy treatment) using the actual true effect sizes, is 97.6% with an alpha of 0.05. Thus power calculations that ignore latent effects may lead to significantly underpowered studies. The actual observed power for detecting a change between these treatment groups, as a function of the strength of the latent effect is shown in figure 5. Once the latent effect reaches 0.7, power drops significantly, and critically is below the commonly used 0.80 threshold. Comparing the low and moderate treatments, power drops into an unacceptable range when the latent treatment effect reaches 0.4. We propose that power calculations may be augmented with simulations such as ours to explore how various unmeasured factors can affect effective power, and provide a more realistic estimate of the necessary sample size.

Effect of variance

Finally, we fix both treatment (RR=0.5) and latent effects (RR=0.7) along with mortality rate and now vary the standard deviation for the latent effect. This captures the scenario when the latent effect is incorporated in power calculations, but its variation is not properly accounted for. In most cases, we assume the treatment will have the same effect in different settings, but provider characteristics may vary in important ways between sites such as degree of adherence

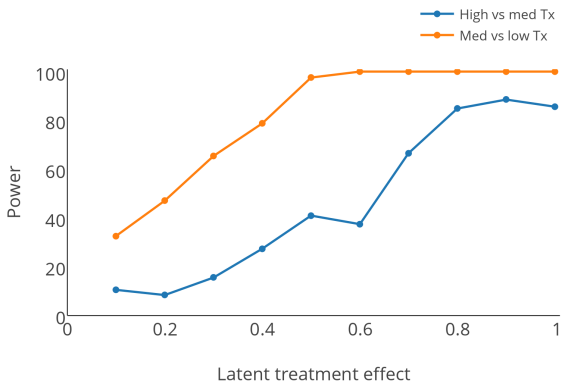


Figure 5: Actual power using $\alpha = 0.05$, when comparing two treatments, as a function of latent treatment effect's impact on relative risk.

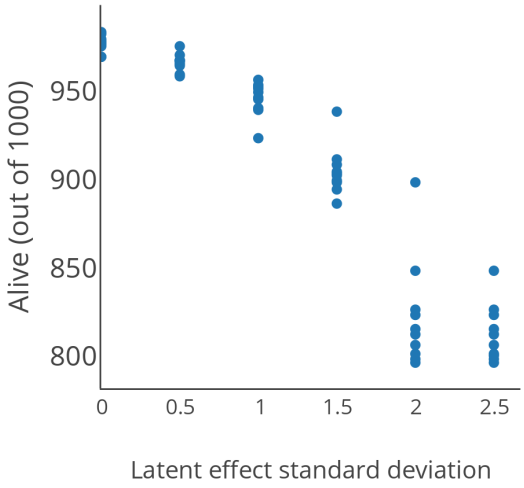


Figure 6: Result of variation in latent effect on mortality rate.

to protocol, level of experience, or workload. Figure 6 shows the result, where increased variation in treatment effect leads to significant differences in mortality rates within and across experiments. With zero standard deviation (so treatment effect is always mean of 0.5 and latent effect is mean of 0.7), mortality averages less than 3%, but this increases to 10% with s.d. 1.5, and 19% with s.d. 2.5. If these were multiple sites in an RCT, then significant differences may seemingly be observed even though the treatments have exactly the same mean effect in every experiment.

Discussion

Our simulations suggest that traditional sample size calculations may lead to underpowered studies in cases where there are latent effects or multiple sites. Further, due to the potentially complex interaction of multiple latent effects, some of which may be positive (high adherence to protocol) and some negative (high variance in clinician skill level), sample size calculations that account for these interactions may be prohibitively complex. Even when actual characteristics of an observed intervention are completely known (versus hypothesized, as in actual practice), we showed that when latent factors are not considered, studies become underpowered as latent effect size grows. Yet, these factors are rarely considered. One review found that only 41% of a set of pediatric ICU RCTs considered these nuisance parameters, leading to reduced estimates of variance and lower sample sizes than would actually be needed[21]. Another review found that many studies were either underpowered or overpowered (which can waste resources and time, and expose more patients to risks) due to incorrect assumptions about nuisance parameters[22], which usually are not precisely known. While computational work on causal inference has aimed to identify latent confounders, the effects observed

here are not of the usual form considered (where a latent common cause may lead to inferring spurious relationships between its effects). This is exactly the type of variation that the randomization in RCTs aims to remove the effect of, but in research (1) from EHRs or (2) multi-site trials where effective sample size is much smaller due to site-related dependence, we cannot easily remove such effects through experiment design.

Recommendations

We propose that there are some key benefits for the simulation of medical processes and make suggestions for future work. First, while we focus here on a single latent treatment, the simulation introduced allows any number of such factors, enabling exploration of the impact of multiple latent effects which may have nonlinear interactions. This can enable better understanding of the effect different variances, treatment effects, and interactions have on sample size determinations. We propose that this can be used to complement standard power analysis and explore how effective sample size may differ under various conditions. Further this approach may be useful for analyzing RCT results (both the trial data and associated EHR data) to distinguish between intervention effects and other effects. As the simulation is extended to include other factors in EHR recording, latent factors like physician variation may have an identifiable signature. Second, it is customary to account for the effect of known co-interventions in any study. However normally only aggregate measures are available and used for analysis. With the advent of EHR documentation, one can envision a study where all of the co-interventions, known or unknown, are documented. Unfortunately, our ability to use this data is limited. The correctness of time varying analyses that aim to account for variations in co-interventions will need to be tested. Only with data for which ground truth is known can the validity of these algorithms be truly tested. By extending our simulation to include intermittent documentation of latent treatments, one can validate these time varying algorithms against with ground truth. Finally, simulation can be used to estimate the distribution of these latent effects to create data patterns that are similar to real world EHR data patterns and thus provide possible hypotheses of the underlying data generation processes that then can be tested. Ultimately, this approach will separate the underlying truth of a patient's state from our sporadic and noisy measurement and documentation of this state, allowing better use of EHRs for research.

Limits of Simulation

The key limitation of our simulation approach is that it is meant to test the behavior of algorithms and contain similar structure to real scenarios, but it cannot be used to answer counterfactual queries about specific treatments. That is, we aim to capture the structure of the system (e.g. features leading to error) rather than to create an exact model of a specific disease process. Thus it cannot be used for developing intervention strategies. Further, some confounding factors may not be removable in practice. However, if they can be explicitly documented, it may be possible to account for their effects using EHR data and computational algorithms. In addition, we have chosen to use normally distributed parameters to model risk of disk and effects, for the sake of simplicity and because RCT power calculations are often based on such distributions. It is possible that the actual structure of the effects have bimodal or multimodal effects, which should be studied in future extensions to this work.

Acknowledgments

This work was supported in part by the NLM of the NIH under Award Number R01LM011826.

References

1. Ioannidis JP. Why most published research findings are false. PLoS medicine. 2005;2(8):e124.
2. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery. 2011 8;10(9):712.
3. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012;483(7391):531–533.
4. Klein RA, Ratliff KA, Vianello M, Adams Jr RB, Bahník S, et al. Investigating Variation in Replicability. Social Psychology. 2014;45(3):142–152.

5. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349(6251):aac4716.

6. Young SS, Karr A. Deming, data and observational studies. Significance. 2011;8(3):116–120.

7. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JPA. Reproducible Research Practices and Transparency across the Biomedical Literature. PLoS Biology. 2016 01;14(1):1–13.

8. Harpaz R, Odgers D, Gaskin G, DuMouchel W, Winnenburg R, Bodenreider O, et al. A time-indexed reference standard of adverse drug reactions. 2014 11;1:140043 EP –.

9. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG; Janssen Research; Room K30205 PO Box 200 Titusville NJ 08560 USA ryanomop.org. Development LLC. Defining a reference set to support methodological research in drug safety. Drug Saf. 2013 10;36 Suppl 1:S33–47.

10. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. Journal of the American Medical Informatics Association. 2012;19(e1):e162–e169.

11. Overby CL, Pathak J, Gottesman O, Haerian K, Perotte A, Murphy S, et al. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. Journal of the American Medical Informatics Association. 2013;20(e2):e243–e252.

12. Kleinberg S, Elhadad N. Lessons learned in replicating data-driven experiments in multiple medical systems and patient populations. In: AMIA Annual Symposium Proceedings; 2013. p. 786–795.

13. Borrelli B. The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials. Journal of public health dentistry. 2011;71(s1):S52–S63.

14. Chesworth BM, Leathley MJ, Thomas LH, Sutton CJ, Forshaw D, Watkins CL. Assessing fidelity to treatment delivery in the ICONS (Identifying Continence OptioNs after Stroke) cluster randomised feasibility trial. BMC Medical Research Methodology. 2015;15(1):1–9.

15. Spirito A, Abebe KZ, Iyengar S, Brent D, Vitiello B, Clarke G, et al. Sources of site differences in the efficacy of a multisite clinical trial: the Treatment of SSRI-Resistant Depression in Adolescents. Journal of consulting and clinical psychology. 2009;77(3):439.

16. Chao DL, Halloran ME, Obenchain VJ, Longini Jr IM. FluTE, a publicly available stochastic influenza epidemic simulation model. PLoS Comput Biol. 2010;6(1):e1000656.

17. Murray RE, Ryan PB, Reisinger SJ. Design and validation of a data simulation model for longitudinal healthcare data. In: AMIA Annual Symposium Proceedings. vol. 2011. American Medical Informatics Association; 2011. p. 1176–1185.

18. Ryan PB, Schuemie MJ. Evaluating performance of risk identification methods through a large-scale simulation of observational data. Drug safety. 2013;36(1):171–180.

19. Gruber S. A Causal Perspective on OSIM2 Data Generation, with Implications for Simulation Study Design and Interpretation. Journal of Causal Inference. 2015;3(2):177–187.

20. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. Bmj. 2004;328(7441):702–708.

21. Nikolakopoulos S, Roes K, van der Lee JH, van der Tweel I. Sample size calculations in pediatric clinical trials conducted in an ICU: a systematic review. Trials. 2014;15(1):274.

22. Tavernier E, Giraudeau B. Sample Size Calculation: Inaccurate A Priori Assumptions for Nuisance Parameters Can Greatly Affect the Power of a Randomized Controlled Trial. PloS one. 2015;10(7):e0132578.