HHS Public Access
Author manuscript
Behav Res Ther. Author manuscript; available in PMC 2020 October 14.

Published in final edited form as:
Behav Res Ther. 2019 September ; 120: 103412. doi:10.1016/j.brat.2019.103412.

Machine learning methods for developing precision treatment rules with observational data

Ronald C. Kessler[a,*], **Robert M. Bossarte**[b,c,d], **Alex Luedtke**[e,f], **Alan M. Zaslavsky**[a], **Jose R. Zubizarreta**[a,g]

[a]Department of Health Care Policy, Harvard Medical School, Boston, MA, USA

[b]West Virginia University Injury Control Research Center, Morgantown, WV, USA

[c]Department of Behavioral Medicine and Psychiatry, West Virginia University, Morgantown, WV, USA

[d]VISN 2 Center of Excellence for Suicide Prevention, Canandaigua VA Medical Center, Canandaigua, NY, USA

[e]Department of Statistics, University of Washington, Seattle, WA, USA

[f]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[g]Department of Statistics, Faculty of Arts and Sciences, Harvard University, Cambridge, MA, USA

Abstract

Clinical trials have identified a variety of predictor variables for use in precision treatment protocols, ranging from clinical biomarkers and symptom profiles to self-report measures of various sorts. Although such variables are informative collectively, none has proven sufficiently powerful to guide optimal treatment selection individually. This has prompted growing interest in the development of composite precision treatment rules (PTRs) that are constructed by combining information across a range of predictors. But this work has been hampered by the generally small samples in randomized clinical trials and the use of suboptimal analysis methods to analyze the resulting data. In this paper, we propose to address the sample size problem by: working with large observational electronic medical record databases rather than controlled clinical trials to develop preliminary PTRs; validating these preliminary PTRs in subsequent pragmatic trials; and using ensemble machine learning methods rather than individual algorithms to carry out statistical analyses to develop the PTRs. The major challenges in this proposed approach are that treatment are not randomly assigned in observational databases and that these databases often lack measures of key prescriptive predictors and mental disorder treatment outcomes. We proposed a tiered case-

[*]Corresponding author. Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA, 02115, USA. Kessler@hcp.med.harvard.edu (R.C. Kessler).

Declarations of interest
In the past 3 years, Dr. Kessler received support for his epidemiological studies from Sanofi-Aventis; was a consultant for Johnson & Johnson Wellness and Prevention, Sage, Shire, Takeda; and served on an advisory board for the Johnson & Johnson Services Inc. Lake Nona Life Project. Kessler is a co-owner of DataStat, Inc., a market research firm that carries out healthcare research. Zubizarreta was a consultant for Johnson & Johnson Real World Data Analytics and Research, New Brunswick, NJ. The remaining authors have no financial disclosures.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

cohort design approach that uses innovative methods for measuring and balancing baseline covariates and estimating PTRs to address these challenges.

Keywords

Clinical decision support; Ensemble machine learning; Personalized treatment; Precision treatment; Super learner

Evidence-based treatments for mental disorders continue to grow in number and complexity. But none of these treatments is optimal for all patients. Furthermore, it remains unclear *which treatments work best for which patients*. As a result, clinicians need to rely on trial-and-error in selecting treatments, and patients often need to suffer through multiple ineffective courses of treatment before achieving remission. A substantial proportion of patients give up before an effective treatment is found, sometimes with tragic results, even though treatments exist that would have led many of these patients to remit. The development of comprehensive precision treatment rules (PTRs) would be of enormous value in addressing these problems.

The goals of precision medicine are to understand how the effects of treatment are modified by patient characteristics and to develop PTRs based on this understanding to optimize treatment selection for each patient or fine-grained patient subgroups. PTRs are developed using information about *prescriptive* predictors of treatment response, by which we mean predictors of the relative effectiveness of different treatments. Numerous clinical features (e.g., illness persistence, severity, subtypes, and comorbidity), socio-demographic characteristics (e.g., patient age, sex, education), and biomarkers (e.g., genetic, neuroendocrine, electrophysiological, brain imaging) have been found to be reliable prescriptive predictors of some mental disorder treatments versus others (Kessler, 2018; Olbrich, van Dinteren, & Arns, 2015; Prendes-Alvarez & Nemeroff, 2018; Saltiel & Silvershein, 2015). However, the associations of the individual predictors with the outcomes in these studies have been too small to be of clinical value when considered one at a time, leading to an interest in developing composite prescriptive prediction scores containing information about a range of individually-significant predictors (DeRubeis et al., 2014; Fabbri, Crisafulli, Calabro, Spina, & Serretti, 2016; Trivedi et al., 2016).

We were asked by the issue editors to discuss the use of machine learning (ML) methods to develop PTRs. Several reviews have been written about these methods both in the general medical literature (Lipkovich, Dmitrienko, & D'Agostino, 2017; Ondra et al., 2016) and the literature on mental disorder treatment (Bzdok & Meyer-Lindenberg, 2018; Menke, 2018). However, much less has been written about the research designs needed to implement these methods or the statistical issues involved in using these designs (Kessler, 2018). We begin with a discussion of these issues, as they are of critical importance in developing PTRs for mental disorders. We then turn to a discussion of the ensemble ML approach we recommend for developing PTRs.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1. Sample size requirements for developing precision treatment rules

We focus initially on research designs because the sample sizes needed to develop PTRs for mental disorders are larger than those that are realistic for controlled clinical trials of these disorders. We have to appreciate that new research designs are needed before we can consider optimal analysis methods. Most comparative effectiveness research on prescriptive predictors of treatment response is based on randomized clinical trials. Predictors in these studies are assessed either at baseline or early in the treatment process. The possibility that a variable is a prescriptive predictor (often referred to as a “moderator”) is evaluated by computing an interaction between the predictor and treatment type (VanderWeele, Luedtke, van der Laan, & Kessler, 2019). Significant interactions are interpreted as evidence that the variable is a prescriptive predictor; that is, that the comparative effectiveness of the alternative treatments differs depending on the predictor. When multiple interactions are statistically significant, a composite PTR is typically created by computing predicted outcome scores for each patient under each treatment regimen based on the coefficients in a multivariate interaction model and then comparing treatment-specific predicted outcome scores at the individual level to select the regimen with the best predicted outcome for each patient (Kovalchik, Varadhan, & Weiss, 2013). This logic is easily extended to more than two treatment options.

Although this simple interaction modeling approach is only one of a great many different statistical approaches that have been developed in recent years to generate individual-level composite prescriptive prediction scores for purposes of personalized treatment planning, all such approaches implicitly estimate interactions and are subject to the constraint that interaction tests have much lower statistical power than tests of main effects (Greenland, 1983). An important implication of this fact is that relatively large samples, shown in a recent simulation to be at least 500 patients per treatment arm, are required if we want to have adequate statistical power to develop accurate PTRs of a strength that is realistic to expect in research on mental disorder treatments (Luedtke, Sadikova, & Kessler, 2019). Even larger samples are needed if the goal is to study the genetic architecture of differential treatment response (Wigmore et al., 2019).

Most mental disorder precision treatment trials are based on much smaller samples than 500 patients per arm (e.g., Bousman, Arandjelovic, Mancuso, Eyre, & Dunlop, 2019; Cohen & DeRubeis, 2018; Trivedi et al., 2016). The researchers who carry out these trials are often aware of the power problem but argue that preliminary PTRs can be developed in these small studies and then tested in larger experimental samples (Petkova et al., 2017). That argument is flawed, as under-powered studies are likely to detect only the largest interactions and will be unable to create stable multivariate PTRs (Judd, Westfall, & Kenny, 2017). Increasing the accepted Type 1 error rate is a strategy sometimes used to identify more interactions as significant in these underpowered studies (e.g., DeRubeis et al., 2014), but this is counter to the usual advice to tighten criteria in reaction to multiplicity of possible tests. The result is over-fitting and reduction in out-of-sample performance when model results are used to make future predictions (Durand, 2013). Internal pseudo-replication methods exist to evaluate the extent of out-of-sample prediction shrinkage of the PTRs (Smith, Seaman, Wood, Royston, & White, 2014). But these methods do not fix the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

problem; they merely confirm its existence. The inevitable conclusion is that samples much larger than those in existing mental disorder randomized clinical trials are required to develop useful PTRs.

2. Using observational data to address the sample size problem

The problem of small sample size has been addressed in other areas of medicine either by using very large trials to develop composite PTRs (Dorresteijn et al., 2011) or by applying and validating external PTR scores in smaller trials after they were developed in large observational samples (Perel, Edwards, Wentz, & Roberts, 2006; Prieto-Merino & Pocock, 2012). The latter would be the more practical approach for the development of mental disorder PTRs. Although estimates of treatment effects are biased in observational studies if treatment assignment is informatively nonrandom, statistical methods exist to adjust for this bias if the baseline (i.e., prior to initiating treatment) assessments include the important determinants of treatment assignment, as the differences in these baseline covariates can be “balanced” statistically to approximate the distributions found in experimental trials (Hirshberg & Zubizarreta, 2017; Zubizarreta, 2015). We discuss this approach here because we consider it a necessary approach to develop mental disorder PTRs.

It has been shown that analyses of balanced databases often yield aggregate results about comparative treatment effects very similar to those obtained in randomized clinical trials (Anglemyer, Horvath, & Bero, 2014; Dahabreh et al., 2012). Extensions exist to develop PTRs (Luedtke & van der Laan, 2017; Zhou, Mayer-Hamblett, Khan, & Kosorok, 2017; Zhu, Zhao, Chen, Ma, & Zhao, 2017). This approach is likely to be of enormous practical value in dealing with the small sample sizes of mental disorder randomized clinical trials, as large observational studies could be carried out at a fraction of the cost of randomized clinical trials. Pragmatic trials based on smaller samples could then be used inexpensively to evaluate the validity of the PTRs developed in the observational studies.

3. Methods to balance observed baseline covariates in observational studies

The large observational studies that are used for comparative effectiveness research in other areas of medicine typically are electronic medical record (EMR) databases (Berger et al., 2017; Ning, Hong, Li, Huang, & Shen, 2017). Such studies are appealing because, unlike randomized clinical trials, the patients in these observational studies are those that exist in the population and the treatments studied are those that are delivered in the population. Matching and weighting are the preferred statistical approaches to balance on baseline differences in observed covariates across groups of patients exposed to different treatments in these studies. Although both matching and weighting have advantages, matching is the more intuitive of the two in that it maintains the unit of analysis intact in an attempt to approximate a randomized experiment hidden within an observational study. Matching also makes it much easier than weighting to carry out certain sensitivity analyses to hidden biases (Rosenbaum, 2017). Matching can be performed with one-to-one or one-to-many fixed or variable matching structures (Hansen, 2004). Matching is often done on a onedimensional summary of the covariates, as in matching based on estimated propensity scores

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(Rosenbaum & Rubin, 1983). However, more recently-developed matching methods balance directly on individual covariates and joint distributions depending on the nature of the data at hand (e.g., Diamond & Sekhon, 2013; Zubizarreta, 2012). These methods are reviewed by Visconti and Zubizarreta (Visconti & Zubizarreta, 2018).

Weighting, in comparison, has advantages over matching in statistical efficiency and computational tractability. Weighting also uses a larger proportion of cases than matching. Two broad approaches exist to weighting (Hirshberg & Zubizarreta, 2017). The first estimates a prediction model for type of treatment and then inverts the individual-level predicted probabilities from this model to weight the data. This inverse probability weight (IPW) approach is used in conventional logistic regression analysis (Cole & Hernán, 2008) and in variants based on more flexible ML methods (e.g., Gruber, Logan, Jarrin, Monge, & Hernan, 2015). But this approach suffers from two practical problems: that balancing can be very difficult to achieve due to model misspecification, small samples, and/or sparse covariates, leading to biased estimates of treatment effects (Imai & Ratkovic, 2014); and that the weights can be highly variable, producing unstable estimates of treatment effects (Kang & Schafer, 2007).

The second approach to weighting uses one of a number of recently-developed methods that weight directly to reduce covariate imbalance and weight dispersion across treatment groups (Chan, Yam, & Zhang, 2016; Hainmueller, 2012; Zubizarreta, 2015). These methods can achieve exact balance across many covariates in large samples, but at the expense of creating high weight variance. Some of these methods can alternatively achieve *approximate* balance while minimizing weight variance (Zubizarreta, 2015). This has great appeal in allowing researchers to make principled decisions about the trade-off between bias and efficiency in estimation. Related methods are reviewed by Yiu and Su (Yiu & Su, 2018). Wang and Zubizarreta (Wang & Zubizarreta, 2019) study the formal properties of this balancing approach to weighting and explain its connection to the classical modeling approach.

It should be noted that conventional linear regression analysis in which baseline covariates are used as control variables can be conceptualized as a special form of weighting in which the implicit weight perfectly balances the means of the covariates included in the regression. However, these implicit weights can take negative values, resulting in extrapolation of estimated treatment effects outside the bounds of observed data. Because of this feature, matching or explicit weighting are preferred to control variable analysis in attempting to balance observed baseline covariates. This issue is discussed in more detail by Hirshberg et al. (Hirshberg, Maleki, & Zubizarreta, 2019).

4. Methods to balance unobserved baseline covariates in observational studies

Although the methods described in the last section can be very useful when the observed baseline covariates include the key determinants of treatment assignment, they do not deal with unmeasured confounders. The best way to do that, of course, is with a large randomized clinical trial. When this is impossible, though, it is sometimes possible to find natural variation in access to treatment in a large sample that mimics an experiment (Handley, Lyles,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

McCulloch, & Cattamanchi, 2018). Many opportunities exist of this sort to study policy interventions using before-after ecological designs. For example, 30 different studies were reported over the past decade that examined the aggregate effects of interventions to reduce suicide by means of restrictions of various kinds, such as prohibiting firearms from being taken home by reservists in the Israeli Defense Force (Lubin et al., 2010) and restricting access to lethal pesticides in Taiwan (Lin & Lu, 2011). These studies are reviewed by Zalsman et al. (Zalsman et al., 2016). Other studies investigated the effects of interventions that changed exposure to financial stress in before-after surveys of populations in which financial stress was either increased for a segment of the sample (e.g., a plant closing in one of two towns, each of which had only one large employer; Penkower, Bromet, & Dew, 1988) or decreased for a segment of the sample (e.g., a casino opening on an Indian reservation leading to tribal members but not others in rural North Carolina receiving ongoing income supplements; Costello, Erkanli, Copeland, & Angold, 2010). Although only aggregate intervention effects were investigated in most of these studies, PTRs could have been developed from these datasets using the methods described below in the section Using machine learning methods to develop precision treatment rules.

In the absence of such natural experiments, it is sometimes possible to make principled causal inferences about aggregate treatment effects by attempting to find an instrumental variable (IV) (Baiocchi, Cheng, & Small, 2014; Swanson, 2017) or a situation in which a regression discontinuity design can be used (Moscoe, Bor, & Bärnighausen, 2015; Venkataramani, Bor, & Jena, 2016). In its simplest form, an instrumental variable is a random encouragement to receive treatment that affects the outcome only through the treatment and consequently tends to balance both observed and unobserved covariates across groups defined by the instrument and identifies the effect of treatment (specifically, the average causal effect on the compliers to treatment assignment; see Angrist et al. [Angrist, Imbens, & Rubin, 1996] for details). A classic example is the work of McClellan et al. (McClellan, McNeil, & Newhouse, 1994) on the association between intensity of acute myocardial infarction (AMI) treatment and long-term survival. In that study, the authors used information about comparative distances between patient homes and hospitals that differed in intensity of treating AMIs as an IV, leading to the important conclusion that long-term survival was not strongly affected by greater use of catherization or revascularization among marginal cases.

Discontinuity designs are related in that they estimate the effects of treatments among patients in a close neighborhood of a cutoff of a variable that governs treatment assignment. When that threshold is inexact, the threshold can be considered an IV. A good example of an opportunity to use a discontinuity design is the US Veterans Health Administration (VHA) REACH VET suicide preventive intervention, which pinpoints and provides intensive case management to the 0.1% of VHA patients (about 35,000 patients each year) who are estimated by a previously-developed ML model to be at highest risk of sucide (VA Office of Public and Intergovernmental Affairs, 2017). The subsequent suicide rate among the 35,000 patients predicted by the ML model to be in the top 0.1% of suicide risk, all of whom received the intervention, could be compared to the suicide rate of the 35,000 additional patients who were predicted by the ML model to be in the second tenth of the first percentile (i.e., > 0.1–0.2%), none of whom received the intervention. The implicit assumption in such

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

a comparison would be these two groups of patients are approximately balanced by design on both observed and unobserved covariates by virtue of being in the neighborhood of the treatment threshold. A preliminary PTR could be developed based on this assumption to evaluate the effect of the intervention among patients in this neighborhood of risk. More formal estimates of aggregate intervention effects near the margin in a regression discontinuity design can be obtained using a modeling approach unique to this design either at the cutoff (Hahn, Todd, & Van der Klaauw, 2001) or in the neighborhood of the cutoff (Cattaneo, Frandsen, & Titiunik, 2015).

Both IV and discontinuity analyses can be conducted with either matching or weighting methods (Keele, Titiunik, & Zubizarreta, 2015; Zubizarreta, Small, Goyal, Lorch, & Rosenbaum, 2013) and combined with additional regression adjustments using more complex estimators (Robins & Rotnitzky, 1995). However, these designs are used currently only to estimate aggregate treatment effects, as it is unclear how to estimate PTRs in these designs unless one is willing to treat cases in the neighborhood of the threshold in the discontinuity design as if they were experimentally assigned. We consequently focus the discussion of PTR development later in this paper on observational studies in which the important determinants of treatment assignment are observed at baseline and balanced using either matching or weighting.

5. Measuring the important baseline covariates

Before turning to the discussion of ML methods to develop PTRs, it is important to be clear about a fundamental challenge in using large EMR databases for this purpose: obtaining accurate measures of the baseline covariates that are hypothesized to be the most important prescriptive predictors of treatment response. Many of these covarates are absent from large administrative databases. Three general approaches exist to augment the data available in EMR databases to overcome this challenge: collecting additional information on hypothesized prescriptive predictors in the same way as in the baseline assessment of a randomized clinical trial; linking other administrative databases to the EMR database; and engaging in an iterative process that combines the first two approaches. We comment briefly on each of these three in the remainder of this section.

6. Direct assessment in the full baseline sample

Measuring baseline covariates in a large observational database in the same way as in the baseline of a randomized clinical trial increases costs, which partially undercuts the advantages of working with a large observational database. However, these costs are nonetheless almost always much lower than in a randomized clinical trial because the other costs of this design are so low. For example, we are carrying out a large observational study comparing the outcomes of patients in primary care treatment of major depressive disorder (MDD) in VHA clinics with and without a measurement-based collaborative care (MBCC) program (Lipschitz et al., 2017). MBCC is available in only about 400 of the roughly 1100 VHA outpatient clinics in the country and is used to treat only about 30% of depressed PCP patients in those clinics due to limited availability of MBCC case managers. This makes it possible for us to select separate control samples of patients both in the same clinics and in

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

clinics where MBCC is not available to develop a PTR for the types of patients most likely to be helped by MBCC.

In order to do this, we are recruiting patients being treated with and without MBCC to fill out a web-based self-administered questionnaire and complete a series of performance-based neurocognitive tests at the beginning of treatment and then to complete a telephone interview three months later to assess short-term treatment outcome. The baseline assessments contain a wide range of patient-reported measures hypothesized to determine nonrandom assignment to MBCC as well as numerous patient-reported measures found to be significant prescriptive predictors of MDD treatment response in previous research (Kessler et al., 2017). This study is being carried out in a sample of 2000 patients (500 receiving MBCC, 500 others being treated in the same clinics with treatment-as-usual, and 1000 treated in clinics that do not have a MBCC program) at a cost only a fraction as high as the cost of a randomized clinical trial of the same size.

7. Linking other archival measures

A second way to augment the assessment of baseline covariates in large EMR studies is to link data at the patient level with other archival data sources that provide information about prescriptive predictors that are not assessed in structured EMR records. This approach is especially useful in situations where the required sample is so large that it is infeasible to carry out direct patient assessments. We faced a situation of this sort in designing a currently ongoing study to develop a PTR for hospitalizing VHA patients in the immediate aftermath of a nonfatal suicide attempt versus treat these individuals as outpatients. This is an important decision because even though hospitalization is the standard of care for patients after a suicide attempt (Hjorthoj, Madsen, Agerbo, & Nordentoft, 2014; U.S. Department of Veterans Affairs, 2013), the high suicide rate among recently-discharged psychiatric inpatients (Chung et al., 2017) is thought to be partly due to hospitalization causing trauma (Paksarian et al., 2014), humiliation (Svindseth, Dahl, & Hatling, 2007) and negative societal reactions (Kinard & Klerman, 1980). Based on these considerations, experts urge clinicians to weigh the potential risks and benefits of hospitalization carefully on a case-by-case basis before deciding whether to hospitalize a patient after a nonfatal suicide attempt (Large & Kapur, 2018). Yet little empirical guidance exists on how this should be done. This is a situation ideally suited to the development of a PTR to provide clinical decision support.

The key practical problem we faced in designing this study was that the sample size needed for powerful analysis was too large for direct baseline patient assessment. About 14,000 nonfatal suicide attempts occur in VHA each year (Hoffmire et al., 2016) and about 8% of these patients experience another suicide-related behavior (SRB; either suicide death or a repeat nonfatal suicide attempt) over the next 12 months. Power analyses based on simulation methods described elsewhere (Luedtke et al., 2019) show that the ability to develop a stable PTR that could lead to a 10% proportional improvement in treatment outcome compared to current practice (i.e., from 8% to 7.2% subsequent SRB) would require a sample of at least 5000 patients. It would be financially and logistically challenging to carry out a direct assessment in such a large sample of patients shortly after a nonfatal suicide attempt.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We began addressing this problem by referring to the limited literature on prescriptive predictors of response to clinical interventions for suicidal patients (Barbui, Esposito, & Cipriani, 2009; Bryan, Peterson, & Rudd, 2018; Harned et al., 2008; Huh et al., 2018; McMain et al., 2018) and the larger literature on prescriptive predictors of response to treatments for mental disorders associated with high SRB risk (Cohen & DeRubeis, 2018; Kessler, 2018; Kessler et al., 2017). The significant prescriptive predictors in these studies include various patient socio-demographics, psychiatric disorder and symptom profiles (e.g., persistence and severity of specific disorders, comorbidity), exposure to stress and adversity, and patient personality-temperament. We then found that a substantial proportion of these measures were not available in the structured EMR data. Based on this fact, we linked our EMR database with data from several other administrative data systems to increase our ability to assess the missing prescriptive predictors:

i. We extracted information from clinical notes in the Suicide Behavior Report for the initial suicide attempt using natural language processing (NLP) ML methods (a topic not covered in this paper, but see McCoy, Castro, Roberson, Snapper, & Perils, 2016; Rumshisky et al., 2016; Zhang et al., 2018) about the clinician's evaluation of the likely course of any triggering events for the suicide attempt, severity of the attempt, and perceived likelihood that the patient would make a repeat attempt if not hospitalized. All of these are likely to influence treatment decisions (i.e., lead to imbalance in baseline covariates that we can correct using matching or weighting once these variables are extracted based on the NLP analysis) as well as treatment outcome. More recent research suggests that NLP methods can also be used to estimate Research Domain Criteria (RDoC) scores from clinical notes and that these scores predict subsequent suicide deaths (McCoy, Pellegrini, & Perils, 2019).

ii. We linked the sample with several small-area geocode datasets that were developed and made available by researchers for all small areas (e.g., zip codes, Census Tracts) in the country. These datasets were created by aggregating information obtained from various kinds of administrative data or data from the US Census Bureau American Community Survey for all small areas in the country (American Community Survey, 2019). The goal was to create measures of neighborhood-level risk factors for SRB that might also predict non-random treatment assignment or prescriptive treatment response. Patient addresses were used for this purpose. We used measures of such potentially important prescriptive predictors as neighborhood disadvantage (Kind & Buckingham, 2018), neighborhood social capital (Rupasingha, Goetz, & Freshwater, 2006), local unemployment rate (Nordt, Warnke, Seifritz, & Kawohl, 2015), the local violent crime rate (Rosellini et al., 2017), and local firearms ownership rate (Anestis & Houtsma, 2018).

iii. We also linked the sample to patient-level data in the LexisNexis Social Determinants of Health (SDOH) database. LexisNexis is an electronic database company that maintains the world's largest repository of public records and credit information (LexisNexis, 2019). Their SDOH database contains more than 450 variables on various aspects of employment, finances, marital status,

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

parenting status, and involvement with the criminal justice system for close to 300 million Americans. The SDOH database was linked to patient records using basic demographic-geographic variables (e.g., date of birth and address).

In addition to these three types of patient-level data, we created measures of structural characteristics of the clinics and treatment centers in which patients were seen that might influence patient treatment assignment as well as treatment outcomes. Included here were measures of such things as number of available inpatient psychiatric beds in the nearest VHA inpatient facility, distance between the patient's home and the nearest VHA inpatient facility, and availability of a MBCC program in the nearest VHA outpatient clinic, all of which might influence the likelihood of the VHA Suicide Prevention Coordinator (SPC) deciding on outpatient treatment rather than hospitalization of a patient in the immediate aftermath of a suicide attempt. We also created a profile of the SPC's history of using outpatient treatment to manage past patients after suicide attempts. Finally, we created measures of VHA inpatient unit characteristics found in previous research to predict SRB after hospital discharge, such as staffing levels, staff turnover, and average length of stay (Kapur et al., 2016; While et al., 2012).

8. Combined use of archival data and direct assessment

What can be done when the researcher has a primary interest in prescriptive predictors that are expensive to collect, such as biomarkers or in-depth clinical assessments, and the study requires a very large sample size for powerful analysis? Projects like the UK Biobank (U.K. Biobank, 2018) and the Precision Medicine Initiative's All of Us Research Program (National Institutes of Health, 2019) provide one answer. Each of these is an omnibus research initiative that recruits very large panels of individuals (between 500,000 and 1 million) who agree to be followed over time to provide ongoing self-reports, biospecimens, and access to EMR data for purposes of monitoring their health and well-being. The data collected in these initiatives will be made available to many different researchers working on many different projects to improve prevention, diagnosis, and treatment of many different disorders. A much more limited, but nonetheless noteworthy, initiative is the UK Genetic Links to Anxiety and Depression (GLAD) Study (Genetic Links to Anxiety and Depression, 2019), which is trying to recruit 40,000 people with a history of anxiety or depression to create a pool of patients for a range of future studies that would combine genetic data with administrative claims data and self-report data collected via web surveys. These initiatives hold considerable promise for advancing research on precision treatment of mental disorders (McIntosh et al., 2016). It will doubtlessly be possible to develop PTRs for biomarker prescriptive predictors from these databases using balanced observational subsamples constructed with the methods described earlier in this paper.

However, even these massive omnibus research initiatives have limits and will not meet the needs of all researchers seeking to use novel measures of hypothesized prescriptive predictors to develop PTRs. A practical way of addressing these gaps when expensive measures are needed in very large samples might be to develop an iterative process of beginning with inexpensive archival data that are analyzed with the goal of targeting a smaller subsample of patients for direct assessment with the expensive measures. A multi-

Author Manuscript

tiered approach might be used instead. A good example of such an approach can be found in research designed to screen psychiatric inpatients and outpatients thought to be at elevated suicide risk to predict future SRB. Clinical practice guidelines call for such patients to receive in-depth clinical evaluations of suicide risk (Bernert, Horn, & Roberts, 2014; Silverman et al., 2015). However, these evaluations are very time-consuming (Rudd, 2014) and are likely to be useful for only a small proportion of patients. Perhaps because of this fact, in-depth clinical evaluations are carried out with only about half the patients even in settings where practice guidelines call for this always to be done (Cooper et al., 2013). It is unclear how clinicians decide which patients should or should not receive these in-depth evaluations, but a principled way to make this decision and to help improve treatment planning based on the evaluations would be to use a three-tiered approach of the following sort based on a series of ML models:

i. In the first tier, passively-collected EMR data, including data extracted from text analyses of clinical notes using NLP, could be used to develop a first-stage ML prediction model to determine which patients have such low predicted suicide risk that they could be excluded from further assessments. There is reason to believe based on the results of previous studies reviewed elsewhere (Kessler et al., 2019) that such models would find a substantial proportion of patients with such low SRB risk that they could reasonably be spared the burden of being subjected to more active suicide risk assessments;

ii. In a second tier, structured self-report suicide risk scales could be administered to the remaining patients for purposes of developing a second-tier ML prediction model designed to determine which of the remaining patients have a sufficiently high SRB risk to be considered for one of the special intensive types of psychotherapy (Jobes et al., 2017; Linehan et al., 2015; Rudd, 2012; Tighe, Nicholas, Shand, & Christensen, 2018) or combination therapies (Chesin et al., 2018; Forkmann, Brakemeier, Teismann, Schramm, & Michalak, 2016) known to have significant aggregate effects in reducing SRB. Numerous scales of this sort exist (Greist, Mundt, Gwaltney, Jefferson, & Posner, 2014; Quinlivan et al., 2016; Runeson et al., 2017). In addition, several self-report assessment tools have recently been developed that predict future SRB even among patients who deny suicidality. Included here are performance-based neurocognitive tests (Nock et al., 2010) and assessments based on linguistic and acoustic features extracted using natural language processing methods from digitally-recorded verbal responses to open-ended questions (Pestian et al., 2017).

It is plausible to think that a battery of such measures, when combined with first-tier predictors, could help clinicians distinguish between patients with a comparatively high predicted SRB risk and patients with lower predicted risk. The former patients could then receive an in-depth clinical evaluation. It is noteworthy that a composite risk prediction score could be used in developing this score based on coefficients generated in the first-tier analysis. This would allow the added strength of prediction from the larger first-tier sample to be imported into the second-tier analysis. This general approach of using information from prior tiers with much larger samples to reduce costs and improve the accuracy of

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

estimation in later tiers is known as *case-cohort analysis*. Case-cohort designs are reviewed by Noma and Tanaka (Noma & Tanaka, 2017).

It should be noted that critics have argued against using structured self-report suicide prediction scales as a basis for making clinical decisions about treating suicidal patients because of the low positive predictive values of these scales (Carter et al., 2017; Large, Ryan, Carter, & Kapur, 2017; Mulder, Newton-Howes, & Coid, 2016). However, it is important to recognize that this criticism does not speak to the likely value of ML models based on such scales helping to determine which patients should receive in-depth evaluations given that many patients currently do not receive such evaluations.

(iii) In the third tier assessment, in-depth clinical suicide risk evaluations could be carried out with patients predicted to be at high SRB risk to determine which of these patients would profit from special treatments designed to reduce SRB. As noted above, practice guidelines currently call for clinicians to make such decisions based on clinical judgment. However, previous research suggests that such judgments are often suboptimal (Nock et al., 2010; Woodford et al., 2017). It is very likely that data-driven PTRs could lead to improvements in these decisions. This should be done ideally using a case-cohort analysis approach to improve the precision of third-tier estimates.

9. Using machine learning methods to develop precision treatment rules

In order to develop a PTR, a prescriptive prediction model needs to be built using baseline variables that are either available for all patients or available in a tiered form of the sort described above. Once developed, such a model can be used to generate individual-level predicted treatment outcome scores for each patient separately for each treatment option under consideration. These scores can then be analyzed using external information about the relative costs of the different treatments to arrive at PTRs. The first part of this section describes our preferred approach to estimating prescriptive prediction models. The second part describes the way in which external information can be used to determine PTRs based on individual-level predictions produced by prescriptive prediction models.

10. Estimating prescriptive prediction models

We noted above in the Sample size requirement section that the conventional approach to estimating a prescriptive prediction model is to include hypothesized prescriptive predictors, dummy variables for treatment types, and interactions between the two classes of variables as predictors of treatment outcomes in a multiple regression analysis. When multiple interactions are statistically significant, predicted outcome scores for each patient are estimated under each treatment regimen based on model coefficients, and these treatment-specific predicted outcome scores are compared at the individual level to select the regimen with the best predicted outcome (VanderWeele, Luedtke, van der Laan, & Kessler, 2019). Cross-validation (CV) is needed to avoid the problem of over-fitting (Abadie, Chingos, & West, 2018). The accuracy of this approach requires correct specification of both the (possibly nonlinear) main effects and the (possibly complex nonlinear and higher-order) interactions.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Other modeling approaches exist that estimate interactions directly and do not require correct specification of the main effects although they do require correct specification of the interaction terms (Murphy, 2003; Robins, 2004). The approach we prefer is one of the latter approaches but has an important advantage over others that focus directly on interactions in that it improves on the ability to specify interactions correctly. It does this by using an ensemble ML approach rather than relying on a single algorithm to estimate the interactions. That is, our preferred approach estimates the interactions a number of different times, each time using a different modeling approach (e.g., conventional logistic regression, penalized regression, random forests, support vector machines, neural networks, etc.), and then uses a weighting procedure to arrive at a final estimate that averages individual-level estimates across the different approaches. This is done as a special case of the super learner (SL) algorithm (van der Laan, Polley, & Hubbard, 2007), an ensemble ML approach that uses CV to select a weighted combination of predicted outcome scores across a collection of candidate algorithms that yields an optimal weighted combination guaranteed to perform as least as well as the best component algorithm according to a pre-specified criterion.

It is important to recognize that the use of an ensemble of both parametric algorithms (which could include, but would not need to be limited to, the conventional linear interaction model described earlier in the paper) and flexible ML algorithms to estimate these interactions makes SL less prone than approaches based on a single algorithm to misspecification of the interactions (van der Laan & Luedtke, 2015). Furthermore, the guarantee that SL performs at least as well as the best candidate algorithm in expectation allows a rich library of parametric and flexible candidate algorithms to be included, with typical ensembles including such algorithms as conventional regression, various types of penalized regression, spline regression, adaptive polynomial splines, decision trees, Bayesian additive regression trees, support vector machines, and neural networks (LeDell, van der Laan, & Petersen, 2016).

Although it is beyond the scope of this paper to present a formal description of SL, an intuitive understanding is easy to grasp:

i. The approach begins by estimating a SL model to predict the treatment outcome separately among patients in each treatment arm under the assumption of randomization of patients to treatments, where this assumption could be based either on actual randomization in a randomized clinical trial, one of the balancing methods described above in an observational sample, or on balancing applied to a randomized clinical trial to adjust for imperfect randomization.

ii. The coefficients in the resulting treatment-specific ensemble models are then used to generate a predicted treatment outcome score under each treatment alternative for each patient in the study regardless of the type of treatment received.

iii. These predicted outcome scores are then used to create individual-level predicted difference scores for each logically possible pair of treatment alternatives, for each alternative compared to an average outcome across all treatment alternatives, or, in an observational study, for each alternative compared to a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

weighted average outcome across the observed distribution of treatment alternatives.

iv. Although these difference scores could be examined directly and interpreted as interactions, a better approach is to estimate a second set of SL models to predict a weighted version of these difference scores (for a discussion of the rationale for the weighting, see Luedtke & van der Laan, 2017) as a way of directly estimates the predictors of interactions (i.e., individual differences in treatment effects) in a way that avoids the need to estimate main effects. This is important step because the main effects might be more difficult to estimate due to the complexity of the underlying biological and psychosocial factors influencing overall treatment response, in which case individual-level difference scores would be estimated more accurately using an approach that does not require main effects to be estimated.

Selection of the predicted optimal treatment for a specific patient boils down to selecting the treatment with the best predicted difference score (either a higher probability of remission or a more desirable continuous outcome score) across the range of alternatives in this second set of SL models. The relative importance of specific predictors in determining these individual-level predicted differences can be examined by using the individual-level predicted differences scores from these models as outcomes in exploratory ML analyses where either penalized regression or regression trees are used to estimate which prescriptive predictors in the models contributed most to the predicted difference scores. Some ML algorithms (e.g. random forests) generate a graph of the relative importance of all such predictors.

It is noteworthy that absolute predicted values (i.e., main effects, including intercepts) are irrelevant to the decision about optimal treatments so long as all the treatment options under consideration are included in the analysis. This is true because individual-level differences in expected outcomes across this complete set of options provide sufficient information for choosing an optimal treatment for each patient. However, main effects become relevant when the option of receiving no treatment is also includes as an alternative, in which case SL PTR models can be expanded to estimate an absolute outcome score (i.e., an individual-level predicted probability of remission or a predicted continuous outcome score), either under one particular type of treatment or averaged across a range of treatment alternatives. As noted above, though, it is important to recognize that these absolute value estimates are likely to be more biased than the estimates of difference scores.

It is also noteworthy that SL is not unique in being able to estimate a PTR that is always approximately optimal for a given set of prescriptive predictors within the (possibly mis-specified) class of functional forms considered. This same property holds for some specific candidate algorithms (Rubin & van der Laan, 2012; Zhang, Tsiatis, Davidian, Zhang, & Laber, 2012; Zhao, Zeng, Rush, & Kosorok, 2012). However, as noted above, chances of estimating the optimal treatment strategy correctly is higher in SL because of the wider range of functional forms it allows to be considered. In cases where the conditional average treatment effects for each treatment condition can be correctly specified with a single algorithm, an approach that begins by using only that algorithm can have better finite-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

sample performance than SL (Luedtke & Chambaz, 2017) even though SL will, in expectation, find the same solution so long as the correct algorithm is included in the SL library. But this disadvantage of SL is more theoretical than real because there can never be a guarantee that the conditional average treatment effects for each treatment condition are correctly specified with a particular algorithm and misspecification would lead to biased estimation of the PTR with no advantage other than an increased efficiency that could be addressed by increasing sample size.

11. Estimating the aggregate impact of using a precision treatment rule

As noted in the previous section, the expected *individual-level effect* of optimal treatment for a specific patient can be conveyed to clinicians by reporting the extent to which the outcome for that patient would be expected to improve under optimal treatment compared to some alternative specified by the researcher. The *aggregate effect* of optimal treatment, in comparison, can be conveyed to policy-makers by using a cross-validated targeted minimum loss-based estimator (CV-TMLE) (van der Laan & Luedtke, 2015) of the attained improvement in the outcome based on optimized treatment compared to the same alternative as in the individual-level analysis. Depending on the policy purposes of the researcher and treatment system considering the PTR, the analysis might focus on a PTR that selected the treatment option with the better predicted outcome for each patient or, in the case of constrained treatment resources, for some proportion of patients compared to the aggregate outcome under some other treatment allocation scheme. The latter could be balanced randomization (i.e., when 50% of patients are randomly assigned to each of two treatment conditions) or, in the case of an observational study, observed allocation across treatment conditions. These CV-TMLEs yield estimators of the attained improvement with minimal bias because they use CV to separate the estimation of the optimal treatment strategy from the assessment of the estimated strategy's performance and allow also for the incorporation of flexible estimation approaches for the regressions and conditional probabilities needed to define the attained improvement.

If, as is sometimes the case, one treatment option is preferable for some patients and another treatment option for other patients, aggregate effects of optimization can be evaluated by comparing predicted outcomes based on optimal assignment vs. random assignment. However, when treatment options are incremental and the more intensive treatment is the preferred one for all patients, it is sometimes useful to evaluate aggregate treatment outcomes under optimization constraints. For example, we could estimate whether significantly more than 25% of the aggregate increase in remission achieved by providing a globally optimal but expensive treatment to all patients could be achieved by providing that treatment only to the 25% of patients predicted to benefit most from that treatment. The distribution of CV predicted individual-level difference scores could be inspected in such a case to determine if a substantial difference existed across patients in the relative benefit of an optimal treatment. A series of incremental decision margins of this sort can be evaluated to determine the extent to which aggregate estimates of predicted benefits deviate from the values expected in the absence of individual-differences in the value of the optimal treatment.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

As suggested above, extensions can easily be made to situations where more than two treatment options exist. An interesting area of investigation in such situations is the separate evaluation of the proportional distribution of optimality across treatment types and the calculation of the extent to which aggregate optimal treatment effects are due to treatments that might or might not be those that are most often optimal. For example, in ongoing collaborative work we are carrying out to evaluate the comparative effectiveness of first-line medications for schizophrenia in a large registry sample with an observational design, we are comparing 15 different medications. Results suggest that a substantial improvement in successful 12-month outcomes (defined by absence of hospitalization) could be achieved over current practice by using optimal treatment allocation across these medications. A small number of medications are optimal for most patients. We decomposed the aggregate expected improvement in outcome due to treatment optimization by type of optimal treatment by summing the product of the proportion of patients for which a given treatment is optimal by the mean expected improvement in outcome associated with that treatment. As it happens, some large contributors to aggregate expected improvement were treatments that were optimal for only small proportions of patients but made large differences to the outcomes of these patients compared to second-best treatments.

When treatment options are not all equivalent in terms of costs and secondary benefits, creation of PTRs for each patient requires the additional step of considering the comparative costs and secondary benefits of each treatment option (Steyerberg et al., 2010). A treatment that would be considered optimal for a specific patient based only on considerations of comparative treatment effectiveness might not be optimal when taking into consideration the small expected improvement in clinical outcome and high cost compared to those of next-best treatments. In a situation of this sort, the clinician needs to use such additional information to decide on the comparative net benefits of treatments in making treatment decisions. A discussion of statistical issues in the evaluation of net benefit in deciding on clinical decision thresholds is presented by Vickers and colleagues (Vickers, Van Calster, & Steyerberg, 2016).

12. Discussion

By allowing the researcher to use all algorithms under consideration when developing PTRs rather than relying on a single algorithm, a SL ensemble can resolve uncertainties about the form of the optimal PTR. Importantly, if it turns out that a simple interaction model (or any other single algorithm) is best, the SL optimality guarantee means that the PTR based on SL will be at least as good as the PTR based on that single algorithm in expectation. Furthermore, SL diagnostics will allow the researcher to detect a situation in which a single algorithm dominates the others in the ensemble, in which case efficiency can be improved by repeating the analysis using only that single algorithm. Given these desirable features of SL, the main challenge in estimating PTRs is that much larger samples are needed than those available in existing randomized clinical trials for mental disorders. The most feasible way of dealing with this problem is to work with large observational databases. As we discussed above, the key challenge in doing this is that observational databases often lack measures of the key prescriptive predictors of interest to the researchers and of the key baseline covariates that are needed to balance samples prior to carrying out causal analyses. We

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

illustrated the ways we are addressing this challenge in ongoing studies designed to develop PTRs for mental disorders from observational data.

As the adoption of measurement-based treatment protocols for common mental disorders becomes more and more common, availability of both prescriptive predictors and mental disorder treatment outcome measures will increase (Peterson, Anderson, & Bourne, 2018; Waldrop & McGuinness, 2017). This will make it much easier than it is currently to analyze large observational databases to develop PTRs for mental disorders. We suspect that patient report and administrative variables are likely to be the key components of these models unless either breakthroughs occur in biomarker assessment methods or, more realistically, new biological treatments come into existence for which useful PTRs can be developed using existing measures. It is likely that biomarkers, which have been a central focus of research on precision treatment of mental disorders up to now (Menke, 2018), will continue to be expensive to collect, although costs for many biomarkers are dropping rapidly, but tiered analyses of observational datasets along the lines described above will increase the efficiency of biomarker studies even if costs remain high by focusing biomarker collection on the segments of the patient population where this information is most likely to be of value. Only time will tell the extent to which the development and use of PTRs based on these developments will be able to improve patient outcomes, but it is almost certainly the case that such improvements will be substantial.

Acknowledgments

Funding

Jose R. Zubizarreta was supported by a grant from the Alfred P. Sloan Foundation (G-2018-10118).

Role of the funding source

The funding source had no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit this article for publication.

References

Abadie A, Chingos MM, & West MR (2018). Endogenous stratification in randomized experiments. The Review of Economics and Statistics, C(4), 567–580.

American Community Survey (ACS) (2019). https://www.census.gov/programs-surveys/acs, Accessed date: 15 May 2019.

Anestis MD, & Houtsma C (2018). The association between gun ownership and statewide overall suicide rates. Suicide and Life-Threatening Behavior, 48(2), 204–217. 10.1111/sltb.12346. [PubMed: 28294383]

Anglemyer A, Horvath HT, & Bero L (2014). Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database of Systematic Reviews, 4, MR000034 10.1002/14651858.MR000034.pub2.

Angrist JD, Imbens GW, & Rubin DB (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91(434), 444–455.

Baiocchi M, Cheng J, & Small DS (2014). Instrumental variable methods for causal inference. Statistics in Medicine, 33(13), 2297–2340. 10.1002/sim.6128. [PubMed: 24599889]

Barbui C, Esposito E, & Cipriani A (2009). Selective serotonin reuptake inhibitors and risk of suicide: A systematic review of observational studies. Canadian Medical Association Journal 180(3), 291–297. 10.1503/cmaj.081514. [PubMed: 19188627]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Berger ML, Sox H, Willke RJ, Brixner DL, Eichler HG, Goettsch W, et al. (2017). Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE special task force on real-world evidence in health care decision making. Pharmacoepidemiology and Drug Safety, 26(9), 1033–1039. 10.1002/pds.4297. [PubMed: 28913966]

Bernert RA, Hom MA, & Roberts LW (2014). A review of multidisciplinary clinical practice guidelines in suicide prevention: Toward an emerging standard in suicide risk assessment and management, training and practice. Academic Psychiatry, 38(5), 585–592. 10.1007/s40596-014-0180-1. [PubMed: 25142247]

Bousman CA, Arandjelovic K, Mancuso SG, Eyre HA, & Dunlop BW (2019). Pharmacogenetic tests and depressive symptom remission: A meta-analysis of randomized controlled trials. Pharmacogenomics, 20(1), 37–47. 10.2217/pgs-2018-0142. [PubMed: 30520364]

Bryan CJ, Peterson AL, & Rudd MD (2018). Differential effects of brief CBT versus treatment as usual on posttreatment suicide attempts among groups of suicidal patients. Psychiatric Services, 69(6), 703–709. 10.1176/appi.ps.201700452. [PubMed: 29493409]

Bzdok D, & Meyer-Lindenberg A (2018). Machine learning for precision psychiatry: Opportunities and challenges. Biological Psychiatry. Cognitive Neuroscience and Neuroimaging, 3(3), 223–230. https://doi.Org/10.1016/j.bpsc.2017.11.007. [PubMed: 29486863]

Carter G, Milner A, McGill K, Pirkis J, Kapur N, & Spittal MJ (2017). Predicting suicidal behaviours using clinical instruments: Systematic review and meta-analysis of positive predictive values for risk scales. The British Journal of Psychiatry: Journal of Mental Science, 210(6), 387–395. 10.1192/bjp.bp.116.182717.

Cattaneo MD, Frandsen BR, & Titiunik R (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. Journal of Causal Inference, 3(1), 1–24. 10.1515/jci-2013-0010.

Chan KC, Yam SC, & Zhang Z (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. Journal of the Royal Statistical Society: Series B, 78(3), 673–700. 10.1111/rssb.12129.

Chesin MS, Brodsky BS, Beeler B, Benjamin-Phillips CA, Taghavi I, & Stanley B (2018). Perceptions of adjunctive mindfulness-based cognitive therapy to prevent suicidal behavior among high suicide-risk outpatient participants. Crisis, 39(6), 451–460. 10.1027/0227-5910/a000519. [PubMed: 29848083]

Chung DT, Ryan CJ, Hadzi-Pavlovic D, Singh SP, Stanton C, & Large MM (2017). Suicide rates after discharge from psychiatric facilities: A systematic review and meta-analysis. JAMA Psychiatry, 74(7), 694–702. 10.1001/jamapsychiatry.2017.1044. [PubMed: 28564699]

Cole SR, & Hernán MA (2008). Constructing inverse probability weights for marginal structural models. American Journal of Epidemiology, 168(6), 656–664. [PubMed: 18682488]

Cohen ZD, & DeRubeis RJ (2018). Treatment selection in depression. Annual Review of Clinical Psychology, 14, 209–236. 10.1146/annurev-clinpsy-050817-084746.

Cooper J, Steeg S, Bennewith O, Lowe M, Gunnell D, House A, et al. (2013). Are hospital services for self-harm getting better? An observational study examining management, service provision and temporal trends in england. BMJ Open, 3(11), e003444 10.1136/bmjopen-2013-003444.

Costello EJ, Erkanli A, Copeland W, & Angold A (2010). Association of family income supplements in adolescence with development of psychiatric and substance use disorders in adulthood among an American Indian population. Journal of the American Medical Association, 303(19), 1954–1960. 10.1001/jama.2010.621. [PubMed: 20483972]

Dahabreh IJ, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, … Kitsios GD (2012). Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. European Heart Journal 33(15), 1893–1901. 10.1093/eurheartj/ehs114. [PubMed: 22711757]

DeRubeis RJ, Cohen ZD, Forand NR, Fournier JC, Gelfand LA, & Lorenzo-Luaces L (2014). The personalized advantage index: Translating research on prediction into individualized treatment recommendations. A demonstration. PLoS One, 9(1), e83875 10.1371/journal.pone.0083875. [PubMed: 24416178]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Diamond A, & Sekhon JS (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. The Review of Economics and Statistics, 95(3), 932–945. 10.1162/REST_a_00318.

Dorresteijn JA, Visseren FL, Ridker PM, Wassink AM, Paynter NP, Steyerberg EW, … Cook NR (2011). Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ, 343, d5888 10.1136/bmj.d5888. [PubMed: 21968126]

Durand CP (2013). Does raising type 1 error rate improve power to detect interactions in linear regression models? A simulation study. PLoS One, 8(8), e71079 10.1371/journal.pone.0071079. [PubMed: 23976980]

Fabbri C, Crisafulli C, Calabro M, Spina E, & Serretti A (2016). Progress and prospects in pharmacogenetics of antidepressant drugs. Expert Opinion on Drug Metabolism and Toxicology, 12(10), 1157–1168. 10.1080/17425255.2016.1202237. [PubMed: 27310483]

Forkmann T, Brakemeier EL, Teismann T, Schramm E, & Michalak J (2016). The effects of mindfulness-based cognitive therapy and cognitive behavioral analysis system of psychotherapy added to treatment as usual on suicidal ideation in chronic depression: Results of a randomized-clinical trial. Journal of Affective Disorders, 200, 51–57. 10.1016/j.jad.2016.01.047. [PubMed: 27128357]

Genetic Links to Anxiety & Depression (GLAD) (2019). https://gladstudy.org.uk/, Accessed date: 15 May 2019.

Greenland S (1983). Tests for interaction in epidemiologic studies: A review and a study of power. Statistics in Medicine, 2(2), 243–251. [PubMed: 6359318]

Greist JH, Mundt JC, Gwaltney CJ, Jefferson JW, & Posner K (2014). Predictive value of baseline electronic Columbia-Suicide Severity Rating Scale (eC-SSRS) assessments for identifying risk of prospective reports of suicidal behavior during research participation. Innovations in Clinical Neuroscience, 11(9–10), 23–31.

Gruber S, Logan RW, Jarrin I, Monge S, & Hernan MA (2015). Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. Statistics in Medicine, 34(1), 106–117. 10.1002/sim.6322. [PubMed: 25316152]

Hahn J, Todd P, & Van der Klaauw W (2001). Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica, 69(1), 201–209. 10.1111/1468-0262.00183.

Hainmueller J (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis, 20(1), 25–46. 10.1093/pan/mpr025.

Handley MA, Lyles CR, McCulloch C, & Cattamanchi A (2018). Selecting and improving quasi-experimental designs in effectiveness and implementation research. Annual Review of Public Health, 39, 5–25. 10.1146/annurev-publhealth-040617-014128.

Hansen BB (2004). Full matching in an observational study of coaching for the SAT. Journal of the American Statistical Association, 99(467), 609–618. 10.1198/016214504000000647.

Harned MS, Chapman AL, Dexter-Mazza ET, Murray A, Comtois KA, & Linehan MM (2008). Treating co-occurring Axis I disorders in recurrently suicidal women with borderline personality disorder: A 2-year randomized trial of dialectical behavior therapy versus community treatment by experts. Journal of Consulting and Clinical Psychology, 76(6), 1068–1075. 10.1037/a0014044. [PubMed: 19045974]

Hirshberg DA, Maleki A, & Zubizarreta J (2019). Minimax linear estimation of the retargeted mean. arXiv.org, arXiv:190110296.

Hirshberg DA, & Zubizarreta JR (2017). On two approaches to weighting in causal inference. Epidemiology, 28(6), 812–816. 10.1097/EDE.0000000000000735. [PubMed: 28817467]

Hjorthoj CR, Madsen T, Agerbo E, & Nordentoft M (2014). Risk of suicide according to level of psychiatric treatment: A nationwide nested case-control study. Social Psychiatry and Psychiatric Epidemiology, 49(9), 1357–1365. 10.1007/s00127-014-0860-x. [PubMed: 24647741]

Hoffmire C, Stephens B, Morley S, Thompson C, Kemp J, & Bossarte RM (2016). VA Suicide Prevention Applications Network: A national health care system–based suicide event tracking system. Public Health Reports, 131(6), 816–821. 10.1177/0033354916670133. [PubMed: 28123228]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Huh D, Jobes DA, Comtois KA, Kerbrat AH, Chalker SA, Gutierrez PM, et al. (2018). The collaborative assessment and management of suicidality (CAMS) versus enhanced care as usual (E-CAU) with suicidal soldiers: Moderator analyses from a randomized controlled trial. Military Psychology, 30(6), 459–506. 10.1080/08995605.2018.1503001.

Imai K, & Ratkovic M (2014). Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B, 76(1), 243–263. 10.1111/rssb.12027.

Jobes DA, Comtois KA, Gutierrez PM, Brenner LA, Huh D, Chalker SA, … Crow B (2017). A randomized controlled trial of the collaborative assessment and management of suicidality versus enhanced care as usual with suicidal soldiers. Psychiatry, 80(4), 339–356. 10.1080/00332747.2017.1354607. [PubMed: 29466107]

Judd CM, Westfall J, & Kenny DA (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. Annual Review of Psychology, 68, 601–625. 10.1146/annurev-psych-122414-033702.

Kang JDY, & Schafer JL (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science, 22(4), 523–539.

Kapur N, Ibrahim S, While D, Baird A, Rodway C, Hunt IM, et al. (2016). Mental health service changes, organisational factors, and patient suicide in England in 1997-2012: A before-and-after study. Lancet Psychiatry, 3(6), 526–534. 10.1016/S2215-0366(16)00063-8. [PubMed: 27107805]

Keele L, Titiunik R, & Zubizarreta JR (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. Journal of the Royal Statistical Society: Series A, 178(1), 223–239.

Kessler RC (2018). The potential of predictive analytics to provide clinical decision support in depression treatment planning. Current Opinion in Psychiatry, 31(1), 32–39. 10.1097/YCO.0000000000000377. [PubMed: 29076894]

Kessler RC, Bernecker SL, Bossarte RM, Luedtke AR, McCarthy JF,, & Nock MK (2019). The role of big data analytics in predicting suicide In Passos IC, Mwangi B, & Kapaczinski F (Eds.). Personalized and predictive psychiatry - big data analytics in mental health (pp. 77–98). Switzerland: Springer Nature.

Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, Ebert DD, … Zaslavsky AM (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. Epidemiology and Psychiatric Sciences, 26(1), 22–36. 10.1017/S2045796016000020. [PubMed: 26810628]

Kinard EM, & Klerman LV (1980). Changes in life style following mental hospitalization. The Journal of Nervous and Mental Disease, 268(11), 666–672.

Kind AJH, & Buckingham WR (2018). Making neighborhood-disadvantage metrics accessible - the neighborhood atlas. New England Journal of Medicine, 378(26), 2456–2458. 10.1056/NEJMp1802313. [PubMed: 29949490]

Kovalchik SA, Varadhan R, & Weiss CO (2013). Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. Statistics in Medicine, 32(28), 4906–4923. 10.1002/sim.5881. [PubMed: 23788362]

van der Laan MJ, & Luedtke AR (2015). Targeted learning of the mean outcome under an optimal dynamic treatment rule. Journal of Causal Inference, 3(1), 61–95. 10.1515/jci-2013-0022. [PubMed: 26236571]

van der Laan MJ, Polley EC, & Hubbard AE (2007). Super learner. Statistical Applications in Genetics and Molecular Biology, 6(Article25)10.2202/1544-6115.1309.

Large MM, & Kapur N (2018). Psychiatric hospitalisation and the risk of suicide. The British Journal of Psychiatry: Journal of Mental Science, 212(5), 269–273. 10.1192/bjp.2018.22.

Large MM, Ryan CJ, Carter G, & Kapur N (2017). Can we usefully stratify patients according to suicide risk? BMJ, 359, j4627 10.1136/bmj.j4627. [PubMed: 29042363]

LeDell E, van der Laan MJ, & Petersen M (2016). AUC-maximizing ensembles through metalearning. International Journal of Biostatistics, 12(1), 203–218. https://doi.org.10.1515/ijb-2015-0035. [PubMed: 27227721]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

LexisNexis (2019). Solutions for professionals who shape the world https://www.lexisnexis.com/en-us/gateway.page, Accessed date: 2 January 2019.

Linehan MM, Korslund KE, Harned MS, Gallop RJ, Lungu A, Neacsiu AD, … Murray-Gregory AM (2015). Dialectical behavior therapy for high suicide risk in individuals with borderline personality disorder: A randomized clinical trial and component analysis. JAMA Psychiatry, 72(5), 475–482. 10.1001/jamapsychiatry.2014.3039. [PubMed: 25806661]

Lin JJ, & Lu TH (2011). Trends in solids/liquids poisoning suicide rates in Taiwan: A test of the substitution hypothesis. BMC Public Health, 11, 712. [PubMed: 21933432]

Lipkovich I, Dmitrienko A, D'Agostino RB, & DAgostiino RB Sr. (2017). Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. Statistics in Medicine, 36(1), 136–196. 10.1002/sim.7064. [PubMed: 27488683]

Lipschitz JM, Benzer JK, Miller C, Easley SR, Leyson J, Post EP, et al. (2017). Understanding collaborative care implementation in the Department of Veterans Affairs: Core functions and implementation challenges. BMC Health Services Research, 17(1), 691 10.1186/s12913-017-2601-9. [PubMed: 29017488]

Lubin G, Werbeloff N, Halperin D, Shmushkevitch M, Weiser M, & Knobler HY (2010). Decrease in suicide rates after a change of policy reducing access to firearms in adolescents: A naturalistic epidemiological study. Suicide and Life-Threatening Behavior, 40(5), 10.1521/suli.2010.40.5.421 424–424.

Luedtke A, & Chambaz A (2017). Faster rates for policy learning. arXiv.org, arXiv:1704.06431.

Luedtke A, Sadikova E, & Kessler RC (2019). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. Clinical Psychological Science, 7(3)10.1177/2167702618815466.

Luedtke AR, & van der Laan MJ (2017). Evaluating the impact of treating the optimal subgroup. Statistical Methods in Medical Research, 26(4), 1630–1640. 10.1177/0962280217708664. [PubMed: 28482779]

McClellan M, McNeil BJ, & Newhouse JP (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. Journal of the American Medical Association, 272(11), 859–866. [PubMed: 8078163]

McCoy TH Jr., Castro VM, Roberson AM, Snapper LA, & Perlis RH (2016). Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. JAMA Psychiatry, 73(10), 1064–1071. 10.1001/jamapsychiatry.2016.2172. [PubMed: 27626235]

McCoy TH Jr., Pellegrini AM, & Perlis RH (2019). Research Domain Criteria scores estimated through natural language processing are associated with risk for suicide and accidental death. Depression and Anxiety, 36(5), 392–399. 10.1002/da.22882. [PubMed: 30710497]

McIntosh AM, Stewart R, John A, Smith DJ, Davis K, Sudlow C, … Group, M. Q. D. S. (2016). Data science for mental health: A UK perspective on a global challenge. Lancet Psychiatry, 3(10), 993–998. 10.1016/S2215-0366(16)30089-X. [PubMed: 27692269]

McMain SF, Fitzpatrick S, Boritz T, Barnhart R, Links P, & Streiner DL (2018). Outcome trajectories and prognostic factors for suicide and self-harm behaviors in patients with borderline personality disorder following one year of outpatient psychotherapy. Journal of Personality Disorders, 32(4), 497–512. 10.1521/pedi_2017_31_309. [PubMed: 28910214]

Menke A (2018). Precision pharmacotherapy: Psychiatry's future direction in preventing, diagnosing, and treating mental disorders. Pharmgenomics and Personalized Medicine, 11, 211–222. 10.2147/PGPM.S146110.

Moscoe E, Bor J, & Bärnighausen T (2015). Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: A review of current and best practice. Journal of Clinical Epidemiology, 68(2), 122–133. 10.1016/j.jclinepi.2014.06.021. [PubMed: 25579639]

Mulder R, Newton-Howes G, & Coid JW (2016). The futility of risk prediction in psychiatry. The British Journal of Psychiatry: Journal of Mental Science, 209(4), 271–272. 10.1192/bjp.bp.116.184960.

Murphy SA (2003). Optimal dynamic treatment regimes. Journal of the Royal Statistical Society - Series B: Statistical Methodology, 65(2), 331–355.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

National Institutes of Health (2019). All of us research program https://allofus.nih.gov/, Accessed date: 15 May 2019.

Ning J, Hong C, Li L, Huang X, & Shen Y (2017). Estimating treatment effects in observational studies with both prevalent and incident cohorts. Canadian Journal of Statistics, 45(2), 202–219. 10.1002/cjs.11317. [PubMed: 29056817]

Nock MK, Park JM, Finn CT, Deliberto TL, Dour HJ, & Banaji MR (2010). Measuring the suicidal mind: Implicit cognition predicts suicidal behavior. Psychological Science, 21(4), 511–517. 10.1177/0956797610364762. [PubMed: 20424092]

Noma H, & Tanaka S (2017). Analysis of case-cohort designs with binary outcomes: Improving efficiency using whole-cohort auxiliary information. Statistical Methods in Medical Research, 26(2), 691–706. 10.1177/0962280214556175. [PubMed: 25348675]

Nordt C, Warnke I, Seifritz E, & Kawohl W (2015). Modelling suicide and unemployment: A longitudinal analysis covering 63 countries, 2000-11. Lancet Psychiatry, 2(3), 239–245. 10.1016/S2215-0366(14)00118-7. [PubMed: 26359902]

Olbrich S, van Dinteren R, & Arns M (2015). Personalized medicine: Review and perspectives of promising baseline EEG biomarkers in major depressive disorder and attention deficit hyperactivity disorder. Neuropsychobiology, 72(3–4), 229–240. 10.1159/000437435. [PubMed: 26901357]

Ondra T, Dmitrienko A, Friede T, Graf A, Miller F, Stallard N, et al. (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. Journal of Biopharmaceutical Statistics, 26(1), 99–119. 10.1080/10543406.2015.1092034. [PubMed: 26378339]

Paksarian D, Mojtabai R, Kotov R, Cullen B, Nugent KL, & Bromet EJ (2014). Perceived trauma during hospitalization and treatment participation among individuals with psychotic disorders. Psychiatric Services, 65(2), 266–269. 10.1176/appi.ps.201200556. [PubMed: 24492906]

Penkower L, Bromet EJ, & Dew MA (1988). Husbands' layoff and wives' mental health. A prospective analysis. Archives of General Psychiatry, 45(11), 994–1000. [PubMed: 3178415]

Perel P, Edwards P, Wentz R, & Roberts I (2006). Systematic review of prognostic models in traumatic brain injury. BMC Medical Informatics and Decision Making, 6, 38 10.1186/1472-6947-6-38. [PubMed: 17105661]

Pestian JP, Sorter M, Connolly B, Bretonnel Cohen K, McCullumsmith C, & Gee JT … STM Research Group. (2017). A machine learning approach to identifying the thought markers of suicidal subjects: A prospective multicenter trial. Suicide and Life-Threatening Behavior, 47(1), 112–121. 10.1111/sltb.12312. [PubMed: 27813129]

Peterson K, Anderson J, & Bourne D (2018). Evidence brief: Use of patient reported outcome measures for measurement-based care in mental health shared decision-making VA evidence-based synthesis program reports. Washington, DC: Department of Veterans Affairs (US).

Petkova E, Ogden RT, Tarpey T, Ciarleglio A, Jiang B, Su Z, … Trivedi MH (2017). Statistical analysis plan for stage 1 EMBARC (establishing moderators and biosignatures of antidepressant response for clinical care) study. Contemporary Clinical Trials Communications, 6, 22–30. 10.1016/j-conctc.2017.02.007. [PubMed: 28670629]

Prendes-Alvarez S, & Nemeroff CB (2018). Personalized medicine: Prediction of disease vulnerability in mood disorders. Neuroscience Letters, 669, 10–13. 10.1016/j.neulet.2016.09.049. [PubMed: 27746310]

Prieto-Merino D, & Pocock SJ (2012). The science of risk models. European Journal of Preventive Cardiology, 19(2 Suppl), 7–13. 10.1177/2047487312448995. [PubMed: 22801064]

Quinlivan L, Cooper J, Davies L, Hawton K, Gunnell D, & Kapur N (2016). Which are the most useful scales for predicting repeat self-harm? A systematic review evaluating risk scales using measures of diagnostic accuracy. BMJ Open, 6(2), e009297 10.1136/bmjopen-2015-009297.

Robins JM (2004). Optimal structural nested models for optimal sequential decisions In Lin DY, & Heagerty PJ (Eds.). Proceedings of the second Seattle symposium in biostatistics: Analysis of correlated data (pp. 189–326). New York: Springer.

Robins JM, & Rotnitzky A (1995). Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association, 90(429), 122–129.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Rosellini AJ, Street AE, Ursano RJ, Chiu WT, Heeringa SG, Monahan J, … Kessler RC (2017). Sexual assault victimization and mental health treatment, suicide attempts, and career outcomes among women in the us army. American Journal of Public Health, 107(5), 732–739. 10.2105/AJPH.2017.303693. [PubMed: 28323466]

Rosenbaum PR (2017). Observation and experiment: An introduction to causal inference. Cambridge, Massachusetts: Harvard University Press.

Rosenbaum PR, & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55.

Rubin DB, & van der Laan MJ (2012). Statistical issues and limitations in personalized medicine research with clinical trials. The International Journal of Biostatics, 8(1), 18 https://doi.org/l0.1515/1557-4679.1423.

Rudd DM (2012). Brief cognitive behavioral therapy (BCBT) for suicidality in military populations. Military Psychology, 24, 592–603.

Rudd MD (2014). Core competencies, warning signs, and a framework for suicide risk assessment in clinical practice In Nock MK (Ed.). The Oxford handbook of suicide and self-injury (pp. 323–336). (1st ed.). Cary: Oxford University Press.

Rumshisky A, Ghassemi M, Naumann T, Szolovits P, Castro VM, McCoy TH, et al. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Translational Psychiatry, 6(10), e921 10.1038/tp.2015.182. [PubMed: 27754482]

Runeson B, Odeberg J, Pettersson A, Edbom T, Jildevik Adamsson I, & Waern M (2017). Instruments for the assessment of suicide risk: A systematic review evaluating the certainty of the evidence. PLoS One, 12(7), e0180292 10.1371/journal.pone.0180292. [PubMed: 28723978]

Rupasingha A, Goetz SJ, & Freshwater D (2006). The production of social capital in US counties. The Journal of Socio-Economics, 35(1), 83–101. 10.1016/j.socec.2005.11.001.

Saltiel PF, & Silvershein DI (2015). Major depressive disorder: Mechanism-based prescribing for personalized medicine. Neuropsychiatric Disease and Treatment, 2015(11), 875–888. 10.2147/NDT.S73261.

Silverman JJ, Galanter M, Jackson-Triche M, Jacobs DG, Lomax JW 2nd, Riba MB, … American Psychiatric Association (2015). The American Psychiatric Association practice guidelines for the psychiatric evaluation of adults. American Journal of Psychiatry, 172(8), 798–802. 10.1176/appi.ajp.2015.1720501. [PubMed: 26234607]

Smith GC, Seaman SR, Wood AM, Royston P, & White IR (2014). Correcting for optimistic prediction in small data sets. American Journal of Epidemiology, 180(3), 318–324. https://doi.org/l0.1093/aje/kwul40. [PubMed: 24966219]

Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, … Kattan MW (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology, 21(1), 128–138. 10.1097/EDE.0b013e3181c30fb2. [PubMed: 20010215]

Svindseth MF, Dahl AA, & Hatling T (2007). Patients' experience of humiliation in the admission process to acute psychiatric wards. Nordic Journal of Psychiatry, 61(1), 47–53. 10.1080/08039480601129382. [PubMed: 17365789]

Swanson SA (2017). Instrumental variable analyses in pharmacoepidemiology: What target trials do we emulate? Current Epidemiology Reports, 4(4), 281–287. [PubMed: 29226066]

Tighe J, Nicholas J, Shand F, & Christensen H (2018). Efficacy of acceptance and commitment therapy in reducing suicidal ideation and deliberate self-harm: Systematic review. Journal of Medical Internet Research Mental Health, 5(2), e10732 10.2196/10732.

Trivedi MH, McGrath PJ, Fava M, Parsey RV, Kurian BT, Phillips ML, … Weissman MM (2016). Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design. Journal of Psychiatric Research, 78, 11–23. 10.1016/j.jpsychires.2016.03.001. [PubMed: 27038550]

Biobank UK (2018). https://www.ukbiobank.ac.uk/, Accessed date: 2 January 2019.

U.S. Department of Veterans Affairs (2013). VA/DoD clinicalpractice guidelines: Assessment and management of patients at risk for suicide, https://www.healthquality.va.gov/guidelines/MH/srb/, Accessed date: 2 January 2019.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

VA Office of Public and Intergovernmental Affairs (2017). VA REACH VET Initiative helps save veterans lives: Program signals when more help is needed for at-risk veterans, www.va.gov/opa/pressrel/pressrelease.cfm7id=2878, Accessed date: 2 January 2019.

VanderWeele TJ, Luedtke AR, van der Laan MJ, & Kessler RC (2019). Selecting optimal subgroups for treatment using many covariates. Epidemiology, 30(3), 334–341. 10.1097/EDE.0000000000000991. [PubMed: 30789432]

Venkataramani AS, Bor J, & Jena AB (2016). Regression discontinuity designs in healthcare research. BMJ, 352, i1216 10.1136/bmj.i1216. [PubMed: 26977086]

Vickers AJ, Van Calster B, & Steyerberg EW (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ, 352, 16 10.1136/bmj.16.

Visconti G, & Zubizarreta JR (2018). Handling limited overlap in observational studies with cardinality matching. Observational Studies, 4, 217–249.

Waldrop J, & McGuinness TM (2017). Measurement-based care in psychiatry. Journal of Psychosocial Nursing and Mental Health Services, 55(11), 30–35. 10.3928/02793695-20170818-01.

Wang Y, & Zubizarreta JR (2019). Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations. Biometrika in press.

While D, Bickley H, Roscoe A, Windfuhr K, Rahman S, Shaw J, et al. (2012). Implementation of mental health service recommendations in England and Wales and suicide rates, 1997-2006: A cross-sectional and before-and-after observational study. Lancet, 379(9820), 1005–1012. 10.1016/S0140-6736(11)61712-1. [PubMed: 22305767]

Wigmore EM, Hafferty JD, Hall LS, Howard DM, Clarke TK, Fabbri C, et al. (2019). Genome-wide association study of antidepressant treatment resistance in a population-based cohort using health service prescription data and meta-analysis with GENDEP. The Pharmacogenomics Journal, 10.1038/s41397-019-0067-3.

Woodford R, Spittal MJ, Milner A, McGill K, Kapur N, Pirkis J, et al. (2017). Accuracy of clinician predictions of future self-harm: A systematic review and meta-analysis of predictive studies. Suicide and Life-Threatening Behavior, 10.1111/sltb.12395.

Yiu S, & Su L (2018). Covariate association eliminating weights: A unified weighting framework for causal effect estimation. Biometrika, 105(3), 709–722. 10.1093/biomet/asy015. [PubMed: 31031408]

Zalsman G, Hawton K, Wasserman D, van Heeringen K, Arensman E, Sarchiapone M, et al. (2016). Suicide prevention strategies revisited: 10-year systematic review. Lancet Psychiatry, 3(7), 646–659. 10.1016/S2215-0366(16)30030-X. [PubMed: 27289303]

Zhang B, Tsiatis AA, Davidian M, Zhang M, & Laber E (2012). Estimating optimal treatment regimes from a classification perspective. Stat, 1(1), 103–114. 10.1002/sta.411. [PubMed: 23645940]

Zhang Y, Zhang OR, Li R, Flores A, Selek S, Zhang XY, et al. (2018). Psychiatric stressor recognition from clinical notes to reveal association with suicide. Health Informatics Journal 146045821879659810.1177/1460458218796598.

Zhao Y, Zeng D, Rush AJ, & Kosorok MR (2012). Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association, 107(449), 1106–1118. 10.1080/01621459.2012.695674. [PubMed: 23630406]

Zhou X, Mayer-Hamblett N, Khan U, & Kosorok MR (2017). Residual weighted learning for estimating individualized treatment rules. Journal of the American Statistical Association, 112(517), 169–187. 10.1080/01621459.2015.1093947. [PubMed: 28943682]

Zhu R, Zhao YQ, Chen G, Ma S& Zhao H (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. Biometrics, 73(2), 391–400. 10.1111/biom.12593. [PubMed: 27704531]

Zubizarreta JR (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. Journal of the American Statistical Association, 107(500), 1360–1371.

Zubizarreta JR (2015). Stable weights that balance covariates for estimation with incomplete outcome data. Journal of the American Statistical Association, 110(511), 910–922. 10.1080/01621459.2015.1023805.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Zubizarreta JR, Small DS, Goyal NK, Lorch S, & Rosenbaum PR (2013). Stronger instruments via integer programming in an observational study of late preterm birth outcomes. Annals of Applied Statistics, 7(1), 25–50.