Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments

Ariel Linden, DrPH[1,2], Paul R. Yarnold, PhD[3]

[1] President, Linden Consulting Group, LLC, Ann Arbor, Michigan, USA

[2] Research Scientist, Division of General Medicine, Medical School - University of Michigan, Ann Arbor, Michigan, USA

[3] President, Optimal Data Analysis, LLC, Evanston, Illinois, USA

Corresponding Author Information:
Ariel Linden, DrPH
Linden Consulting Group, LLC
1301 North Bay Drive
Ann Arbor, MI USA 48103
Phone: (971) 409-3505
Email: alinden@lindenconsulting.org

Key Words: multivalued treatments, machine learning, propensity score, inverse probability of treatment weighting, marginal mean weighting through stratification, doubly robust, observational studies, causal inference

Running Header: machine learning and multivalued treatments

Acknowledgement: We wish to thank Julia Adler-Milstein for her review and feedback on the manuscript.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/jep.12610

ABSTRACT

Rationale, aims and objectives: Interventions with multivalued treatments are common in medical and health research; examples include comparing the efficacy of competing interventions, and contrasting various doses of a drug. In recent years there has been growing interest in the development of methods that estimate multivalued treatment effects using observational data. This paper extends a previously described analytic framework for evaluating binary treatments to studies involving multivalued treatments utilizing a machine learning algorithm called optimal discriminant analysis (ODA).

Method: We describe the differences between regression-based treatment effect estimators and effects estimated using the ODA framework. We then present an empirical example using data from an intervention including three study groups to compare corresponding effects.

Results: The regression-based estimators produced statistically significant mean differences between the two intervention groups, and between one of the treatment groups and controls. In contrast, ODA was unable to discriminate between distributions of any of the three study groups.

Conclusions: ODA offers an appealing alternative to conventional regression-based models for estimating effects in multivalued treatment studies because of its insensitivity to skewed data and use of accuracy measures applicable to all prognostic analyses. If these analytic approaches produce consistent treatment effect *P* values, this bolsters confidence in the validity of the results. If the approaches produce conflicting treatment effect *P* values, as they do in our empirical example, the investigator should consider the ODA-derived estimates to be most

robust, given that ODA uses permutation *P* values that require no distributional assumptions and are thus, always valid.

INTRODUCTION

Interventions with multivalued treatments -- those including more than two discrete conditions (e.g., comparing the efficacy of competing drugs or interventions) or multiple levels of one treatment (e.g., various doses of a particular drug) -- are common in medical and health research. In experimental studies of multivalued treatments, outcomes may be analyzed by simply regressing the outcome on a set of indicator variables representing each treatment, followed by contrasts between treatments to estimate treatment effects. This analytic approach is sufficient to provide unbiased treatment effect estimates when subjects are randomized. However, when analyzing observational data, investigators estimate treatment effects by applying causal-inferential methods to control for threats to validity [1]. Selection bias is a particularly prominent threat to validity when evaluating health management programs, because individuals with high levels of health care utilization or costs are likely to be assigned to a particular treatment. Given their high outlier status at baseline, these individuals' outcomes naturally regress to the mean on their follow-up measurement, giving the false impression of a treatment effect [2,3].

In recent years, there has been a growing interest in the development of multivalued treatment effect estimators using observational data. The seminal work of Imbens [4] and Lechner [5] gave rise to this flourishing area by extending Rosenbaum and Rubin's [6] propensity score framework for binary treatments to multivalued treatments. Subsequently, several methods designed for binary treatments -- including regression, matching, weighting, and stratification -- have been reformulated to accommodate multivalued treatments (see for example, [7,8,9,10,11]).

Unlike randomized studies in which treatment groups are inherently comparable on both observed and unobserved pre-intervention characteristics, observational studies of multivalued treatments can only endeavor to generate treatment groups that are comparable on observed characteristics, and must assume that any unmeasured variables will not bias the results [12]. Thus, in evaluating a health management program with multivalued treatments, the investigator would ensure that all treatment groups were comparable on pre-intervention levels of health care utilization and cost, but must assume, for example, that unmeasured motivation to change health behaviors will not confound the outcomes [13,14]. Accordingly, an essential condition for assuming the validity of treatment effects in multivalued treatment studies is that all treatment groups are comparable on their observed pre-intervention characteristics [15,16].

Recently, a novel machine-learning approach was introduced for both assessing covariate balance on observed pre-intervention characteristics [17], and estimating treatment effects [18] in studies with binary treatments. This methodology employs an algorithm called optimal discriminant analysis (ODA) [19,20] to determine if, and to what degree, treatment groups can be distinguished based on the distributions of the covariates [17], and then subsequently on the outcomes [18].

In this paper we extend this machine-learning framework from the studies of binary treatments to those involving multivalued treatments. By framing the treatment-outcome relationship as a classification problem (i.e., how accurately does the outcome variable classify individuals as being in their actual treatment group), ODA offers several benefits over the

conventional statistical methods typically employed to assess both covariate balance and treatment effects in multivalued treatment studies. These include the ability to handle an outcome variable measured using any metric (from categorical to continuous), insensitivity to skewed data or outliers, and the use of accuracy measures that can be widely applied to all classification analyses. ODA also offers the unique ability to ascertain if individuals are likely to be responding to the treatment level as assigned (or self-selected) based on optimized (maximum-accuracy) cut-points on the outcome variable. Moreover, ODA accepts analytic weights, thereby allowing the evaluation of observational studies using any algorithm that produces weights for covariate adjustment [17,18]. Finally, ODA provides the capability to use cross-validation in assessing the generalizability of the model to individuals outside of the original study sample, or to identify solutions that cross-generalize with maximum accuracy when applied across multiple samples [20].

To illustrate the ODA-multivalued treatment framework, and compare it to other commonly-used methods, the paper is organized as follows. In the Methods section we provide a brief introduction to ODA as it is operationalized in the context of multivalued treatments, and describe the data source and analytic framework employed in the current study. The Results section reports and compares the results from the ODA-multivalued treatment framework to several other widely-used methods. The Discussion section describes the specific advantages of the ODA-multivalued treatment framework for assessing covariate balance and evaluating

treatment effects compared with alternative methods, and discusses how machine-learning can be applied more broadly within the causal inferential framework.

METHODS

A brief introduction to optimal discriminant analysis for analyzing multivalued treatments

ODA is a machine learning algorithm that was introduced over 25 years ago to offer an alternative analytic approach to conventional statistical methods commonly used in research [21]. Its appeal lies in its simplicity, flexibility, and accuracy as compared to conventional methods [20,22,23].

An ODA model for multivalued treatments first orders the outcome variable from low to high. It then seeks a specific combination of *cutpoint*(s) and *direction* with respect to the ordered outcome data [19,20,21]. In order to identify potential model cutpoints, the ODA algorithm begins by finding every point along the outcome continuum in which two successive values belong to individuals from different treatment categories (e.g. the previous value belongs to a subject in treatment category 3 whereas the next value belongs to a subject in treatment category 1). For a treatment variable with T categories, an ODA model would generate a total of $T - 1$ cutpoints. For a multivalued treatment with $T = 3$ treatment categories (dummy-coded as 1, 2, 3), for example, the ODA model will have $T = 3 - 1 = 2$ cutpoints. The value of each cutpoint is computed as the mean of the successive outcome values: cutpoint = (previous value + current value) / 2.

Directionality defines how cutpoints are used to classify individual observations. A unidirectional "confirmatory" approach is used when the investigator hypothesizes the order of the treatment categories with respect to the value of the outcome. For example one might hypothesize that observations in treatment category 3 have the lowest values, observations in treatment category 2 have the highest values, and observations in treatment category 1 have values that fall between those of the other class categories. A non-directional "exploratory" approach is used when the investigator has no hypothesis about the order of the treatment categories with respect to the value of the outcome, and the alternative hypothesis tested is that at least two of the categories can be discriminated on the basis of observations' values on the outcome variable. For a directional hypothesis, only the specified ordering of the treatment categories is evaluated, and for a non-directional hypothesis all possible orderings are evaluated.

For the outcome continuum, ODA assesses how well the confirmatory model—consisting of T - 1 cutpoints (that are identified by ODA as described below) in combination with the researcher-specified direction (or "ordering"), or how well the exploratory model—consisting of T - 1 cutpoints identified by ODA for every possible ordering of the class categories, correctly predicts that individuals within a given range of the outcome are in a particular treatment. [19,20]

ODA relies on three measures of accuracy to identify the optimal (maximum-accuracy) model – that is, the exact combination of cutpoint(s) and direction that produces the most accurate predictions possible for the sample. In the multivalued treatment case, *sensitivity* or true

positive rate [24] is the proportion of actual subjects in a given treatment level that are correctly predicted by the ODA model to be in that level -- that is, those who have a value on the outcome that lies within the range specified by the *T* - 1 cutpoints identified by the ODA algorithm [19,20]. The second measure of accuracy combines the sensitivity estimates for each treatment and is called the *effect strength for sensitivity* or ESS [19,20]. ESS is a chance-corrected (0 = the level of accuracy expected by chance) and a maximum-corrected (100 = perfect prediction) index of predictive accuracy. The formula for computing ESS in a multivalued treatment study (multi-category) is:

ESS = [(Mean Percent Accuracy in Classification – *T**)] / (1 – *T**) x 100% (1),

where

Mean Percent Accuracy in Classification = $\sum Sensitivity_t$ / *T* (2)

where *t* is the treatment level in the set of treatments = {0,1,…,*T*} and *T** is the inverse of *T*.

The ODA algorithm explicitly determines the ESS associated with every possible solution under the alternative hypothesis for the sample. The maximally-accurate ("optimal") model is that which has the cutpoint(s) and direction with the highest associated value of ESS. Based on simulation research, ESS values <25% conventionally indicate a relatively weak, <50% indicate a moderate, 50-75% indicate a relatively strong, and ≥75% indicate a strong effect [19,20].

ODA also computes *P*-values to assess the statistical reliability (or "significance") of the maximally-accurate ODA model. *P*-values are estimated using Monte Carlo permutation

experiments. In multivalued treatment models, this involves repeatedly shuffling subjects' treatment assignment at random, holding their outcome value fixed at its true value. In each permuted dataset the ESS is recorded, and the permutation *P*-value represents the proportion of all permuted datasets in which the ESS is higher than the ESS of the maximally-accurate ODA model [19,20,21].

Finally, ODA can be implemented using cross-validation to assess the generalizability of the model, using methods including *k*-fold cross-validation, bootstrapping, and leave-one-out jackknife cross-validation [20,25,26]. This typically entails first estimating a model using a training sample and calculating the accuracy measures, followed by applying the same model to one or more hold-out (test) samples and then recalculating the accuracy measures. If the accuracy measures remain consistent with those of the original model obtained using the training sample, then the model is considered generalizable. This may be important, for example, if the goal of the analysis is to assist health researchers with the identification of new candidates for participation in an ongoing intervention, or initiate the intervention in other settings. Cross-validation is less important if the goal is only to estimate treatment effects of the existing set of interventions [17,27,28].

Data

The data for our empirical example come from a disease management program designed for patients with congestive heart failure and implemented in a large health plan located in the western United States. Individuals with the condition were contacted and invited to enroll in the

program. Those agreeing to participate received one of the following interventions: (1) periodic telephone calls from a nurse to discuss self-management behaviors, or (2) remote tele-monitoring (RTM), which entailed daily electronic transmission of the participant's disease-related symptoms to a database followed by a call from the nurse if symptoms appeared to indicate the onset of an acute exacerbation. Assignment to either intervention arm was conducted by the program nurse and based largely on a subjective assessment of the patient's psycho-social needs, past levels of health care utilization, and the patient's preferred level of contact. The primary goal of the intervention was to reduce avoidable hospitalizations [28]. Patients with congestive heart failure, but not participating in the program, received their usual medical care and served as controls in this study (see [29], and [30] for a more comprehensive description).

The retrospectively collected data consist of observations for 1,359 program participants who completed a full 12 months of the intervention, and 6,612 non-participants who were health-plan members during the same period but were not exposed to the intervention. The sample was divided according to treatment assignment: (a) 6,612 non-participants, (b) 654 participants in the telephonic intervention, and (c) 705 participants in the RTM intervention. Each individual in the dataset has 12 months of pre-intervention data and 12 months of intervention-period data. Pre-intervention characteristics of participants in the three study arms include patient demographic characteristics (age and gender), the Charlson comorbidity index and associated comorbidities [31], and key measures of health care utilization (prescription filled, office visits, emergency

department visits, hospital admissions and hospital days). The primary outcome for all analyses used in this paper is the number of all-cause hospitalizations in the intervention year.

Analytic approach

For the purpose of this empirical example, we repeat the regression-based analyses conducted in Linden et al [30] that used five common methods to estimate multivalued treatment effects. These estimation methods fall into three general categories: (1) estimators based on a model for the outcome variable using conventional regression adjustment (RA); (2) estimators based on a model for the treatment assignment, using inverse probability of treatment weighting (IPTW) [32,33] and marginal mean weighting through stratification (MMWS) [29,34]; and (3) 'doubly-robust' estimators that model both the treatment assignment and outcome variable within the same framework, using an augmented IPTW approach (A-IPTW) [35,36,37] and IPTW combined with RA (IPTW-RA) [10,38,39].

The RA approach was implemented by regressing the outcome -- the number of all-cause hospitalizations in the intervention year – on the set of pre-intervention covariates (described above) separately for each treatment level, after which the predicted outcomes for each subject and treatment level were computed using data only from the individuals receiving the relevant treatment level. The average of those predicted values estimates the potential outcome means, and were then contrasted (Bonferroni corrected) to estimate average treatment effects between all treatment levels [30].

The IPTW approach was implemented by first estimating the generalized propensity score (GPS) [4] using multinomial logistic regression. The levels of treatment (untreated, nurse calls, RTM) served as the outcome, and were regressed on all pre-intervention covariates. Next, the IPT weight was derived by taking the inverse of the propensity score that corresponded with that individual's true treatment assignment. The IPTW was then used as a probability weight within the outcomes model -- which was estimated by regressing the outcome on the treatment indicator variable. The average of those predicted values estimates the potential outcome means, and were then contrasted (Bonferroni corrected) to estimate average treatment effects [30].

The MMWS approach was implemented as follows: First, the GPS was estimated as described for the IPTW. Next, each GPS was stratified into five equal sized quantile categories, separately for each of the estimated probabilities. The marginal mean weights were computed based on the formula by Hong [34]. The MMWS was then used as a probability weight within the outcomes model -- which was estimated by regressing the outcome on the treatment indicator variable (which included the three levels of treatment). The Bonferroni corrected contrasts of these weighted averages provide the treatment effect estimates [30].

The A-IPTW and IPTW-RA approaches belong to a class of estimators that model both the probability of treatment and the outcome simultaneously, within the same framework, providing asymptotically unbiased estimates when only one of the two models is correctly specified. These estimators are called 'doubly robust' because they provide the investigator two opportunities to derive consistent treatment effects [33,36]. The A-IPTW model was

operationalized in a three-step process: First, the parameters of the GPS model were estimated and the IPT weights computed as described previously. Next, separate regression models of the outcome were estimated for each treatment level, and the treatment-specific predicted outcomes for each individual were obtained. Next, unconditional means were estimated using the estimated GPS from the first step, as well as the estimated conditional mean functions. The Bonferroni corrected contrasts of these weighted averages provide the treatment effect estimates [30].

The IPTW-RA model was also operationalized in a three-step process: First, the GPS was estimated, and the IPT weights were computed for each level of treatment. Next, using the estimated IPTW, the outcome models were fitted by a weighted regression for each treatment level, and treatment-specific predicted outcomes for each individual were obtained using the estimated coefficients from this weighted regression. Finally, the means of the treatment-specific predicted outcomes were computed. The Bonferroni corrected contrasts between these averages provide the estimates of the treatment effects [30].

Stata 14.1 (StataCorp, College Station, TX, USA) was used to conduct all regression-based statistical analyses: (1) Naïve treatment effect estimates were derived by regressing the outcome on indicator variables representing the levels of treatment. (2) The RA estimator was implemented using the *teffects ra* command. (3) The IPTW estimator with adjusted weights was implemented using the *teffects ipw* command. (4) MMWS estimates were derived by dividing the sample equally into five strata based on the estimated GPS, computing the MMWS weights by implementing a user-written command for Stata *mmws* [40], and then by regressing the outcome

on indicator variables representing the treatment levels, with the MMWS weights used as sampling weights and applying robust standard errors [29]. (5) The A-IPTW estimator was implemented using the *teffects aipw* command and, (6) The IPTW-RA estimator was implemented using the *teffects ipwra* command. Additionally, pairwise contrasts (treatment effects) were estimated between all treatment levels, and across all estimators studied, using Stata's *pwcompare* command. *pwcompare* performs Wald tests using linear combinations of marginal linear predictions and uses the delta method to estimate the variance. *P* values were then Bonferroni adjusted to account for multiple comparisons. Covariate balance was calculated by implementing the user-written command for Stata *covbal* [41].

The ODA framework was operationalized as follows: First, in order to be consistent with the conventional approaches, the parameters of the GPS model were estimated and the IPT weights computed as described previously. Next an IPT-weighted ODA model was obtained in which the outcome was specified as the attribute and the multivalued treatments were specified as the class variable without assuming *a priori* directionality. Finally, *post hoc* contrasts between treatment level effects were obtained by conducting all possible (Bonferroni corrected) pair-wise comparisons. Exact *P* values were estimated using 25,000 Monte Carlo experiments [20]. All ODA analyses were performed using UniODA Software [19].

The effectiveness of the GPS-based weighting approach in reducing bias across the three treatment conditions was examined by assessing covariate balance using conventional and ODA-based approaches: the conventional method compares the standardized difference in means [42],

and ODA assesses the aforementioned measures of accuracy – sensitivity and ESS [19,20,24]. The expectation is that well-matched cohorts across treatments will have standardized differences close to zero, and poor (i.e., low) measures of accuracy [17]. Analyses were performed on the unweighted population (naïve estimate) and on the GPS weighted sample (adjusted), in order to assess the degree to which weighting reduced confounding and altered the treatment effect estimates.

RESULTS

Assessment of Covariate Balance

Tables of covariate balance using standardized differences are replicated from Linden et al [30] and are presented in Tables 1 and 2, for before and after weighting, respectively. As shown in Table 1, many of the standardized differences are substantially greater than zero and nine of the 69 standardized differences are greater than the 0.25 cutoff recommended by Rubin [43]. In general, the participants in the RTM intervention arm were older and had a higher prevalence of comorbidities than the other two groups. However, all groups were comparable on key measures of health care utilization. Table 2 presents the same pre-intervention characteristics of the study participants after IPT weighting. As shown, all standardized differences are much closer to zero, and no value is greater than 0.25 [30]. In other words, IPT weighting achieved covariate balance and reduced confounding.

Tables 3 and 4 present covariate balance testing using ODA, before and after weighting, respectively. As shown in Table 3, there are several covariates identified as being imbalanced,

based on permuted *P* values <0.05, and most of them concur with those imbalanced covariates identified using standardized differences (Table 1). However, ODA did identify imbalances in four of the 5 utilization and cost measures (the exception was office visits). As seen, the statistically significant imbalances in these utilization and cost variables are driven by moderate to high sensitivity in the RTM group, with much lower sensitivity in the other two treatment groups. In other words, ODA was able to predict assignment to the RTM group rather well, while it was not able to discriminate very well between individuals in the control or nursing call groups. For such a pattern of results, pairwise comparisons used to disentangle statistically significant multivalued treatment omnibus effects generally show that the treatment having higher model sensitivity is significantly different with respect to the outcome variable compared to the treatments with lower model sensitivity -- and also that the latter treatments do not differ significantly with respect to the outcome variable. This general pattern was observed across all covariates, where ODA was able to accurately predict assignment of one the treatment arms, and less accurately in the other two arms. As a consequence, ESS for the omnibus model (which is reported as a measure of "clinical" importance for which higher percentage values represent better classification accuracy and ability to discriminate between groups) is very weak across covariates.

Table 4 presents the same pre-intervention characteristics of the study participants after IPT weighting. The results here are consistent with those of the standardized differences (Table 2). We found consistently weak ESS values throughout, and all permuted *P* values > 0.05. These

results indicate that covariate balance was achieved across all pre-intervention observed covariates.

Assessment of Treatment Effect

Table 5 provides pairwise treatment effect estimates between all treatment arms, by estimator [30]. Here, treatment effects represent the difference between groups in all-cause hospital admissions. In the naïve model (unadjusted for confounding or bias), both intervention arms (calls and RTM) had significantly higher rates of hospital admission than the control arm ($P < 0.001$ and $P = 0.001$ for calls versus controls, and RTM versus controls, respectively), but no statistically significant difference between the intervention arms themselves. All of the regression adjusted methods trended toward similar results. Irrespective of adjustment method, the arm receiving nursing calls had statistically higher hospital admissions than controls, the RTM arm was not statistically different than controls, and the RTM arm had statistically fewer admissions than the arm receiving nursing calls [30].

Table 6 summarizes findings of between-treatment omnibus comparisons of all-cause hospitalizations in the intervention period by treatment level, both unadjusted (naïve) and IPT weighted, using ODA as the analytic tool. Summary values represent the cutoff point on the outcome for each treatment level, and is presented together with the sensitivity of the cutpoint for each treatment level. As shown for the naïve estimate, the ODA model predicted that an individual was in the control group if they had ≤ 0.5 hospitalizations in the intervention period (because number of hospitalizations is an integer-based count, here "$\leq$ 0.05 hospitalizations"

indicates a count of zero hospitalizations), a participant in the nursing call intervention if they had more than 0.5 and less than 4.5 (i.e., between one and four) hospitalizations, and in the RTM intervention if they had greater than 4.5 (i.e., five or more) hospitalizations. The ODA model correctly classified 81% of controls, 29% of participants receiving nursing calls, and 1.6% of participants on RTM. Classification performance was weak (ESS = 5.73%) but statistically significant ($P < 0.0001$) overall. *Post hoc* tests indicated that both RTM and calls had statistically higher hospitalizations than controls, but were not significantly different from each other. However, after controlling for confounding via IPT weighting, ODA reported a miniscule clinical effect (ESS = 0.06%) that was not statistically significant ($P < 0.860$). In other words, after controlling for confounding, ODA found no treatment effect between any of the treatment conditions.

DISCUSSION

Our results demonstrate that ODA can be combined with GPS-based weighting to provide an alternate strategy to regression-based methods for estimating treatment effects in evaluations of multivalued treatments. And while we used IPTW for multivalued treatments in this particular example, the ODA algorithm can be extended to any design where weights are used for covariate adjustment (see for example [29,30,44,45,46,47]).

As our results illustrate, conventional regression-based models and ODA analyses do not always produce consistent results. This is supported by other studies comparing the two methods

that have also obtained strongly divergent findings in a wide variety of real-world data and research designs [19,20]. Thus, a good rule of thumb for investigators is to perform the program evaluation using both conventional and ODA frameworks, and then compare the resulting treatment effect estimates.

If both methods provide consistent results (vis-à-vis statistical significance), then the investigator should be confident that, at the very least, the estimate is insensitive to distributional assumptions required for the validity of *P* values estimated using a regression-based model, and also more likely to be a reflection of the true statistical significance of the treatment effect estimate. However, if the approaches result in conflicting statistical conclusions (as occurred in our empirical example), the investigator should consider the ODA-based *P* values to be most robust, given that ODA uses permutation *P* values that require no distributional assumptions and are therefore always valid [19,20,21].

Of course, in any specific application it is possible that statistical assumptions underlying the validity of effect estimates made by conventional linear methods (that are designed to compare differences in central tendencies) are satisfied for the sample. In the present study statistically significant mean differences were found between treatments, but ODA (that is designed to assess distributional overlap between the different groups) failed to find a statistically significant difference between the distributions of individual scores within each of the multiple groups. If the opposite pattern of results was found (i.e., that there were no significant mean differences (regression), but significant distributional differences (ODA)), this

would indicate that although means do not differ between groups, observations in the different groups can be successfully discriminated on the basis of their individual scores. The remaining potential patterns (i.e., that both mean and distributional differences exist, or that neither mean nor distributional differences exist) are unambiguous in terms of their interpretation.

ODA is an appealing alternative statistical framework in program evaluation because it holds several advantages over conventional methods for assessing covariate balance, outcomes, or both, in observational studies. First, the ODA algorithm, with its associated measure of normed classification performance (ESS) and non-parametric permutation tests, can be universally applied to any variable type and number of study groups, and is not affected by skewed data or outliers – a concern that may arise in the context of meeting assumptions underlying the validity of the estimated *P*-value using conventional statistics alone.

Second, within the proposed treatment effects framework, ODA can also help explain (a) how individuals self-select in observational studies (by identifying group membership based on the cut-point on any given covariate) [LY2; LY5], and (b) how individuals are likely to respond to various levels of the intervention (by identifying where individuals scores are relative to the cutpoint on the outcome) [48]. In the multivalued treatment context, such detail can allow administrators to fine-tune the enrollment criteria to target and assign individuals who will most likely benefit from various levels of the intervention, while concomitantly allowing administrators to improve their estimates of which individuals actually benefit from the various levels of the intervention [49].

Finally, ODA can be implemented using cross-validation to assess the generalizability of the model to new candidates for participation in the existing intervention, or to initiate the intervention in other settings [20]. Cross-validation is less important if the goal is only to estimate treatment effects of the intervention [27,49].

While this paper specifically focused on creating a framework in which machine learning and weighting approaches can be combined to improve causal inference in the evaluation of multivalued treatments, there are several additional ways in which machine learning techniques can be applied in causal inferential work. For example, Linden and Yarnold [49] use classification tree analysis (CTA) to characterize the nature of individuals who choose to participate in observational studies. Athey & Imbens [50] modify the conventional classification and regression trees (CART) approach to estimate heterogeneous causal effects in such studies. CTA has also been proposed as an approach to identify potential instrumental variables (IV) that may provide an unbiased estimate of the causal effect of intervention on the outcome [17]. An IV is a variable that is correlated with the intervention, but not correlated with unobserved confounders of the outcome [51]. Similarly, CTA can be used to identify causal mediation effects. A mediator is an intermediate variable that lies on the casual pathway between treatment and outcome [52]. As indicated by these examples, the application of machine-learning techniques to improve causal inference in observational studies is open to much further exploration. Particular emphasis should be placed on determining the most appropriate algorithm for a given problem -- or a generalization to all algorithms, extension to outcomes with censored

data [20,53], and the development of specific sensitivity analyses for these applications [54] to ensure that the resulting models remain robust to changes in assumptions and inputs

In summary, this paper demonstrates that ODA can be combined with GPS-based weighting to provide an alternate strategy to regression-based methods for estimating treatment effects in evaluations of multivalued treatments. In the present data, the results of this framework were inconsistent with those derived using the conventional approaches. However, given that ODA uses permutation *P* values that require no distributional assumptions and are always valid, ODA-derived *P* value estimates should be considered most robust. ODA provides additional information (e.g., class category sensitivities, ESS, cross-generalizability) than is currently included in conventional approaches. More broadly, health researchers should consider the many potential uses of machine learning algorithms to improve causal inference in observational studies.

REFERENCES

1. Campbell, D. T., & Stanley, J. C. (1966) *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand McNally.
2. Linden, A. (2007) Estimating the effect of regression to the mean in health management programs. *Disease Management and Health Outcomes,* 15, 7-12.
3. Linden, A. (2013) Assessing regression to the mean effects in health care initiatives. *BMC Medical Research Methodology,* 13, 1-7.
4. Imbens, G.W. (2000) The role of the propensity score in estimating dose-response functions. *Biometrika* 87, 706–710.
5. Lechner, M. (2001) Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies*, Lechner M, Pfeiffer F (eds). Physica: Heidelberg, 43–58.
6. Rosenbaum, P.R., & Rubin, D.B. (1983) The central role of propensity score in observational studies for causal effects. *Biometrika,* 70, 41–55.
7. Robins, J,M,, Hernán, M.A., & Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology,* 11, 550–560.
8. Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P.R. (2001) Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96, 1245–1253.

9. Frölich, M. Programme evaluation with multiple treatments. (2004) *Journal of Economic Surveys*, 18, 181–224.
10. Wooldridge, J.M. (2010) *Econometric Analysis of Cross Section and Panel Data. 2nd edn.* Cambridge, MA: MIT Press.
11. Hong, G. (2010) Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics*, 35, 499–531.
12. Rubin, D. (2007) The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20–30.
13. Linden, A., & Roberts, N. (2004) Disease management interventions: What's in the black box? *Disease Management*, 7, 275-291.
14. Linden, A., Butterworth, S., & Roberts, N. (2006) Disease management interventions II: What else is in the black box? *Disease Management*, 9, 73-85.
15. Linden, A. & Samuels, S. J. (2013) Using balance statistics to determine the optimal number of controls in matching studies. *Journal of Evaluation in Clinical Practice*, 19, 968–975.
16. Linden, A. (2015) Graphical displays for assessing covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*, 21, 242–247.
17. Linden, A., & Yarnold, P. R. (*In Print_2*) Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice.*

18. Linden, A., & Yarnold, P. R. (*In Print_5*) Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice.*
19. Yarnold, P.R., & Soltysik, R.C. (2005) *Optimal data analysis: A Guidebook with Software for Windows* Washington, DC: APA Books.
20. Yarnold, P.R., & Soltysik, R.C. (2016) *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286
21. Yarnold, P.R., & Soltysik, R.C. (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences*, *22*, 739-752.
22. Grimm, L.G., & Yarnold, P.R. (Eds.). *Reading and Understanding Multivariate Statistics*. Washington, D.C.: APA Books, 1995.
23. Grimm, L.G., & Yarnold, P.R. (Eds.). *Reading and Understanding More Multivariate Statistics*. Washington, D.C.: APA Books, 2000.
24. Linden, A. (2006) Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice*, 12, 132-139.
25. Witten, I.H., Frank, E., & Hall, M.A. (2011) *Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition*. San Francisco: Morgan Kaufmann.

26. Linden, A., Adams, J., & Roberts, N. (2005) Evaluating disease management program effectiveness: An introduction to the bootstrap technique. *Disease Management and Health Outcomes,* 13, 159-167.
27. Linden, A., Adams, J., & Roberts, N. (2004) The generalizability of disease management program results: getting from here to there. *Managed Care Interface,* 17, 38-45.
28. Linden, A. (2006) What will it take for disease management to demonstrate a return on investment? New perspectives on an old theme. *American Journal of Managed Care,* 12, 217–222.
29. Linden, A. (2014) Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions. *Journal of Evaluation in Clinical Practice,* 20, 1065–1071.
30. Linden, A., Uysal, S.D., Ryan, A., & Adams, J.L. (2016) Estimating causal effects for multivalued treatments: A comparison of approaches. *Statistics in Medicine,* 35, 534-552.
31. Charlson, M.E., Pompei, P., Ales, K.L., McKenzie, C.R. (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Disease,* 40, 373–383.
32. Rosenbaum, P.R. (1987) Model-based direct adjustment. *Journal of the American Statistical Association*, 82, 387–394.
33. Robins, J.M., & Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122–129.

34. Hong, G. (2012) Marginal mean weighting through stratification: a generalized method for evaluating multi-valued and multiple treatments with non-experimental data. *Psychological Methods*, 17, 44–60.
35. Robins, J.M. (2000) Robust estimation in sequentially ignorable missing data and causal inference models. In *1999 Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association: Alexandria, VA, 2000, 6–10.
36. Bang, H., & Robins, J.M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.
37. Cattaneo, M.D. (2010) Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics,* 155, 138 –154.
38. Wooldridge, J.M. (2007) Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141, 1281–1301.
39. Uysal, S.D. (2015) Doubly robust estimation of causal effects with multivalued treatments: an application to the returns to schooling. *Journal of Applied Econometrics*, 30, 763–786.
40. Linden A. (2014) MMWS: Stata module for implementing mean marginal weighting through stratification. Statistical Software Components s457886, Boston College Department of Economics, 2014. Downloadable from http://ideas.repec.org/c/boc/bocode/s457886.html [Accessed on 6 June 2016].
41. Linden, Ariel (2016). COVBAL: Stata module for generating covariate balance statistics. http://ideas.repec.org/c/boc/bocode/s458188.html [Accessed on 6 June 2016].

42. Flury, B.K. & Reidwyl, H. (1986) Standard distance in univariate and multivariate analysis. *The American Statistician*, 40, 249–251.

43. Rubin, D.B. (2001) Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, 2, 169–188.

44. Linden, A., & Adams, J.L. (2010) Using propensity score-based weighting in the evaluation of health management programme effectiveness. *Journal of Evaluation in Clinical Practice, 16,* 175-179.

45. Linden, A., & Adams, J.L. (2010) Evaluating health management programmes over time. Application of propensity score-based weighting to longitudinal data. *Journal of Evaluation in Clinical Practice, 16,* 180-185.

46. Linden, A., & Adams, J.L. (2011) Applying a propensity-score based weighting model to interrupted time series data: improving causal inference in program evaluation. *Journal of Evaluation in Clinical Practice,* 17, 1231-1238.

47. Linden, A., & Adams, J.L. (2012) Combining the regression-discontinuity design and propensity-score based weighting to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice,* 18, 317-325.

48. Linden, A., Yarnold, P.R., & Nallomothu, B.K. (*In Print*) Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice*

49. Linden, A., & Yarnold, P.R. (*In Print_1*) Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice.*

50. Athey, S., & Imbens, G. (2015) Recursive partitioning for heterogeneous causal effects. *Working Paper*. Downloadable from http://arxiv.org/abs/1504.01132 [Accessed on 14 May 2016].
51. Linden, A., & Adams, J. (2006) Evaluating disease management program effectiveness: an introduction to instrumental variables. *Journal of Evaluation in Clinical Practice,* 12, 148-154.
52. Linden, A., & Karlson, K.B. (2013) Using mediation analysis to identify causal mechanisms in disease management interventions. *Health Services and Outcomes Research Methodology*, 13, 86-108.
53. Linden, A., Adams, J., & Roberts, N. (2004) Evaluating disease management program effectiveness: An introduction to survival analysis. *Disease Management*, 7, 180-190.
54. Linden, A., Adams, J., & Roberts, N. (2006) Strengthening the case for disease management effectiveness: unhiding the hidden bias. *Journal of Evaluation in Clinical Practice*, 12, 140-147.

Table 1: Unadjusted baseline (prior 12 months) characteristics of program participants and non-participants in a multivalued treatment study (From [30])

				Absolute Standardized Differences		
Variables	Control	Calls	RTM	Calls vs Controls	RTM vs Controls	RTM vs Calls
N	6612	654	705			
Female	2976 (45.0%)	308 (47.1%)	343 (48.7%)	0.042	0.073	0.031
Age, mean (SD)	62.96 (15.77)	66.17 (14.55)	72.31 (12.42)	0.212	0.659	0.454
Charlson index score , mean (SD)	2.64 (2.52)	3.28 (2.57)	3.67 (2.63)	0.250	0.399	0.150
Diabetes (non-comp)	1723 (26.1%)	244 (37.3%)	281 (39.9%)	0.244	0.297	0.052
Diabetes (comp)	697 (10.5%)	122 (18.7%)	130 (18.4%)	0.231	0.226	0.006
Acute MI	782 (11.8%)	112 (17.1%)	162 (23.0%)	0.151	0.297	0.147
Chronic Pulmonary	1468 (22.2%)	177 (27.1%)	251 (35.6%)	0.113	0.299	0.185
Liver (mild)	396 (6.0%)	35 (5.4%)	32 (4.5%)	0.028	0.065	0.038
Liver (Mod/Severe)	48 (0.7%)	3 (0.5%)	5 (0.7%)	0.035	0.002	0.033
Cancer	784 (11.9%)	80 (12.2%)	97 (13.8%)	0.012	0.057	0.045
Cancer - metastatic	140 (2.1%)	11 (1.7%)	10 (1.4%)	0.032	0.053	0.021
Rheumatoid	228 (3.4%)	30 (4.6%)	26 (3.7%)	0.058	0.013	0.045
Cerebrovascular	952 (14.4%)	107 (16.4%)	132 (18.7%)	0.054	0.117	0.062
Peripheral vascular	874 (13.2%)	115 (17.6%)	150 (21.3%)	0.121	0.214	0.093
Renal	1083 (16.4%)	160 (24.5%)	214 (30.4%)	0.202	0.335	0.132
Dementia	164 (2.5%)	10 (1.5%)	8 (1.1%)	0.068	0.101	0.034
Hemi or Paraplegia	130 (2.0%)	10 (1.5%)	11 (1.6%)	0.033	0.031	0.003
Peptic ulcer	105 (1.6%)	12 (1.8%)	13 (1.8%)	0.019	0.020	0.001
Prescriptions, mean (SD)	41.10 (37.42)	49.37 (38.90)	55.32 (37.18)	0.217	0.381	0.156
Office visits, mean (SD)	0.42 (0.93)	0.47 (0.83)	0.44 (0.84)	0.056	0.014	0.044
ED visits, mean (SD)	0.49 (1.30)	0.51 (1.04)	0.44 (0.95)	0.017	0.046	0.072
Hospitalizations, mean (SD)	0.64 (1.15)	0.74 (1.07)	0.64 (1.04)	0.088	0.006	0.099
Hospital days, mean (SD)	3.66 (11.61)	3.74 (8.60)	3.21 (16.09)	0.008	0.032	0.041

Note: All variables are reported as N (%) unless otherwise noted. RTM is remote telemonitoring.

Table 2: Weighted baseline (prior 12 months) characteristics of program participants and non-participants in a multivalued treatment study (From [30])

				Absolute Standardized Differences		
Variables	Control	Calls	RTM	Calls vs Controls	RTM vs Controls	RTM vs Calls
N	6612	654	705			
Female	1227 (45.5)	1227 (45.9)	1182 (45.5)	0.007	0.001	0.008
Age, mean (SD)	64.11 (15.85)	64.50 (15.05)	65.72 (13.89)	0.026	0.109	0.086
Charlson index score , mean (SD)	2.79 (2.59)	2.83 (2.43)	2.90 (2.51)	0.014	0.044	0.029
Diabetes (non-comp)	764 (28.3%)	792 (29.6%)	757 (29.1%)	0.028	0.018	0.010
Diabetes (comp)	323 (12.0%)	332 (12.4%)	343 (13.2%)	0.013	0.037	0.021
Acute MI	359 (13.3%)	368 (13.8%)	352 (13.6%)	0.014	0.008	0.006
Chronic Pulmonary	642 (23.8%)	638 (23.9%)	634 (24.4%)	0.001	0.014	0.013
Liver (mild)	157 (5.8%)	164 (6.1%)	147 (5.7%)	0.013	0.007	0.021
Liver (Mod/Severe)	19 (0.7%)	18 (0.7%)	22 (0.8%)	0.002	0.016	0.020
Cancer	325 (12.1%)	302 (11.3%)	312 (12.0%)	0.024	0.001	0.022
Cancer - metastatic	54 (2.0%)	52 (1.9%)	50 (1.9%)	0.006	0.007	0.002
Rheumatoid	95 (3.5%)	92 (3.4%)	96 (3.7%)	0.006	0.008	0.013
Cerebrovascular	405 (15.0%)	426 (15.9%)	444 (17.1%)	0.024	0.056	0.031
Peripheral vascular	388 (14.4%)	408 (15.3%)	396 (15.2%)	0.025	0.024	0.001
Renal	495 (18.4%)	493 (18.4%)	513 (19.7%)	0.002	0.035	0.031
Dementia	62 (2.3%)	70 (2.6%)	87 (3.3%)	0.022	0.063	0.046
Hemi or Paraplegia	51 (1.9%)	56 (2.1%)	81 (3.1%)	0.014	0.078	0.068
Peptic ulcer	44 (1.6%)	42 (1.6%)	47 (1.8%)	0.007	0.013	0.019
Prescriptions, mean (SD)	43.24 (39.07)	43.92 (36.00)	46.22 (35.14)	0.019	0.082	0.062
Office visits, mean (SD)	0.43 (0.92)	0.44 (0.82)	0.44 (0.84)	0.015	0.012	0.003
ED visits, mean (SD)	0.49 (1.26)	0.50 (1.07)	0.47 (0.98)	0.015	0.014	0.033
Hospitalizations, mean (SD)	0.65 (1.15)	0.67 (1.02)	0.65 (1.07)	0.018	0.000	0.019
Hospital days, mean (SD)	3.64 (11.20)	3.47 (8.63)	3.50 (18.13)	0.017	0.009	0.002

Note: All variables are reported as N (%) unless otherwise noted. RTM is remote telemonitoring

Table 3: Unadjusted baseline (prior 12 months) characteristics of program participants and non-participants in a multivalued treatment study using ODA

						Bonferroni Adjusted *P*-values		
Variables	Control	Calls	RTM	ESS (%)	*P*-value	Calls vs Controls	RTM vs Controls	RTM vs Calls
N	6612	654	705					
Female	= 0 (83.70)	= 1 (17.95)	= 1 (17.95)	1.65	0.058	--	--	--
Age	≤ 59.5 (42.45)	> 59.5 & ≤ 64.5 (21.56)	> 64.5 (70.35)	17.18	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Charlson index score	≤ 1.5 (41.62)	> 1.5 & ≤ 3.5 (33.94)	> 3.5 (45.53)	10.55	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Diabetes (non-comp)	= 0 (85.43)	= 1 (23.35)	= 1 (23.35)	8.78	< 0.0001	--	--	--
Diabetes (comp)	= 0 (84.24)	= 1 (26.55)	= 1 (26.55)	10.79	< 0.0001	--	--	--
Acute MI	= 0 (84.31)	= 1 (25.95)	= 1 (25.95)	10.26	< 0.0001	--	--	--
Chronic Pulmonary	= 0 (84.67)	= 1 (22.57)	= 1 (22.57)	7.25	< 0.0001	--	--	--
Liver (mild)	= 0 (17.21)	= 1 (85.53)	= 0 (17.21)	2.74	0.185	--	--	--
Liver (Mod/Severe)	= 0 (8.22)	= 1 (94.64)	= 0 (8.22)	2.78	0.757	--	--	--
Cancer	= 0 (83.14)	= 1 (18.42)	= 0 (83.14)	1.56	0.305	--	--	--
Cancer - metastatic	= 1 (86.96)	= 0 (17.13)	= 0 (17.13)	4.09	0.252	--	--	--
Rheumatoid	= 0 (83.05)	= 1 (19.72)	= 1 (19.72)	2.77	0.306	--	--	--
Cerebrovascular	= 0 (83.48)	= 1 (20.07)	= 1 (20.07)	3.55	0.003	--	--	--
Peripheral vascular	= 0 (83.99)	= 1 (23.27)	= 1 (23.27)	7.25	< 0.001	--	--	--
Renal	= 0 (84.88)	= 1 (25.67)	= 1 (25.67)	10.55	< 0.001	--	--	--
Dementia	= 1 (90.11)	= 0 (17.22)	= 0 (17.22)	7.33	0.009	--	--	--
Hemi or Paraplegia	= 1 (86.09)	= 0 (17.11)	= 0 (17.11)	3.20	0.412	--	--	--
Peptic ulcer	= 0 (82.99)	= 1 (19.23)	= 1 (19.23)	2.22	0.693	--	--	--
Prescriptions	≤ 17.5 (32.03)	> 17.5 & ≤ 21.5 (5.05)	> 21.5 (84.26)	10.67	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Office visits	≤ 0.5 (72.78)	> 0.5 & ≤ 3.5 (31.04)	> 3.5 (1.13)	2.48	0.093	0.032	1.000	0.309

ED visits	> 2.5 (4.25)	> 0.5 & ≤ 2.5 (29.20)	≤ 0.5 (73.62)	3.54	0.009	0.095	1.000	0.061
Hospitalizations	> 5.5 (0.67)	> 0.5 & ≤ 5.5 (43.88)	≤ 0.5 (62.84)	3.69	0.012	0.005	1.000	0.015
Hospital days	> 28.5 (2.51)	> 0.5 & ≤ 28.5 (42.20)	≤ 0.5 (62.84)	3.77	0.026	0.011	0.770	0.050

Note: Values represent cut-points on the covariate, and values in parentheses represent sensitivity. For binary covariates (e.g. female), the model is specified with the covariate as the class and treatment level as the attribute. Thus, pairwise comparisons are not relevant and noted as "--". Comp is complicated, MI is myocardial infarction, mod is moderate, ED is emergency department, ESS is effect strength for sensitivity.

Table 4: Weighted baseline (prior 12 months) characteristics of program participants and non-participants in a multivalued treatment study using ODA

						Bonferroni Adjusted *P*-values		
Variables	Control	Calls	RTM	ESS (%)	*P*-value	Calls vs Controls	RTM vs Controls	RTM vs Calls
N	6612	654	705					
Female	= 1 (67.5)	= 1 (67.5)	= 0 (32.7)	0.19	1.000	--	--	--
Age	≤ 20.5 (0.5)	> 20.5 & ≤ 64.5 (58.8)	> 64.5 (50.0)	4.61	0.074	0.178	0.080	0.215
Charlson index score	> 15.5 (0.1)	> 0.5 & ≤ 15.5 (89.9)	≤ 0.5 (12.2)	1.08	0.501	1.000	1.000	1.000
Diabetes (non-comp)	= 1 (100)	= 1 (100)	= 1 (100)	0.00	1.000	--	--	--
Diabetes (comp)	= 1 (100)	= 1 (100)	= 1 (100)	0.00	1.000	--	--	--
Acute MI	= 1 (100)	= 1 (100)	= 1 (100)	0.00	1.000	--	--	--
Chronic Pulmonary	= 1 (100)	= 1 (100)	= 1 (100)	0.00	1.000	--	--	--
Liver (mild)	= 1 (68.6)	= 1 (68.6)	= 0 (32.7)	1.24	1.000	--	--	--
Liver (Mod/Severe)	= 0 (67.4)	= 0 (67.4)	= 1 (67.4)	4.41	1.000	--	--	--
Cancer	= 0 (100)	= 0 (100)	= 0 (100)	0.00	1.000	--	--	--
Cancer - metastatic	= 1 (34.9)	= 0 (66.2)	= 0 (66.2)	1.08	1.000	--	--	--
Rheumatoid	= 0 (67.4)	= 0 (67.4)	= 1 (33.9)	1.35	1.000	--	--	--
Cerebrovascular	= 1 (100)	= 1 (100)	= 1 (100)	0.00	1.000	--	--	--
Peripheral vascular	= 1 (100)	= 1 (100)	= 1 (100)	0.00	1.000	--	--	--
Renal	= 1 (100)	= 1 (100)	= 1 (100)	0.00	1.000	--	--	--
Dementia	= 1 (100)	= 1 (100)	= 1 (100)	0.00	1.000	--	--	--
Hemi or Paraplegia	= 1 (100)	= 1 (100)	= 1 (100)	0.00	1.000	--	--	--
Peptic ulcer	= 1 (68.7)	= 0 (33.6)	= 1 (68.7)	2.32	1.000	--	--	--
Prescriptions	> 186 (0.6)	≤ 57.5 (70.9)	> 57.5 & ≤ 186 (33.0)	2.30	0.744	0.821	0.513	1.000
Office visits	> 6.5 (0.2)	≤ 2.5 (96.9)	> 2.5 & ≤ 6.5 (4.2)	0.65	0.350	0.448	0.801	1.000

ED visits	> 8.5 (0.2)	≤ 4.5 (99.1)	> 4.5 & ≤ 8.5 (1.6)	0.44	0.465	1.000	0.546	1.000
Hospitalizations	> 6.5 (0.4)	≤ 3.5 (97.9)	> 3.5 & ≤ 6.5 (4.0)	1.13	0.163	0.284	0.230	0.510
Hospital days	> 91.5 & ≤ 109.5 (0.1)	≤ 91.5 (100)	> 109.5 (0.5)	0.29	0.693	0.516	1.000	1.000

Note: Values represent cut-points on the covariate, and values in parentheses represent sensitivity. For binary covariates (e.g. female), the model is specified with the covariate as the class and treatment level as the attribute. Thus, pairwise comparisons are not relevant and noted as "--". Comp is complicated, MI is myocardial infarction, mod is moderate, ED is emergency department, ESS is effect strength for sensitivity.

Table 5: Contrasts (Bonferroni adjusted) between treatment levels on all-cause hospitalizations during the intervention period, by regression-based causal estimator (From [30])

Estimator	Contrast	SE	z	P>|z|	[95% Conf.	Interval]
Naive						
Calls vs Control	0.215	0.037	5.87	<0.001	0.127	0.303
RTM vs Control	0.128	0.035	3.63	0.001	0.044	0.213
RTM vs Calls	-0.087	0.049	-1.78	0.223	-0.203	0.030
Regression adjustment						
Calls vs Control	0.179	0.046	3.93	<0.001	0.070	0.288
RTM vs Control	0.023	0.039	0.59	1.000	-0.070	0.115
RTM vs Calls	-0.156	0.058	-2.69	0.021	-0.295	-0.017
MMWS						
Calls vs Control	0.193	0.052	3.68	0.001	0.067	0.319
RTM vs Control	0.013	0.037	0.36	1.000	-0.075	0.102
RTM vs Calls	-0.180	0.062	-2.88	0.012	-0.329	-0.030
IPTW						
Calls vs Control	0.180	0.043	4.16	<0.001	0.077	0.284
RTM vs Control	0.029	0.039	0.74	1.000	-0.064	0.122
RTM vs Calls	-0.152	0.057	-2.68	0.022	-0.287	-0.016

A-IPTW						
Calls vs Control	0.179	0.045	4.01	<0.001	0.072	0.285
RTM vs Control	0.015	0.035	0.44	1.000	-0.069	0.100
RTM vs Calls	-0.163	0.055	-2.97	0.009	-0.294	-0.032
IPTW-RA						
Calls vs Control	0.180	0.045	4.03	<0.001	0.073	0.286
RTM vs Control	0.014	0.033	0.43	1.000	-0.065	0.094
RTM vs Calls	-0.165	0.054	-3.09	0.006	-0.293	-0.037

Note: MMWS is marginal mean weighting through stratification, IPTW is inverse probability of treatment weighting, A-IPTW is augmented inverse probability of treatment weighting, and IPTW-RA is inverse probability of treatment weighting with regression adjustment, RTM is remote telemonitoring.

Table 6: Contrasts (Bonferroni adjusted) between treatment levels on all-cause hospitalizations during the intervention period, using ODA

						Bonferroni Adjusted *P*-Values		
Model	Control	Calls	RTM	ESS (%)	*P*-value	Calls vs Controls	RTM vs Controls	RTM vs Calls
N	6612	654	705					
Naïve ODA	≤ 0.5 (81.00)	> 0.5 & ≤ 4.5 (28.90)	> 4.5 (1.56)	5.73	< 0.0001	< 0.0001	< 0.0001	0.122
Weighted ODA	> 7.5 (0.15)	≤ 6.5 (99.82)	> 6.5 & ≤ 7.5 (0.14)	0.06	0.860	1.000	0.786	1.000

Note: Values represent cut-points on the covariate, and values in parentheses represent sensitivity. ESS is effect strength for sensitivity.