Maximizing Predictive Accuracy By P. R. Yarnold R. C. Soltysik. ODA Books, Chicago, IL, 2016, $98.00, 396 pp. ISBN 0 692 70092 7.

Ariel Linden

Received: 12 July 2016 | Accepted: 13 July 2016

DOI 10.1111/jep.12625

WILEY Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

BOOK REVIEW

MAXIMIZING PREDICTIVE ACCURACY

By P. R. Yarnold | R. C. Soltysik

ODA Books, Chicago, IL, 2016, $98.00, 396 pp. ISBN 0 692 70092 7.

1 | INTRODUCTION

With the ever-growing quantity and availability of health care data, many health scientists are increasingly turning to automated analytic processes to identify meaningful patterns in the data to improve diagnostic accuracy, identify high-risk patients, and extract concepts in unstructured data.[1] However, some traditionalists still argue that automated algorithms are not a viable substitute for the tried-and-true approach to solving research problems, which relies on a mixture of content expertise, experience, intuition, and ultimately, conventional statistics (see Breiman[2] for an excellent discussion on the discord between these 2 cultures).

The optimal data analysis (ODA) paradigm described by Yarnold and Soltysik in *Maximizing Predictive Accuracy*[3] is a machine learning algorithm that was introduced over 25 years ago to offer an alternative to conventional statistical methods commonly used in research.[4] It bridges the divide between data mining and statistics, easily overcoming many of the concerns put forth by traditional health researchers. Its appeal lies in its simplicity, accuracy, versatility, and transparency, compared with conventional methods. By framing the relationship between the outcome variable and independent variable as a classification problem (ie, how accurately does the outcome variable classify individuals as belonging to their actual level of the independent variable?), ODA offers several benefits over the conventional statistical methods typically employed in most health research studies. These include the ability to handle an outcome variable measured on any scale (from categorical to continuous), insensitivity to skewed data or outliers, the use of accuracy measures that can be widely applied to all classification analyses, and *P* values estimated using Monte Carlo permutation tests.

Using this classification approach, ODA additionally offers the unique ability to ascertain if individuals are likely to respond to the assigned treatment (such as doses of a drug[5] or adherence to behavioral-change interventions[6]) based on maximally accurate cutpoints on the outcome variable, thus making this an ideal approach for evaluating dose-response relationships,[7,8] or interventions with multivalued treatments.[9]

Moreover, ODA accepts analytic weights, thereby allowing the evaluation of observational studies using any algorithm that produces weights for covariate adjustment.[9–16] Finally, ODA provides the capability to use cross-validation in assessing the generalizability of the model to individuals outside of the original study sample,[17] or to identify solutions that cross-generalize with maximum accuracy when applied across multiple samples.[3]

Before describing the book *Maximizing Predictive Accuracy* in greater detail, it would be helpful to briefly explain how an ODA model is obtained. Assume we are evaluating the effectiveness of an intervention with 2 treatment levels (treatment and control) and a continuous outcome variable. First, the ODA algorithm orders the outcome variable from low to high. Next, ODA finds all the points along the continuum of the outcome in which the next value belongs to an individual from the alternate treatment than that of the previous value (eg, the next value belongs to a treated subject, whereas the previous value belongs to a control). The *cutpoint* thus represents the mean value of the outcome at this point: cutpoint = (previous value + current value)/2. *Directionality* defines how cutpoints are used to classify individual observations. The 2 directions are "less than" (controls have lower values on the outcome than treated subjects) and "greater than" (controls have higher values on the outcome than treated subjects). For an exploratory "2-tailed" hypothesis (controls and treated subjects have different values on the outcome), and both directions are evaluated by the ODA algorithm. For a confirmatory "1-tailed" hypothesis (controls have lower values), only the appropriate direction (less than) is evaluated. For each cutpoint along the continuum of the outcome, ODA assesses how well the model—that is, the combination of cutpoint and direction—correctly predicts (in the current example) that controls have values of the outcome less than or equal to the cutpoint, and treated subjects have values of the outcome greater than the cutpoint.

Optimal data analysis relies on 3 measures of accuracy to identify the optimal (maximum-accuracy) model—that is, the exact combination of cutpoint and direction that produces the most accurate predictions possible for the sample. *Sensitivity* or true positive rate is the proportion of actual treated subjects that are correctly predicted by the ODA model—that is, those who have a value on the outcome that lies above the cutpoint. *Specificity* or true negative rate is the proportion of actual control subjects that are correctly predicted by the ODA model—that is, those who have a value on the outcome that lies at or below the cutpoint.[18] The third measure of accuracy combines these 2 metrics and is called the effect strength for sensitivity or ESS.[3,19] The ESS is a chance-corrected (0 = the level of accuracy expected by chance) and maximum-corrected (100 = perfect prediction) index of predictive accuracy. The formula for computing ESS for a binary (2-category) case classification result is

$$\text{ESS} = [(\text{Mean Percent Accuracy in Classification}-50)]/50\times 100\%, \quad (1)$$

where

$$\text{ESS} = [(\text{Mean Percent Accuracy in Classification} - 50)]/50 \times 100\%. \quad (2)$$

The ODA algorithm iterates through each successive cutpoint and calculates ESS. The maximally accurate model is that which has the cutpoint and direction with the highest associated value of ESS. Based on simulation research, ESS values < 25% conventionally indicate a relatively weak effect, <50% indicate a moderate effect, 50% to 75% indicate a relatively strong effect, and ≥75% indicate a strong effect.[3,19]

The 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 tests. For example, in models with a binary treatment, this involves repeatedly shuffling subjects' treatment assignment at random, holding their outcome value fixed at its true value. In each permuted data set, the ESS is recorded, and the permutation *P* value represents the proportion of all permuted data sets in which the ESS is higher than the ESS of the maximally accurate ODA model.[3,19]

Finally, ODA can be implemented using cross-validation to assess the generalizability of the model, using leave-1-out cross-validation. Leave-1-out is simply *n*-fold cross-validation, where *n* is the number of observations in the data set. Each observation in turn is left out, and the model is estimated for all remaining observations. The predicted value is then calculated for the 1 hold-out observation, and the accuracy is determined as success or failure in predicting the outcome for that observation. The results of all n predictions are used to calculate the final accuracy estimates displayed in the classification tables, which are then compared with the original estimates.[20] If the accuracy measures remain consistent with those of the original model using the entire sample, then the model is considered generalizable. This may be important, for example, if the goal of the analysis is to assist health researchers identify new candidates for participation in an ongoing intervention, or initiate the intervention in other settings.[17] Other methods used for assessing reproducibility include hold-out validity assessment in which the model developed using a "training" sample is applied to classify observations in one or more independent samples, and the "generalizability" algorithm that identifies a model that—when independently applied to 2 or more samples—maximizes the minimum ESS value obtained across samples: the model is said to cross-generalize if the minimum value of ESS meets or exceeds the a priori specification of the researcher for adequate fit.[3,19]

The ODA framework just described is implemented identically for all models, regardless of variable type or number of categories. In fact, the most sophisticated of models—classification tree analyses (CTAs)—are nothing more than a series of linked simple ODA models. Taken together, the ODA framework offers health researchers a powerful machine learning alternative to conventional approaches to solving research problems.

2 | OVERVIEW OF THE BOOK

Maximizing Predictive Accuracy unfolds in 4 major sections. The Introduction consists of 3 chapters that lay the groundwork for the rest of the work. The book begins with a brief history of the ODA paradigm—starting with its genesis from the field of operations research (mathematically enabling the algorithm to identify the most accurate and parsimonious model possible for a given data set) and evolving vis-à-vis its wedding to a statistical methodology for which no distributional assumptions are required (*P* values are always correct). The Discussion is presented concerning how best learn the paradigm, about publishing journal articles and teaching the paradigm to students, and about obtaining research funding as well as powering commercial applications. The Chapter 1 ends by describing the ultimate objective of the authors—offering an improved statistical framework to increase the speed and precision of research design and statistical discovery.

Chapter 2 discusses fundamental concepts central to every research study. First, the UniODA algorithm is clearly described and illustrated without the use of mathematical formulas—it is demonstrated how a maximum-accuracy model is identified. Once the maximum-accuracy solution is identified, the next step is assessing the statistical reliability of the model, and once again, without the use of formulas, the reader learns how to assess the exact probability of a given result. Then the chapter turns to the heart of the paradigm—defining predictive accuracy. The multivariable version of the ODA algorithm—CTA—is then introduced. Crucial data transformations and various methodologies for assessing the cross-generalizability (reproducibility) of models to independent samples are then discussed (and illustrated throughout the rest of the book). The chapter ends with a description of the Simpson paradox, which the authors argue may be the most important challenge to and shortcoming of the published literature.

Chapter 3 discusses aspects of measurement that "make or break" empirical research: scales, analytic weights, precision, algorithm adaptability, and instrumentation. Statistical power analysis is covered next, followed by pragmatic issues such as data set design and construction, missing data and residual analysis, and reporting of analytic findings in research reports and articles.

The second section of the book consists of 2 chapters: Chapter 4 concerns categorical "attributes" (known as dependent variables in prior statistical paradigms), and Chapter 5 concerns ordered attributes. Both chapters feature many worked examples of optimal (maximum-accuracy) analogues to a myriad of earlier statistical methods. These examples clearly demonstrate that the ODA algorithm is capable of addressing all of these designs reflected in the myriad of legacy methods—the universal applicability of the algorithm to any data geometry is unique.

The third section of the book consists of 4 chapters focusing on multivariable *linear* models. Chapter 6 describes the optimization (maximization of predictive accuracy) of models developed using the general linear model paradigm. It is demonstrated how the phenomenon of "regression toward the mean" can be remedied, and how the predictive accuracy of analysis of variance and linear discriminant functions can be maximized for any given sample. Chapter 7 addresses models developed using the maximum likelihood paradigm, demonstrating how to maximize the predictive accuracy of legacy methods such as the log-linear model, logistic regression, and probit analysis. Chapter 8 describes linear models that specifically maximize predictive accuracy—and achieve greater accuracy than their general linear model–

and maximum likelihood–based counterparts. However, a host of specific algorithms are also described that yield even more accurate predictions for various forms of hypotheses that differ in specificity and in structure. Chapter 9 concludes by demonstrating that all linear models are susceptible to paradoxical confounding, whether attributable to covariates, or to pooling of groups and/or time periods, and shows that this problem exists even in single-case designs.

The fourth and final section of the book consists of 3 chapters focusing on multivariable *nonlinear* CTA models. Chapter 10 discusses the first generation of CTA model, which has produced many of the most accurate and parsimonious models ever identified in a host of applications. It is demonstrated how this analysis can be conducted manually using ODA software. Chapter 11 describes the second generation of CTA model, which has bested the first-generation CTA model in most applications that compared the methods—identifying more accurate and typically more parsimonious solutions. This is accomplished by enumerating all possible orderings of the initial 3 attributes included in the model—since these are the attributes that dominate the ultimate predictive accuracy that is achieved by the model. At this point, the book reveals that through all of this development, there remain unsolved issues that exist for the optimal methods introduced thus far, and that incurably cripple all legacy statistical methods.

Chapter 12 answers the enigma by describing the third and final generation CTA model—the only analysis capable of identifying a globally optimal model (ie, accomplished by effectively enumerating all possible models) for a given application. It is discovered that the relationship between X and Y is not the same as the relationship between Y and X except in specific geometries, and that for many samples there is a discrete family of models relating X and Y, or Y and X. The idea of a statistically ideal model—that achieves perfect accuracy and does so with maximum possible parsimony—is defined, and a new statistical index is introduced that allows one to compare any model with respect to their distance from the theoretical ideal. This is presented in the context of novometric (ie, "new measurement") theory, consisting of 4 axioms that parallel the basic axioms of quantum mechanics, but that apply to classical rather than to atomic phenomena. Exact discrete confidence intervals are described for models as well as for chance. These methods are illustrated and show, in a study of gender and cancer mortality for example, that there are more than 1 type (strata) of male and more than 1 type of female: not all males are alike, and not all females are alike.

3 | COMMENT

Needless to say, learning an entirely new methodological approach is not always easy, especially one as comprehensive as the ODA paradigm described in this book. However, the book is well organized, building from a simple model onward to the most sophisticated of classification algorithms available. The many brief examples are useful and generalizable, allowing researchers from any discipline to contemplate the application of the ODA framework to their own work.

At 396 pages in length, this book provides an encyclopedic level of detail on maximum-accuracy models. Yet for all that, the book is not exhaustive. An entirely new volume could be devoted to novel extensions of the paradigm to specific areas of research. For example, a recent series of papers has described ways in which ODA can be applied to improve causal inference in observational studies.[7–11,21] However, CTA also has many potential applications for improving causal inferential work. For example, CTA should be investigated as an approach for modeling heterogeneous causal effects in observational studies. One can also envision the use of CTA to identify potential instrumental variables that may provide an unbiased estimate of the causal effect of an *intervention* on *the outcome* (IV). An IV is a variable (Z) that is correlated with the intervention (X) but not associated with unobserved confounders of the outcome (Y).[22] Potential IVs may be identified by first generating a CTA model predicting participation and then generating a second model predicting the outcome—allowing the same set of covariates in both models. Covariates that appear in the first (selection) model, but not in the second (outcome) model, may be suggestive of potential IVs, which can then be used within the IV framework. Similarly, CTA should be considered as an approach for identifying causal mediation effects. A mediator is an intermediate variable that lies on the casual pathway between treatment and outcome.[23] A CTA model would be generated to predict the outcome, forcing the inclusion of the mediator after the treatment (to ensure correct temporal alignment), as well as including other covariates to control for confounding. In such a model, the extent of mediation effects can be elucidated by assessing the ESS and *P* values for each node along the pathway from treatment to outcome via the mediator. As indicated by these examples, the application of maximum-accuracy 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,[24] and the development of specific sensitivity analyses for these applications[25] to ensure that the resulting models remain robust to changes in assumptions and inputs.

In summary, I strongly recommend this book for any health researcher interested in learning about an entirely novel approach to evaluating their research—one that combines the power of machine learning with permutation *P* values that require no distributional assumptions, to deliver models with *maximum predictive accuracy*.

Ariel Linden DrPH[1,2]*
[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

REFERENCES

1. Iavindrasana J, Cohen G, Depeursinge A, Müller H, Meyer R, Geissbuhler A. Clinical data mining: a review. *Yearb Med Inform*. 2009;121–133.

2. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci*. 2001;16:199–231.

3. Yarnold PR, Soltysik RC. *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books; 2016.

4. Yarnold PR, Soltysik RC. Theoretical distributions of optima for univariate discrimination of random data. *Decis Sci*. 1991;22:739–752.

5. Couto J, Webster L, Romney M, Leider H, Linden A. Using an algorithm applied to urine drug screening to assess adherence to an OxyContin regimen. *J Opioid Manag*. 2009;5:359–364.

6. Linden A, Butterworth S, Roberts N. Disease management interventions II: what else is in the black box? *Dis Manag*. 2006;9:73–85.

7. Linden A, Yarnold PR, Nallomothu BK. Using machine learning to model dose-response relationships. *J Eval Clin Pract*. 2016;22(6):860–867.

8. Yarnold PR, Linden A. Using machine learning to model dose–response relationships via ODA: eliminating response variable baseline variation by ipsative standardization. *Opt Data Anal*. 2016;5:41–52.

9. Linden A, Yarnold PR. Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *J Eval Clin Pract*. 2016;22(6):875–885.

10. Linden A, Yarnold PR. Using machine learning to assess covariate balance in matching studies. *J Eval Clin Pract*. 2016;22(6):848–854.

11. Linden A, Yarnold PR. Combining machine learning and matching techniques to improve causal inference in program evaluation. *J Eval Clin Pract*. 2016;22(6):868–874.

12. Linden A, Adams JL. Using propensity score-based weighting in the evaluation of health management programme effectiveness. *J Eval Clin Pract*. 2010;16:175–179.

13. Linden A, Adams JL. Evaluating health management programmes over time: application of propensity score-based weighting to longitudinal data. *J Eval Clin Pract*. 2010;16:180–185.

14. Linden A, Adams JL. Applying a propensity-score based weighting model to interrupted time series data: improving causal inference in program evaluation. *J Eval Clin Pract*. 2011;17:1231–1238.

15. Linden A, Adams JL. Combining the regression-discontinuity design and propensity-score based weighting to improve causal inference in program evaluation. *J Eval Clin Pract*. 2012;18:317–325.

16. Linden A. Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions. *J Eval Clin Pract*. 2014;20:1065–1071.

17. Linden A, Adams J, Roberts N. The generalizability of disease management program results: getting from here to there. *Manage Care Interface*. 2004;17:38–45.

18. Linden A. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *J Eval Clin Pract*. 2006;12:132–139.

19. Yarnold PR, Soltysik RC. *Optimal data analysis: A Guidebook with Software for Windows*. Washington, DC: APA Books;2005.

20. Witten IH, Frank E, Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. San Francisco: Morgan Kaufmann;2011.

21. Linden A, Yarnold PR. Using data mining techniques to characterize participation in observational studies. *J Eval Clin Pract*. 2016;22(6):839–847.

22. Linden A, Adams J. Evaluating disease management program effectiveness: an introduction to instrumental variables. *J Eval Clin Pract*. 2006;12:148–154.

23. Linden A, Karlson KB. Using mediation analysis to identify causal mechanisms in disease management interventions. *Health Serv Outcomes Res Methodol*. 2013;13:86–108.

24. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to survival analysis. *Dis Manag*. 2004;7:180–190.

25. Linden A, Adams J, Roberts N. Strengthening the case for disease management effectiveness: unhiding the hidden bias. *J Eval Clin Pract*. 2006;12:140–147.