Improving the analysis and modeling of substance use

David A. Gorelick, MD, PhD[1] and **Sterling McPherson, PhD**[2,3]

[1]Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD

[2]Program of Excellence in Addictions Research, Washington State University, Spokane, WA

[3]Providence Medical Research Center, Providence Sacred Heart Medical Center, Spokane, WA, USA

Substance use disorders (addiction; abuse or dependence in DSM-IV terms) are chronic, relapsing conditions. Our current state of knowledge does not allow for accurate prediction of the trajectory of substance use patterns or development of the disorder in the general population, in a small subgroup (e.g. participants in a clinical trial), or in individual patients. At the population level, this limits our ability to efficiently allocate limited clinical, public health, and law enforcement resources to prevent the growth in use and abuse of existing and novel psychoactive substances (NPS). At the clinical trial level, this limits our ability to accurately identify factors influencing treatment outcome in subgroups of patients. At the patient level, this limits the clinician's ability to identify and promptly intervene during periods of heightened risk of relapse. Four articles in this issue of *The American Journal of Drug and Alcohol Abuse* by Wagner et al. (1), Westover et al. (2), Wakeland et al. (3) and Stogner (4) describe statistical approaches that may improve the current state of the art, ranging from how to handle the raw data that is collected in order to extract the maximum information (in terms of level of aggregation and the best outcome distribution for modeling) (1) to the optimal model to use when evaluating (retrospectively or prospectively) substance use outcomes, whether in a circumscribed group of users (e.g. clinical trial participants) (2) or in the population as a whole (3,4). Overall, the examples in these papers show how sensitive the results of statistical analysis can be to the choice of data aggregation and modeling, highlighting the importance of using the best possible statistical approaches.

Wagner et al. (1) make several recommendations for improved handling of data on self-reported substance use to extract the maximum information. They illustrate their recommendations with data collected using the timeline follow-back (TLFB) method (5), a semi-structured, calendar-based interview widely used in clinical studies. Their recommendations address two common issues with such data: treatment of substance use as a single variable, ignoring which specific substance(s) was used, and which mathematical

distribution best approximates the actual data, especially when data are missing or not collected over the same time interval from all individuals.

In their example, Wagner et al. (1) convincingly demonstrate what is lost when use of any and all substances is aggregated into a single variable. Disaggregation of substance use data improved the findings in two ways. First, it revealed significantly different predictors associated with use patterns of the various individual abused substances (in part because use data for different substances follow different distributions [see below]). Second, it improved the effectiveness of "sensitivity analyses" (i.e. analyzing the same dataset multiple times under different assumptions about the dataset to evaluate how much the resulting outcomes vary with different assumptions) in evaluating the validity of statistical results. Given the promise and increased information resulting from disaggregation, we recommend also applying it, when possible, to the quantification of use, i.e. by analyzing use of each substance as a quantitative, rather than dichotomous (used/abstained) variable, e.g. "number of cigarettes smoked" or "number of drinks consumed". Such quantification should significantly increase the accuracy and precision of the treatment effect estimate.

Wagner et al. (1) also point out that data on substance use, whether dichotomous (used/abstained) or continuous (e.g. counts of use), are often not normally distributed, e.g. because of the frequency of zero values. They recommend a 4-step approach to selecting the best frequency distribution to assume for model-fitting to such data: (i) identify those frequency distributions that are theoretically appropriate based on knowledge of the outcome's characteristics, (ii) visually inspect the outcome data and compare the fit of possible distributions to the outcome data using both (iii) a standard statistical model selection criterion (such as Akaike Information Criterion [AIC]) (6,7) and (iv) visual confirmation (e.g. by superimposing the model based on each distribution onto a histogram of the actual data distribution).

In their example, Wagner et al. (1) again convincingly demonstrate how their 4-step approach efficiently identified the best outcome distribution for effectively modeling the outcome, allowing identification of significant clinical associations that would have been missed using other distributions. Put another way, the validity of study findings was highly sensitive to the model used for analysis. We support their recommendation, and have some suggestions of our own to improve its utility. Model selection would be better informed by using multiple global "fit" indices, rather than only one. We suggest that use of additional criteria would improve decision-making, e.g. the Bayesian Information Criterion (BIC) (8,9) and the Vuong (10,11) test (used for identifying whether controlling for a high frequency of zero values is appropriate). Most statistical software packages (e.g. R, SAS®, and Stata®) can readily calculate such tests. Stata's® "countfit" program automates the process and produces tables for quick comparison across several frequency distributions (e.g. Poisson, negative binomial, and their zero-inflated counterparts) of their model fit using several fit indices (e.g. AIC and BIC), thereby helping identify which model might be best for use in statistical analysis (12).

While not discussed by the authors, we also believe it would be useful to apply their 4-step procedure when addressing the longitudinal (i.e. repeated measures) aspect of most clinical

datasets by use of longitudinal modeling techniques. This allows for a more flexible, and presumably more accurate, representation of the data, resulting in findings with greater validity. For example, a "piecewise growth model" can break up change over time into distinct segments (13,14). This is relevant because many longitudinal data sets show different patterns of change over time during different time intervals. For example, a clinical trial may show significantly more change during the initial treatment period than later on; uptake of a NPS may be much more rapid when it initially appears on the market than later on. Another useful model might be the two-part or semi-continuous model (15,16). This model breaks up the data frequency distribution into two distinct pieces: a binary component (i.e. was alcohol used at all, yes/no?) and a continuous component (i.e. if yes to drinking, how many drinks were consumed?), allowing for a detailed portrait of use.

Westover et al. (2) present an innovative, systematic, stepwise approach for conducting sub-group analyses on a clinical dataset. Typical currently used approaches to evaluating effects in specific subgroups of an overall study population include stepwise selection procedures (e.g. stepwise regression) or testing each subgroup separately. Limitations of the former approach include evaluation of only a small proportion of possible hypotheses (models), use of a biased sequential model selection process, and use of hypothesis-testing methods that do not control for effects of model misspecification or experiment-wise error rates. The latter approach has two limitations: inflating the probability of false positive findings (type I error) because of multiple comparisons and masking true findings (type II error) that may be apparent only in combinations of subgroups.

Building on the concept of a weighted single-indicator moderator score (17), Westover et al. (2) propose a "maximum likelihood type" search process that they term "Best Approximating Model" (BAM). BAM can encompass literally millions of candidate models (with numerous covariates [predictors] and moderator variables) and identifies the "best" model based on a sequence of several specified statistical criteria, including "goodness of fit" measures such as the Generalized AIC. BAM can be considered a highly sophisticated version of stepwise regression which minimizes problems from data multi-collinearity, model misspecification, and multiple comparisons (i.e. family-wise error).

Westover et al. (2) applied their BAM approach to data from a controlled clinical trial of long-acting oral methyl-phenidate as treatment for nicotine dependence (tobacco use disorder) in adults with attention deficit hyperactivity disorder (ADHD) (18). The original investigators found no overall significant effect of medication on cigarette smoking; using conventional subgroup analysis, they found a significant treatment effect associated with ADHD symptom severity and subtype and a trend toward association with lifetime history of substance use disorder. The BAM approach replicated these findings and identified two more treatment associations – with patient age (younger patients did better) and across clinic sites. Thus, use of the BAM approach for this study yielded additional clinically applicable insight into factors associated with treatment response to methylphenidate for smoking cessation in this patient population.

We agree on the promise of the BAM approach, but have some concern about an issue it does not address in its current form. While the final model selected is based on robust model

comparison indices, one factor not considered is how much better the final model is compared to the second-best model. Given the massive number of potential models considered, could the final selected model be just one of a large group of approximately comparable models, making the final selection somewhat arbitrary? We suggest consideration of expanding BAM to include the factor of "clinical utility," i.e. how likely is the generated model to influence (change) current clinical practice? In more specific terms, does incorporation of the proposed risk factor or treatment moderator produce a clinically meaningful change in the outcome variable?

One method of integrating clinical utility into the choice among several comparable, well-fitting statistical models might be assessment of three related indices to guide model selection: area under the curve (AUC) of the receiver operating characteristic (ROC) curve, integrated discrimination improvement (IDI), and net reclassification improvement (NRI) (19,20). Each of these indices provides a distinct measure of improvement in model prediction. Change in AUC when a new predictor (covariate) is introduced into the model (compared to the model without the predictor) may be considered comparable to an improvement in "variance accounted for", but does not specifically quantify and test the difference between models nor indicate whether the new model reclassifies subjects into a different category. IDI is a measure of the difference in discrimination slopes between two models in terms of discrimination slope, i.e. the difference between the mean predicted probabilities of an event for subjects actually experiencing the event and the corresponding mean for those not experiencing the event (19,21). NRI is a measure of the relative increase in correctly predicted probabilities (e.g. model predicted substance use yes/no) for participants in the sample. For example, a model that shows an AUC improvement from 0.75–0.76 or that reclassifies an additional 5 out of 1000 patients into their correct category represents only minimal improvement that likely does not warrant changing clinical practice. Minimal criteria have been proposed (e.g. 3–5% increase in IDI), based on empirical simulation (19,20), for judging the clinical importance of a risk/protective factor. Use of these indices does not involve conventional null hypothesis testing, but rather is a comparison among potential models.

ROC curve values are presented by Westover et al. (2) for the final model that was selected, but were not utilized as model selection criteria. Calculation of model differences based on AUC, IDI, or NRI is readily done in several available commercial software packages. We recommend that the assessment of clinically meaningful improvement become a routine part of any model comparison and selection approach.

Wakeland et al. (3) employed a "closed-loop" dynamic system model to better understand the development of a significant current public health problem in the US population, i.e., misuse of prescription opioid analgesics. Their comprehensive model has several advantages over typical approaches to modeling large-scale substance use problems. First, it incorporates all phases of the development of substance use and misuse, starting with entry processes such as peer influence on initiation of use, substance availability, transitions from informal sharing to paying for drugs, from oral to injecting administration, and to a more potent illegal drug (heroin, in this example) by some users; and exit processes such as quitting use and death. Second, it uses both qualitative and quantitative data. Third, the

dynamic systems feature allows modeling (simulation) of feedback effects in response to changes (single or multiple) in key transition points, e.g. how might strengthening of anti-diversion measures (such as tamper-proof formulations) influence the transition from oral to injecting use?

Wakeland et al. (3) applied their model to several large data sets drawn from nationally representative surveys and published studies. The model accurately retrospectively replicated historical patterns of prescription opioid analgesic use in various population subgroups (to within 8–12% of actual values), including important outcomes such as overdose deaths. More importantly, the model simulations generated predictions about interventions that might significantly reduce prescription opioid analgesic use and adverse consequences, some already being implemented (e.g. tamper-proof formulations) and some worthy of future empirical evaluation (e.g. reducing informal drug sharing with drug buy-back programs).

We believe that this dynamic simulation modeling approach offers major advantages over more conventional approaches. It helps identify gaps in the existing literature, e.g. where there are weak or no estimates for key parameters being utilized in the model. More importantly, it offers the ability to conduct a virtually unlimited number of "thought experiments" that evaluate possible public health and public policy changes (both singly and in combination), thereby better informing clinicians and policy makers. The ability to conduct sensitivity analyses that provide upper and lower bounds for predicted effects makes any model prediction more plausible and therefore more useful to policy makers. For example, Wakeland et al. (3) generate upper and lower limits for the number of overdose deaths caused by prescription opioid analgesics after (theoretical) implementation of two public health measures. The lower limit was half the existing number of such deaths, convincingly suggesting the benefit of these two measures.

Despite the enormous promise of the dynamic simulation modeling approach, we must point out a key weakness of this methodology, i.e., it is completely dependent on the assumptions made. Because it is a closed feedback loop system, even small errors in some putative estimates could generate drastically unrealistic predictions. Also, as with virtually all statistical models, accuracy depends on the degree to which all relevant covariates (predictors) are included in the model. These weaknesses are common in other statistical domains (e.g. power analyses often rest heavily on certain assumptions) and do not diminish our enthusiasm for this approach. They are important for users and their audience to keep in mind when judging the resultant models.

Stogner (4) proposes a 5-step algorithm for predicting whether a novel psychoactive substance will develop into a public health concern. The algorithm comprises five factors: availability of a user base, costs of substance use (price, legal, health, etc.) relative to other available psychoactive substances, quality of the subjective intoxication experience, potential for addiction, and ease of acquiring the substance. A major strength of the approach is comparing the NPS to alternative abusable drugs already in the marketplace. If demonstrated to be accurate, such a predictive algorithm should greatly improve the allocation of finite clinical, public health, and law enforcement resources, resulting in

improved early detection, diagnosis, and treatment efforts. In addition, it would direct public (and political) attention to where it is most appropriate and away from unjustified concerns, thereby reinforcing the more appropriate allocation of societal resources.

Stogner (4) validates his algorithm by applying it retroactively to three past NPS (synthetic cannabinoids, *Salvia divinorum*, and crystal methamphetamine). We agree with his assessment that the algorithm was clearly accurate with regard to the first NPS and arguably fairly accurate with regard to the latter two. To his credit, Stogner (4) argues that the true test of his algorithm is the accuracy of its future predictions; thus, he generates predictions regarding three recent NPS: the synthetic hallucinogen "bromo-dragonfly," the potent opiate agonist acetyl fentanyl, and electronic cigarettes (e-cigarettes). In the 11 months since this manuscript was first submitted for publication, we interpret the initial data as clearly confirming the prediction about e-cigarettes (a growing public health concern) and at least partially confirming the prediction that acetyl fentanyl "warrants attention" (it was reclassified into CSA schedule I by the DEA on 21 May 2015). We consider the jury still out on bromo-dragonfly.

The framework provided in this article is a good starting point, but an apparent weakness is lack of specificity. The current method is qualitative, making it difficult to demonstrate utility to clinicians, policy makers, law enforcement officials, and others. We suggest formalizing the process, similar in rationale to conducting a meta-analysis (22). This should include incorporating comprehensive, quantitative data from existing "early warning" data collection networks (such as the Community Epidemiology Work Group [CEWG] organized by NIDA, Researched Abuse, Diversion and Addiction-Related Surveillance [RADARS®] System operated by Denver Health, National Poison Data System operated by the American Association of Poison Control Centers, and US DEA System to Retrieve Information from Drug Evidence [STRIDE*]*) and specifying what variables will be used and how they will be coded. Overall, the Stogner (4) algorithm might benefit from incorporating many of the statistical modeling features employed by Wakeland et al. (3).

Dynamic simulation modeling would be one such useful feature, a kind of "sensitivity" analysis, wherein different parameters are modified and the prediction changed in order to run several "what if" scenarios based on scenarios being discussed in the media. For example, one could assume a 50% increase in hospitalizations due to a new NPS and see how this influences 5-year cost projections. This would allow for calculation of confidence bounds around the projected estimate, a measure of prediction accuracy. One could imagine a sort of "moving" projection based on newly acquired data as it becomes available. This might be an important tool for avoiding the bias that is all too common in these types of forecasts, especially in the face of popular media reporting.

We commend the authors of these four papers for addressing the seemingly mundane, technical, and often overlooked issue of analysis and modeling of substance use data. Their useful recommendations do not require extensive new resources to implement, but, if refined and widely adopted, would significantly improve the interpretation of existing and future clinical and epidemiological research data, to the betterment of patients with substance use disorders and society at large.

References

1. Wagner B, Riggs P, Mikulich-Gilbertson S. The importance of distribution-choice in modeling substance use data: a comparison of negative binomial, beta binomial, and zero-inflated distributions. Am J Drug Alcohol Abuse. 2015; 41:489–497. [PubMed: 26154448]

2. Westover AN, Kashner TM, Winhusen T, Golden RM, Nakonezny PA, Adinoff B, Henley SS. A systematic approach to subgroup analyses in a smoking cessation trial. Am J Drug Alcohol Abuse. 2015; 41:498–507. [PubMed: 26065433]

3. Wakeland W, Nielsen A, Geissert P. Dynamic model of nonmedical opioid use trajectories and potential policy interventions. Am J Drug Alcohol Abuse. 2015; 41:508–518. [PubMed: 25982491]

4. Stogner JM. Predictions instead of panics: the framework and utility of systematic forecasting of novel psychoactive drug trends. Am J Drug Alcohol Abuse. 2015; 41:519–526. [PubMed: 25774440]

5. Sobell, LC.; Sobell, MB. Timeline followback user's guide: a calendar method for assessing alcohol and drug use. Addiction Research Foundation; Toronto: 1996.

6. Akaike H. Likelihood of a model and information criteria. J Econometrics. 1981; 16:3–14.

7. Bozdogan H. Akaike's information criterion and recent developments in information complexity. J Mathemat Psychol. 2000; 44:62–91.

8. Henson JM, Reise SP, Kim KH. Detecting mixtures from structuralmodel differences using latent variable mixture modeling: a comparison of relative model fit statistics. Struct Equat Modeling. 2007; 14:202–226.

9. Schwarz G. Estimating the dimension of a model. Annals Statistics. 1978; 6:461–464.

10. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica. 1989; 57:307–333.

11. Coulter RW, Blosnich JR, Bukowski LA, Herrick AL, Siconolfi DE, Stall RD. Differences in alcohol use and alcohol-related problems between transgender- and nontransgender-identified young adults. Drug Alcohol Depend. 2015; 154:251–259. [PubMed: 26210734]

12. Scott, LJ.; Freese, J. Regression models for categorical dependent variables using Stata. StataCorp LP; College Station, TX: 2006.

13. Li F, Duncan TE, Duncan SC, Hops H. Piecewise growth mixture modeling of adolescent alcohol use data. Struct Equat Modeling. 2001; 8:175–204.

14. Mamey MR, Barbosa-Leiker C, McPherson S, Burns GL, Parks CD, Roll J. An application of analyzing the trajectories of two disorders: a parallel piecewise growth model of substance use and attention deficit/hyperactivity disorder. Experim Clin Psychopharmacol. Forthcoming.

15. Olsen MK, Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. J Am Stat Assoc. 2001; 96:730–745.

16. McPherson S, Barbosa-Leiker C. An example of a two-part latent growth curve model for semicontinuous outcomes in the health sciences. J Appl Stat. 2012; 39:2113–2128.

17. Wallace ML, Frank E, Kraemer HC. A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. JAMA Psychiatry. 2013; 70:1241–1247. [PubMed: 24048258]

18. Winhusen TM, Somoza EC, Brigham GS, Liu DS, Green CA, Covey LS, Croghan IT, et al. Impact of attention-deficit/hyper-activity disorder (ADHD) treatment on smoking cessation intervention in ADHD smokers: a randomized, double-blind, placebo-controlled trial. J Clin Psychiatry. 2010; 71:1680–1688. [PubMed: 20492837]

19. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008; 27:157–172. discussion 207–112. [PubMed: 17569110]

20. Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P. Interpreting incremental value of markers added to risk prediction models. Am J Epidemiol. 2012; 176:473–481. [PubMed: 22875755]

21. Yates JF. External correspondence: decompositions of the mean probability score. Organizat Behav Human Perform. 1982; 30:132–156.

22. Lipsey, MW.; Wilson, DB. Practical meta-analysis. Sage Publications; Thousand Oaks(CA): 2001.