



III. GOOD RESEARCH PRACTICES FOR COMPARATIVE EFFECTIVENESS RESEARCH: ANALYTIC METHODS TO IMPROVE CAUSAL INFERENCE FROM NON-RANDOMIZED STUDIES OF TREATMENT EFFECTS USING SECONDARY DATA SOURCES - Report of the ISPOR Retrospective Database Analysis Task Force – Part III

Comments received from Reviewer/Leadership/ISPOR membership :

### Respondent #1

Thank you for the opportunity to review and comment on the documents pertaining to good research practices for non-randomized studies of treatment effects using secondary databases.

This is very important work and the authors have put together very thoughtful and useful information. I will comment on the documents generally and then each document separately.

#### Overall comments

It would benefit the audience if the authors could first define their key terms such as “effectiveness research” “comparative effectiveness research” as “secondary data sources.” “observational data” “epidemiologic studies” and to use these terms and definitions consistently across the three documents. Providing common definitions would also be extremely helpful to the field. It’s not clear whether these series is limited to “comparative effectiveness” or “effectiveness.” The authors seem to use “observational research” and “epidemiologic research” interchangeably.

In discussing the relative merits of observational studies versus RTCs, the authors should acknowledge that most evidence hierarchies consider RTCs the gold standard of evidence and rank case control and cohort studies further down the evidence later. The main reason for this is that unmeasured or unobservable confounding brings into question the validity of observational studies. This does not mean that observational studies are not useful, it just acknowledges how they are currently being viewed and why. The role of unobservable confounding and the limited clinical information to control for confounding should be acknowledged early on in the documents.

### Respondent #2

#### **III. Comments Specific to Good Research Practices for Analyses of Non-Randomized Studies of Treatment Effects Using Secondary Databases.**

Prior to discussing propensity score matching it would be useful to discuss “direct matching” where case and control or exposed and unexposed are directly matched on characteristics rather than a propensity score.

Page 285: The statement here is controversial. Case and controls may not always be matched but the response of throwing out those cases may limit the external validity of the study. For example, if the sickest patients who responded least well to a therapy are excluded from the analysis, the study may erroneously conclude that the medication was effective. This issue may deserve more discussion Page 299. It’s my understanding that limited research to date has focused on how to select variables for inclusion into a propensity score and that the recommendations are evolving. This discussion could be made more tentative rather since it does not seem entirely clear that how to select variables for inclusion in the propensity score model.

### Respondent #3

Overall- it would be nice to have a short commentary/executive summary talking about all of the papers together and outlining the recommendations from them.

### Paper 3

Quote Rothman and Greenland 1998- there was a new edition out this year.

P5 & P9 - there is no mention on how to treat missing observations. This is especially important when constructing propensity scores with a lot of covariates that have missing information. Perhaps a paragraph discussing these methods (from basic to complex) would be useful. To impute or not to impute - is that the question?

In several places in the document GEEs are mentioned (e.g. P6 line 160) as the analysis for categorical outcomes. Why not use the framework of generalized linear mixed models?

P7 line 192 -195. I though it would be good to talk about examining interaction effects here.

P8. Section on diagnostics. It would be good to have some discussion on treating outliers & examining influential observations via sensitivity analysis. It would also be good to get clearer guidelines on what to report in terms of diagnostics - what are the must haves(line 230)

P10 line 280 . mentions excluding patients that are not matched - in some instances this could be quite a lot of patients excluded from the analysis. So the limitations should be discussed

P12 line 350. It would be good to get a bit more guidance on how to assess the overlap e.g. assessing absolute standardized differences before and after matching

P15, line 421 - it might be worth mentioning that MMRM can be used for continuous outcomes as well (easily implemented in SAS with weight statement)

### Respondent #4

My read on the propensity scoring section is that it is being presented as slightly superior to traditional modeling. Given that most of the literature has not shown any real benefit to propensity scores, I would suggest that the proper place for propensity scores is as a sensitivity analysis or a secondary analysis.

Many of the “strengths” of propensity scoring as described are not unique to propensity scoring. In particular the same linearity assumptions and parsimony considerations apply both to propensity score creation and to traditional regression models. There is no reason one can’t apply the same approaches to either model – they are both regression models with similar assumptions and goals. Much of model building is based on personal preferences (e.g. the use of continuous vs. categorical variables, whether to include only significant effects or not, etc.) so I would not draw a distinction between the propensity score and a traditional regression model...

One limitation of propensity scores is that information about the effects of covariates on outcomes (e.g..., age, gender, race, etc) cannot be easily interpreted in the final analysis if these covariates are also in the propensity score. Often times the propensity score is not used this way, but effect modification (interaction) is important to be able to evaluate. While one could, for example, add age to a regression model with a propensity score in it (or stratify), the interpretation of age would be challenging if age were also part of the propensity score (i.e., there would be both direct and indirect effects of age in such a model).

Finally, a traditional regression analysis can provide much more face validity to the model. Particularly for a clinical audience, this is a very important issue in explaining how the model is working.

Thanks for the opportunity to review the document, and for putting such an excellent report together.

### Respondent #5

Overall comments: Excellent job. The authors need to be commended.

Comments on “III. GOOD RESEARCH PRACTICES FOR THE ANALYSIS OF NON-RANDOMIZED STUDIES OF TREATMENT EFFECTS USING SECONDARY DATABASES: Report of the ISPOR Retrospective Database Analysis Task Force – Part III”

Line 787 is missing the first author (McWilliams)

References in all three papers are not uniformly presented in a single, consistent style. Presumably, that will be rectified.

## Respondent #6

Thanks for this important work!

It is very helpful for evaluating routine data. I plan to use it as a sort of checklist so I won't forget important questions. You should take into account that a lot of work with this type of data is never published because it is used for decision making or negotiations in settings which are not very transparent. There it is of high priority to answer questions in time and fitting the system you are working in. Scientific accuracy in most cases is not so important because results are not published. Publication only occurs in transparent settings or if some people think it supporting for their career. So it would be helpful to have these reports published so they can be referred to.

## Respondent #7

III. Good Research Practices for the Analysis of Non-randomized Studies of Treatment Effects Using Secondary Databases  
Illustrating explanation of regression techniques, needs to clarify the "structural equation model" section, basically to define better the model of "drug selection, patient adherence & outcome". May include multi-level analysis as a technique for results assessment.

## Respondent #8

After reading the articles, I think it is a novel study which will help us to utilize the data from non randomized trials and will help to reduce the expenditure.

In my opinion, following points can be useful for the document:

1. To check reliability of the study, grading on the scale of 10 should be done.
2. Unpublished and published case series can be helpful for the preparation of the databases. Pulling data from different case series (may be from same region) can form a bigger sample size.
3. Stratification of case series, for a particular disease can be performed to target a research question.
4. Addition of the drug utilization studies into the task can be helpful to make a database to find out cost effective treatment for a disease.

## Respondent #9

I appreciated the opportunity to review the three working papers from the ISPOR Good Research Practices - Retrospective Database Analysis Task Force. I found all three well written and informative.

My one comment would be to encourage this ISPOR task force to consider **cross trial comparisons** as part of your mandate to "recommend good research practices for non-randomized studies of treatment effects using secondary databases". As you know, cross trial comparisons are routinely conducted by Industry and HTA agencies such as NICE and DERP to explore both issues of comparative and cost-effectiveness. Whilst the original clinical trial may have met the definition of a RCT, when they are aggregated for these cross trial comparisons the database (a collection of clinical trials) is no longer a randomized trial and face many of the same methodological challenges you discuss in your task force papers – as a simple example, the publication may not contain sufficient information to determine that patients are comparable at baseline. Using placebo controlled trials to make head-to-head comparisons certainly qualifies the analysis as secondary.

A white paper that offers guidance on designing and conducting scientifically robust cross trial comparisons would greatly benefit the field.

## Respondent #10

Thank you very much for the opportunity to briefly comment on these excellent reports.

As more and more transnational research is performed, and in the light of increasingly limited resources, a comment on pre-specified country-specific secondary analyses of such studies - if performed in a multinational setting - may be of help for country-specific decision making, however:

- Justification for such secondary country-specific analyses need to be provided a priori
- Meaningful, pre-defined minimum patient numbers from these countries need to be reached (condition sine qua non) and CID should be determined
- Country-specific results need to be put into perspective, i.e., discussed in the light of the results of the entire cohort

## Respondent #11

On document 3, analyzing:

Stratification's benefits extend beyond epidemiology; would suggest replacing the term with "observational study".

Missing from the regression section is the notion of "common support". It is not enough to include covariates that may or may not differ on average between two treatment groups. Scatter-plotting must also be done to make sure the distribution of each variable has enough overlap between groups -- if these distributions occupy very different areas of the covariate space, a bias will be imparted to any regression model (Dehejia R, J Economics 2005; Heckman JJ, Rev Econ Stud 1997)

## Respondent #12

I'm getting in on this at the tail end having just rejoined ISPOR after several years. I was wondering if mixture models were given any consideration for Part III? I have over thirty years using them, especially for observational or real world data, and have found these models to offer valued clinical insight as to what really happened in the study.

## Respondent #13

I glanced through the report that you and your colleagues have put together. I think it is definitely a very important issue and will have a substantial impact.

The third report very nicely lays out various techniques. However, I think most of them are applicable for cross-sectional design. What are some methods that can be applied to longitudinal panel or time series data? I think, may be, there should be also some examples of clinical data sets that researchers can use, such as the SEERs.

## Respondent #14

If I may add a few comments for these great papers, I would like to point out to add a section how to integrate patient reports. I think patients' self-reported dataset which will fill in the bias of secondary datasets, mostly likely to be claims based information.

I think the third paper is excellent to summarize all popular methods in this field. However, in reality, the best strategy is to use the simplest method, i.e. OLS.

I hope this comment helps.

## Respondent #15

Thank you for considering me to make comments on these documents. In general, I think that they are well written and understandable. I just make few comments and I hope they can be useful.

Comments to the following documents:

Good research practices for the analysis of non-randomized studies of treatment effects using secondary data bases.

In general, to what extent, would it be useful to consider the issue of the missing values in the statistical analysis of the information?

Could it (missing values) represent an important issue to consider in the outcomes results?

## Respondent #16

### III. GOOD RESEARCH PRACTICES FOR THE ANALYSIS OF NON-RANDOMIZED STUDIES OF TREATMENT EFFECTS USING SECONDARY DATABASES: Report of the ISPOR Retrospective Database Analysis Task Force – Part III

Line 42 I was of the opinion that stratification is performed on the basis of observables, thus confounding attributable to unobservables or unmeasured X'tics due to the heterogeneous nature of subjects may still plague the analysis.

Line 56 The report might be enhanced by clarifying that stratification (as used in epidemiology) and multivariate analysis are two separate techniques. The latter involves only one independent variable. In econometrics, stratification is just categorizing or grouping the subjects in the data by some observable X'tics and then estimating your model by one of many multivariate approaches.

Line 63 The assertion that dissimilar results from a stratified analysis and a rigorous multivariate model points to a mistake on the part of the analyst is simply inaccurate. One may arrive at markedly different estimated coefficients by stratifying the model vis-à-vis performing an unstratified analysis. A casual look at Line 53 (reduced precision, loss of information, etc.) as well as unobserved heterogeneity may all ensure that the analysis from the stratified and unstratified approaches are markedly different. For instance, if one relies on panel data methods using fixed effects or linear mixed models that controls for time-invariant unobserved heterogeneity the estimated coefficients would be different from that of a model that has been stratified (in the epidemiological sense).

Line 66 I believe the covariates to be included in a model should be informed by theory and previous literature

Line 78-83 I believe an appropriate approach would be to include interaction terms (X1X2) in the model to be estimated if one believes that the effects of X1 on Y is modified by X2. Insights could also be gleaned from theory or prior literature.

Line 88 The Mantel-Haenszel estimators are accurate only to the extent that the assumptions inherent in its use hold. The Mantel-Haenszel assumes that either exposure or disease is measured on an ordinal (or interval) scale, when you have more than 2 levels. Also, the Mantel-Haenszel proceeds on the assumption that there are no confounders (it cannot control for the effects of confounding factors).

Line 133 It appears model selection normally refers to tools such as AIC, BIC, etc, for selecting between a restricted and unrestricted version of the same econometrics or statistical estimation method. Line 206 The dataset needs not be large, and one needs not assert that the parameter values are questionable. The reason for the low R-squared may stem from the nature of the epidemiological data (longitudinal or panel). That is, it is a combination of cross-sectional and time-series datasets, and the R-squared reported from an analysis using this type of data is rather "cross sectional-like". For instance, if one compares person A's outcomes with his/her own outcomes at different times (time-series) one can certainly explain much of the variation with just a few variables. However, if one compares person A's outcomes with person B's outcomes (cross-sectional) those same few variables will explain less, if any, of the variation. Thus, a cause for concern on a low R-squared depends on the type of dataset employed and the model.

Line 221-230 This is perhaps one of the most important parts of the statistical analysis and should be given additional attention. Nine "lines" may not be adequate to capture the magnitude of the issues associated with model diagnostics and the extent to which it's been ignored in industry publications. Also, a commentary on a few diagnostics tests as well as issues related model specification and identification could enhance the report.

Line 333-335 I believe differences between propensity score methods and regression techniques estimates would depend on the type of regression technique employed (OLS and logistic regressions are but two of the numerous regression techniques available to the analyst)

Line 342-343 I believe not all regression analysis imposes linearity assumption.

Line 344-346 This may not be quite accurate. The report would be enhanced if it were to outline the type of regression analysis that fails to admit more covariates, interaction and quadratic terms. The report may refer the reader to applied microeconomics texts such as Greene, W.H. (1997), *Econometric Analysis*, (3rd Edition), Maxwell-MacMillan: New York; Wooldridge, J. (2002) *Econometric Analysis of Cross-Section and Panel Data*, Cambridge: MIT Press; Amemiya, T. (1985), *Advanced Econometrics*, Basil Blackwell; Hsiao, C. (1986), *Analysis of Panel Data*, Econometric Society Monographs, Cambridge; Baltagi, B.H. (1995), *Econometric Analysis of Panel Data*, Wiley.

Line 374-420 Clarification on time-dependent confounders should be clarified. For example, in a regression of the probability of Ivy school admission for a high school graduate [ $pr(Y)$ ], unmeasured parent's education (Z) may influence Ivy school admission (Y), students' test scores (X1) and even the quality of high school attended (X2). However, parents' education is not time dependent – here Z (which the report refers to as time-dependent confounders) affects Y, X1 and X2 Line 421 Transitions to different methods in the report seems quite abrupt at times Line 456 Not quite! The IV approach relies on finding at least one variable that is correlated with the endogenous variable but uncorrelated with the error term (see line 482) and independent of the outcome.

Line 462-475 It might be advisable to test indirectly whether the RHS variable of interest is indeed endogenous, using the Wu-Hausman test before implementing an IV approach. Indeed In instrumental variables approach to correct for endogeneity, there are no provisions for the inverse mills ratio or the inclusion of the IMR as additional regressor in the model (please refer the reader to the suggested texts above). The two-step method being described in the report is rather the Heckman's selection correction model used to correct for sample selection bias. Sample selection bias, and endogeneity bias refer to two distinct concepts, requiring distinct solutions. The former occurs where the dependent variable is observed only for a restricted, nonrandom sample while the latter refers to an independent variable included in the model being potentially a choice variable, correlated with unobservables relegated to the error term; the dependent variable is however observed for all observations in the data. Thus, the two-step approach described here may not be the appropriate approach to correct for endogeneity. It might be advisable to the analyst to use the Sargan Test of over-identifying restrictions to ascertain if the model is over-identified.

Line 500-505 This assertion might be questionable. Recall, OLS rest in part on the assumption that  $E(u|X) = 0$  but this "orthogonality" assumption is not likely to be satisfied. Indeed, IV is consistent whether or not the regressors and disturbance are correlated. Inconsistent estimates has implications for inference. Also, pooling cross-sections across time (used extensively in epidemiology) and estimating the model by OLS may yield incorrect standard errors (OLS assumes that the error term is IID across individuals and over time and may lead to biased and inconsistent estimates of the coefficients).

Line 511-512 IV is not necessarily a test for endogeneity. IV methods allow for consistent estimation when the explanatory variables (covariates) are correlated with the error terms. One cannot test endogeneity or exogeneity directly. However the Wu-Hausman test can be used to test the exogeneity assumption indirectly.

Overall, the report could benefit from some proofing (grammar, punctuation, etc.) Also, the report seems to dwell on logistic regression, OLS and Cox PH to the exclusion of other estimation techniques within the same 'class' of estimators. The report would have been enhanced if commentary/attention were given to other econometrics/statistical approaches. For instance, within limited dependent variable models, we have ordered logit and probit, multinomial logit and probit, conditional logit and probit, tobit, etc. Within survival analysis or duration models, there are non-parametric (K-M), semi-parametric (Cox PH) and parametric approaches (Weibull, Gompertz, gamma, log-logistic, etc). For continuous dependent variables (dynamic panel data methods using GMM, FGLS, WLS, SUR, Hierarchical or mixed models, GLM, etc.). A commentary on model specification, identification and types of tests to aid model selection (LR test, AIC, BIC, etc.) could enhance the report.

## Respondent #17

Please find comments as requested below:

### III. Good Research Practices for the Analysis of Non-randomized Studies of Treatment Effects Using Secondary Databases

Clear prioritisation of methods and guidance of 'best practice'. Acknowledging the need to reporting the, for example, coefficient of determination (R<sup>2</sup>) (line 208) is an important point. Propensity scoring gives the reader a good insight into

addressing selection bias. The paper provides a good overview and adequately answer its objective.

I have read these three papers as a consumer looking for guidance and have found them very useful. I hope these comments will in some way be help.

### Respondent #18

We have reviewed all three reports sent by you. Our comments for these three Task Force reports are as follows  
Comments for report 3:

In this study authors have given more focused information on modern methods of statistical tools and its implications in analytical studies. Discussion under the heading and subheadings of "Stratification", "Variable selection", "model selection" is very helpful and informative. I suggest some more points should be discussed under performance measurement and diagnostic tools. A case study is needed to support the study.

### Respondent #19

Overall- it would be nice to have a short commentary/executive summary talking about all of the papers together and outlining the recommendations from them.

#### Paper 3

Quote Rothman and Greenland 1998- there was a new edition out this year.

P5 & P9 - there is no mention on how to treat missing observations. This is especially important when constructing propensity scores with a lot of covariates that have missing information. Perhaps a paragraph discussing these methods (from basic to complex) would be useful. To impute or not to impute - is that the question?

In several places in the document GEEs are mentioned (e.g. P6 line 160) as the analysis for categorical outcomes. Why not use the framework of generalized linear mixed models?

P7 line 192 -195. I thought it would be good to talk about examining interaction effects here.

P8. Section on diagnostics. It would be good to have some discussion on treating outliers & examining influential observations via sensitivity analysis. It would also be good to get clearer guidelines on what to report in terms of diagnostics - what are the "must haves" (line 230)

P10 line 280 mentions excluding patients that are not matched - in some instances this could be quite a lot of patients excluded from the analysis. So the limitations should be discussed

P12 line 350. It would be good to get a bit more guidance on how to assess the overlap e.g. assessing absolute standardized differences before and after matching

P15, line 421 - it might be worth mentioning that MMRM can be used for continuous outcomes as well (easily implemented in SAS with weight statement)

### Respondent #20

#### III. Good Research Practices for the Analysis of Non-randomized Studies of Treatment Effects Using Secondary Databases

Line 24: Full stop (.) after the word confounding.

Line 34: Put the year after the authors in the references.

Line 129-132: It would be helpful if references for the techniques that the authors describe are provided here.

Line 141: "GLM Models" may be replaced with "GLMs".

Line 146: I am not sure comment 1 is exactly correct. Health care costs and utilization variables are often very skewed and therefore the logarithms of those variables are often modeled if one opts for OLS paradigm. However, since estimated costs or utilization from those models are in logs, they need to be transformed their original scale (eg. dollar as opposed to log dollar). Now, since GLM framework models the outcome variable (e.g., cost or utilization) in its own scale, "retransformation" problem is obviated. This is true for all GLM models with any error distribution (as opposed to what is stated in comment (i)). *Also, it might be helpful if a sentence stating that Modified Park Test needs to be carried out to determine the appropriate error distribution in a GLM model.*

Line 167: "know" should be replaced by "known".

Line 179-180: Please note that for logistic model, the error term is assumed to be extreme value distributed but not normally distributed.

Line 206: Suggest that a discussion about goodness of model fit for GLM models be included this section.

Line 232: Replace "increasing" by "increasingly"

Line 251:  $E(x_i)$  should be  $E(Z_i|X_i)$ .