Polydesigns in Causal Inference

Fan Li, Constantine E. Frangakis, and Ravi Varadhan
Department of Biostatistics, Johns Hopkins University
January 4, 2004

Abstract. In an increasingly common class of studies, the goal is to evaluate causal effects of treatments that are only partially controlled by the investigator. In such studies there are two conflicting features: (1) a model on the full design and data can identify the causal effects of interest, but the model's use in extreme regions of the data (e.g., where the outcome of interest is rare) can be sensitive to model misspecification; and (2) the induced model on a reduced design, i.e., of a subset of data (e.g., conditional likelihood on matched pairs) can be more robust to a full model's misspecification, but it does not generally identify the causal effects. We propose to assess inference sensitivity to designs by exploring combinations of both the full and reduced designs. We show that using such a "polydesign" generates a rich class of methods that can identify the causal effect and that can also be more robust to misspecification than the full model and design. We also discuss implementation of polydesign inference.

Key Words: anchor function; causal effects; needle exchange program; partially controlled studies; polydesign; principal stratification.

1. Introduction

A frequent goal is to evaluate causal effects of treatments in studies that can only be partially controlled by the investigator. Such partially controlled studies have two conflicting features: (1) a model on the full design of data can identify the causal effect of interest, but using the model to extreme regions of the data (e.g., when the outcome of interest is rare) can be sensitive to model misspecification; (2) the model induced to a reduced design, i.e., a rule for selecting a subset, of data (e.g., a conditional likelihood on matched subsets of data), which can be more robust to a full model's misspecification, does not identify the causal effects.

To make analogies to a motivating example of a partially controlled study, we consider the Baltimore's Needle Exchange Program (NEP), as originating from the combination of the ALIVE and NEP studies (Vlahov et al., 1997; Strathdee et al., 1999). In the NEP study, a cohort of injection drug users (IDUs) has been enrolled and is being followed, with regular six month (a semester) visits in which the subjects are offered clinic services, including blood tests for HIV. In parallel to the clinic, the NEP study operates sites in Baltimore where IDUs can exchange a used needle for a sterile one, with the hope of reducing HIV transmission. The goal in the NEP is to evaluate the degree to which exchanging versus not exchanging needles reduces (if at all) HIV incidence among IDUs.

For evaluating the goal in the NEP, it is important to recognize that the study is partially controlled in the sense that (a) it controls neither who exchanges nor who provides outcome HIV blood tests at the clinic, but (b) it controls the placement of the NEP sites offering needles, hence the distance of the NEP sites from the IDUs. If distance of the NEP sites to the subjects influences who exchanges at the NEP and who provides HIV blood tests, this relation can be useful to evaluate the effect that distance has on HIV incidence and that is attributable to exchanging needles. However, general results for such studies (Frangakis and Rubin, 1999) imply that it is not appropriate to use standard evaluation of the NEP, e.g., by comparing exchangers with nonexchangers on observed HIV outcomes (e.g., Keende et al., 1993; van Ameijden et al. 1994; Drucker et al., 1998), or by using distance of NEP to IDUs as a standard instrumental variable. In contrast, the framework of "principal stratification (Frangakis and Rubin, 2002) has been shown to allow both definition of more appropriate estimands, and also to allow separation of the assumptions made for the mechanisms of the effects (structural assumptions) from the assumptions made on the design of what data is being collected.

Part(1) of the conflict stated in the first paragraph arises from the necessary relative complexity of the models involved in such partially controlled studies as

This work was the basis of a presentation at the 2003 Joint Statistical Meetings under the sponsorship of the ASA Section of Statistics in Epidemiology and will appear in the 2003 ASA Proceedings volume of that section. We thank Don Rubin for valuable discussions, and the NIH (NEI) grant RO1 EY 014314-01 for partial support.

the NEP. This complexity makes it important to investigate whether the results are robust to the design used. Of particular relevance to alternative designs is often that the outcome of interest is rare. For example, in the NEP study, over an average follow-up of 9 semesters for 1170 subjects, 52 subjects became HIV positive (cases). With such low incidence, results can be sensitive to the model, for example, in the region of controls' covariates that does not overlap with the covariates of the cases. A possible way to address this can be to attempt inference through a semiparametric model (e.g., Robins, Rotnitzky, and Bonetti , 2001, in other settings). However, in problems with such relatively complex structure, current semiparametric formulations cannot guarantee estimability and miss the advantages of efficiency of parametric models which can be justified even from a non-parametric perspective (e.g., Frangakis and Rubin, 2001). For this reason, we choose to explore sensitivity of the results to different designs, which is an approachmoften preferred in epidemiology in other settings. Specifically, we consider designs for selecting a subset of the full data, and which we call here "reduced designs".

Part (2) of the above stated conflict arises because, in partially controlled studies, the information lost in "reduced designs" can be necessary for estimating the causal effects well. Suppose, for example, for a "reduced design" in the NEP, we choose all the HIV cases, and, for each case, we keep one control who closely matches the case on covariates. Then the conditional likelihood induced by the model of the full design on the reduced design, unlike a standard conditional logistic regression likelihood, does not necessarily identify the causal effect, mainly because of a latent structure inherent in the framework of principal stratification of partially controlled studies (Frangakis and Rubin, 2002). The problem we address in partially controlled studies is how to explore results from different designs and in a way that preserves enough information to well estimate the causal effects.

We propose a class of methods that are based on the combination of the full design together with reduced designs. The key of such "polydesign" methods is that they provide a continuum between the full design and reduced designs. The members of this continuum can identify the causal effects and also achieve better robusteness than the full design under model misspecification. Section 2 defines polydesigns, discusses inferential perspectives, and provides main general properties. Section 3 discusses practical implementation of polydesigns, and Section 4 provides remarks.

2. Polydesigns

Consider a design I_{FULL} for selecting units from the population and assigning them to treatments, and let D_{FULL} be the data arising from this design. We call $I_{\text{FULL}}, D_{\text{FULL}}$ the "full" design and data. Let also $\text{pr}_{\text{FULL}}(D_{\text{FULL}} \mid \theta)$ denote the likelihood of the full data based on the full design, where θ represents the parameters describing the population of study, and where the estimand of interest is some function of θ. We consider cases where: θ is identifiable from the likelihood of the full data; where there is concern that inference from the full data can be sensitive to model misspecification; but where parametric modelling is, nevertheless, required for stable estimation, efficiency and interpetation.

A common approach to assessing sensitivity to misspecification has been to consider a "reduced" design I_{REDU}, that is, a design that describes how to select a subset of the units from the full design I_{FULL}, e.g., through a certain matching rule as in case-control studies (Breslow and Day, 1980). Let D_{REDU} denote the "reduced" data, including indicators of which units have been selected, that arise from this design. For the reduced data, we consider a likelihood

$$\text{pr}_{\text{REDU}}(D_{\text{REDU}} \mid \theta) \tag{1}$$

that is induced from the full likelihood when also taking into account the reduced design. We allow that the reduced likelihood can be the marginal likelihood of the reduced data, but also allow that it can be a conditional or a partial likelihood (Cox and Oakes, 1984) of the reduced data, which is the reason for distinguishing between the distributions pr_{FULL} and pr_{REDU}.

The reduced design can avoid sensitivity to mispecification of the full model on those features of the full data that are not modelled in the likelihood (1). Nevertheless, estimation of the causal estimands of interest in frameworks for partially controlled studies, such as principal stratification, generally requires most of the information provided by the full data. This can be formalized in that, in the contingency table of discretized summaries of the observed data, all cell probabilities are generally needed to estimate the causal effect (Frangakis et al. 2003, Section 3), and is reflected by the fact that a reduced likelihood (1) may not be sufficient to identify the causal effect. In fact, identifiability and consistency of estimation in a reduced likelihood holds only for special cases, such as with the standard design and model for a case-control study, which are not applicable in partially controlled studies.

The goal for these cases here is to provide an approach that assesses sensitivity of inference on the parameter θ to different designs. The idea for doing this is that, although a reduced design may not identify the full parameter θ, it may identify some functions of the parameter from (1), and thus can provide robust estimates under misspecification of the model's part that is omitted in that reduced likelihood. This point suggests that it is useful to consider the following definitions.

DEFINITION

(a) An *anchor* function $A(\theta)$ is a function that is identifiable from the reduced data by the reduced likelihood (1).

(b) A *polydesign* with respect to a full design I_{FULL} of a population is a collection of the full design together with reduced designs, $(I_{\text{FULL}}, I_{\text{REDU},1}, I_{\text{REDU},2}..., I_{\text{REDU},k}), k \geq 1$.

In a polydesign, the reduced designs are, by definition, nested within the full design, but do not need to satisfy a nesting structure among each other, and the data they produce can overlap. We focus discussion here on basic polydesigns with one reduced design, as the discussion that follows can be extended easily to the more general case.

The purpose of a polydesign is to synthesize inference of functions of the anchor function of parameters, identifiable from the reduced likelihood, with inference of the remaining parts of the parameter from the full likelihood. The former inference helps reduce sensitivity to misspecification of the full model, whereas the latter, if needed, completes identification of the causal effect. Such synthetic inference can be expressed with a Bayesian or a likelihood perspective. In the following, we assume that standard regularity conditions such as for consistency of parametric maximum likelihood estimation hold when a parameter is identifiable.

Bayesian perspective.

From the full likelihood and a prior distribution, we obtain the full posterior distribution, $\text{pr}(\theta|D_{\text{FULL}})$. For a function $A(\theta)$ of the parameters, consider the decomposition of the full posterior distribution, into the marginal distribution of $A(\theta)$ and the distribution of the remaining functions, by denoted $\theta - A(\theta)$, required to specify θ,

$$\text{pr}_{\text{FULL}}(\theta|D_{\text{FULL}}) = \text{pr}_{\text{FULL}}(\theta - A(\theta) \mid A(\theta), D_{\text{FULL}}) \cdot \text{pr}_{\text{FULL}}(A(\theta)|D_{\text{FULL}}). \quad (2)$$

Also, from the reduced likelihood and the prior distribution, we can obtain the reduced posterior distribution, $\text{pr}_{\text{REDU}}(\theta|D_{\text{REDU}})$, as if we had only observed reduced data D_{REDU}. For an anchor function $A(\theta)$, the resulting reduced posterior distribution $\text{pr}_{\text{REDU}}(A(\theta)|D_{\text{REDU}})$ is not sensitive to specification of the prior because $A(\theta)$ is identifiable from the reduced likelihood. It is then helpful to consider the "polydesign" distribution

$$\text{pr}_{\text{POLY}}(\theta) := \text{pr}_{\text{FULL}}(\theta - A(\theta) \mid A(\theta), D_{\text{FULL}}) \cdot \text{pr}_{\text{REDU}}(A(\theta)|D_{\text{REDU}}). \quad (3)$$

The distribution (3) is a recalibration of (2) to the distribution of the anchor function that arises by the reduced design. In the special case where the reduced likelihood $\text{pr}_{\text{REDU}}(D_{\text{REDU}} \mid \theta)$ is a marginal distribution of the full likelihood, the factor $\text{pr}_{\text{REDU}}(A(\theta)|D_{\text{REDU}})$ of (3) is proportional to an integrated likelihood.

The polydesign distribution can form a basis of estimating θ in a number of ways. To demonstrate, we discuss here operating characteristics of the mode of the distribution because of its analogy to maximum likelihood. For any given polydesign I_{POLY}, the above structure suggests the following properties.

PROPERTY 1

(a) If the assumed full likelihood is correct then, for any anchor function, the mode of the polydesign distribution (3) is consistent for the true parameter θ_0.

(b) For any misspecification of the full likelihood, and with respect to any loss function for a true parameter, there exists an anchor function so that the mode of the polydesign distribution (3) performs uniformly at least as well as the mode of the posterior distribution (2).

A proof of the first part can be established if it is shown that the right multipliers in the right hand side of both distributions (3) and (2) converge to the point mass at the anchor function's value at θ_0. The proof of the second part is straightforward because the class of distributions (3) generated by a polydesign provides a continuum that contains the distribution (2) when $A()$ is free of θ. In fact, part (b) would hold even if we

relaxed identifiability of $A(\theta)$ from the reduced likelihood, which is, nevertheless, desired to allow for use of improper prior distributions, and, hence, for more dependence on the data. Moreover, the optimum is expectedly achievable within the continuum, i.e., with a nonconstant anchor function. Finding an analytic function giving the optimum depends on how a misspecification is expressed, and is not here our goal. In practice, finding the optimum anchor function can practically be replaced by selection of a plausible candidate in the class of polydesigns and by simulation under misspecifications of concern to judge the performance of the candidate.

Likelihood perspective.

Analogously to the Bayesian perspective, we consider a synthetic estimation of $\theta = (\theta - A(\theta), A(\theta))$ by basing estimation of the anchor function $A(\theta)$ on the reduced likelihood, and estimation for the remaining part of the parameter using the full likelihood. Specifically, define:

$$\theta_{\text{POLY}} := ((\theta - A)_{\text{FULL}}, A_{\text{REDU}}), \quad \text{where} \tag{4}$$

$\text{pr}_{\text{REDU}}\{D_{\text{REDU}} \mid [\theta - A(\theta), A(\theta)]\}$ is maximum at $[(\theta - A)_{\text{REDU}}, A_{\text{REDU}}]$, and

$\text{pr}_{\text{FULL}}\{D_{\text{FULL}} \mid [\theta - A(\theta), A_{\text{REDU}}]\}$ is maximum at $[(\theta - A)_{\text{FULL}}, A_{\text{REDU}}]$

The estimator θ_{POLY}, therefore, is the maximizer of the full likelihood after having profiled (Murphy and van der Vaart, 2000) the likelihood on the anchor function that maximizes the reduced likelihood. In the special case where the reduced likelihood $\text{pr}_{\text{REDU}}(D_{\text{REDU}} \mid \theta)$ is a marginal distribution of the full likelihood, the estimator $A(\theta)_{\text{REDU}}$ is a "marginal" (or "restricted", Patterson and Thompshon, 1971) likelihood estimator.

In analogy to the Bayesian perspective, for any given polydesign I_{POLY}, the above structure suggests the following property.

PROPERTY 2

(a) If the assumed full likelihood is correct then, for any anchor function, the polydesign estimator θ_{POLY} defined in (4) is consistent for the true parameter θ_0.

(b) For any mispecification of the full likelihood, and with respect to any loss function for a true parameter, there exists an anchor function so that the polydesign estimator θ_{POLY} performs uniformly at least as well as the maximum likelihood estimator of the full likelihood.

Part (a) can be shown using a variation of Wald's proof for consistency of maximum likelihood, and the proof of part (b) is again straightforward and is followed by analogous arguments to those of Result 1(b).

3. Implementation

Obtaining the value of θ_{POLY} defined in (4) is a maximization problem, although possibly a challenging one in terms of computational stability if the anchor function is complicated.

In contrast, the polydesign distribution (3) can be relatively easily obtained through the following simulation.

Step 1: Approximate $\text{pr}_{\text{FULL}}(A(\theta)|D_{\text{FULL}})$. To do so, one can simulate random draws from the normal approximation based on the MLE of θ of the full likelihood, and then simulate from $\text{pr}_{\text{FULL}}(\theta|D_{\text{FULL}})$ using sampling importance resampling (SIR, Rubin, 1987). For each draw, calculate $A(\theta)$ and estimate $\text{pr}_{\text{FULL}}(A(\theta)|D_{\text{FULL}})$, e.g., with a kernel or a normal approximation.

Step 2: Approximate $\text{pr}_{\text{REDU}}(A(\theta)|D_{\text{REDU}})$. Because maximization of the reduced likelihood to obtain a normal approximation may be unstable, one can start, as in step 1, by simulating from the normal approximation based on the MLE of θ of the full likelihood, and then simulate from $\text{pr}_{\text{REDU}}(\theta|D_{\text{REDU}})$ using SIR. For each draw, calculate $A(\theta)$ and then estimate $\text{pr}(A(\theta)|D_{\text{REDU}})$ as in step 1.

Step 3: Approximate the function $r(\theta) = \text{pr}_{\text{REDU}}(A(\theta)|D_{\text{REDU}})/\text{pr}_{\text{FULL}}(A(\theta)|D_{\text{FULL}})$, i.e., the importance ratio of the polydesign distribution $\text{pr}_{\text{POLY}}(\theta)$ in equation (3) to $\text{pr}_{\text{FULL}}(\theta|D_{\text{FULL}})$ in equation (2). This is obtained by simply dividing the two functions obtained in steps 1 and 2.

Step 4: Simulate from the polydesign distribution $\text{pr}_{\text{POLY}}(\theta)$. To do so, use simulations from step 1, and use SIR by noting that the ratio of the target to candidate distributions is the function $r(\theta)$ of step 3. When enough draws from the target distribution have been obtained, quantiles, the posterior mode, mean, standard deviation, and other summaries can be computed.

The fourth step is alternative description, from an

implementation perspective, that the polydesign distribution is a recalibration of $\text{pr}_{\text{FULL}}(\theta|D_{\text{FULL}})$ to be such that the marginal distribution of the anchor function $A(\theta)$ be equal to the posterior distribution arising from the reduced likelihood $\text{pr}_{\text{REDU}}(A(\theta)|D_{\text{REDU}})$. In that sense, $A(\theta)$ is the intermediate device connecting the full and reduced designs.

4. Discussion

In settings such as partially controlled studies, simply reducing the original (full) design generally does not retain enough information for estimating well the quantities of interest. In such settings, we have proposed a framework that can still assess sensitivity of results to different designs, by combining them with the full design.

Polydesigns can help assess sensitivity of results to the modelling of different features of the data. f the model for the data of the full design is correct, then sensitivity of results to a range of polydesigns is not expected. Thus, evidence of sensitivity to a range of polydesigns is evidence of model mispecification in the structures of the data where the polydesigns differ.

The properties discussed in Section 2 indicate a related but more direct utility of polydesigns. A particular polydesign provides inference that is valid if the model is correct, and that can be more accurate to a model mispecification than the full design, if the reduced design is chosen in a way that it reduces dependency on those feautures of the full data for which that mispecification is a concern.

More specifically, the properties discussed in Section 2 indicate that there can be a precise map, so that if one expresses the type of mispecification that is of concern along with a loss function for estimation, then that map would point precisely at the optimal polydesign and anchor function. Choosing intuitive candidates of the polydesign and anchor function, and then verifying their properties with a simulation around small mispecifications can go a long way in exploring and obtaining polydesigns and anchor functions that are close to the optimal ones. Nevertheless, an analytic map as described above is also of interest for further study.

References

Breslow, N. E. and Day, N. E. (1980). The analysis of case-control data. In *Statistical methods in Cancer Research*, vol 1, World Health Organization.

Cox, D. R., and Oakes, D. (1984). Analysis of survival data. London: Chapman and Hall.

Drucker E., Lurie P., Wodak A. and Alcabes P. (1998). Measuring harm reduction: the effects of needle and syringe exchange programs and methadone maintenance on the ecology of HIV. *AIDS* **12**, Suppl A: S217–230.

Frangakis, C.E., Brookmeyer, R.S., Varadhan, R., Safaeian, M., Vlahov, D. and Strathdee, S.A. (2003). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. Forthcoming in the *Journal of the American Statistical Association*.

Frangakis, C.E. and Rubin, D.B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86**, 365–379.

Frangakis, CE, and Rubin, DB (2001). Addressing the idiosyncrasy in estimating survival curves using double-sampling in the presence of self-selected right censoring. *Biometrics* (with discussion), **57**, 333–353.

Frangakis, C.E. and Rubin, D.B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.

Keende, J. Stimson,G. V, Jones, S., and Parry-Langdon, N. (1993). Evaluation of syringe-exchange for HIV prevention among injecting drug users in rural and urban areas of Wales *Addiction* **88**, 1063-1070.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* (with discussion) **95**, 449–485.

Patterson H.D, and Thompson, R. (1971). Recovery of interblock information when blocks are unequal. *Biometrika* **58**, 545-554.

Robins, J. M., Rotnitzky, A., and Bonetti, M. (2001). Comment on "Addressing an idiosyncrasy in estimating survival curves using double-sampling in the presence of self-selected right censoring ", by CE Frangakis and DB Rubin. *Biometrics*, **57**, 343-347.

Rubin, D. B. (1987). The Calculation of Posterior Distributions by Data Augmentation: Comment: A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm. *Journal of the American Statistical Association*, **82**, 543-546.

Strathdee, S.A., Celentano, D.D., et.al. (1999). Needle-exchange attendence and health care utilization promote entry into detoxification. *Journal of Urban Health* **76**, 448–460.

van Ameijden, E.J, van den Hoek, J.A., Hartgers, C. and Coutinho, R.A. (1994). Risk factors for the transition from noninjection to injection drug use and accompanying AIDS risk behavior in a cohort of drug users. *American Journal of Epidemiology* **139**, 1153–1163.

Vlahov, D., Anthony, J.C., Munoz, A., et. al. (1991). The ALIVE study. A longitudinal study of HIV infection among injection drug users: Description of methods. NIDA Research Monograph **107**, 75–100.