Johns Hopkins University, Dept. of Biostatistics Working Papers

2-1-2005

Designs in Partially Controlled Studies: Messages from a Review

Fan Li

Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, fli@jhsph.edu

Constantine E. Frangakis

Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, cfrangak@jhsph.edu

Designs in partially controlled studies: messages from a review

Fan Li and Constantine E. Frangakis [1]
Department of Biostatistics, Johns Hopkins University, USA
February 4, 2005

The ability to evaluate effects of factors on outcomes is increasingly important for a class of studies that control some but not all of the factors. Although important advances have been made in methods of analysis for such partially controlled studies, work on designs for such studies has been limited. To help understand why, we review main designs that have been used for such partially controlled studies. Based on the review, we give two complementary reasons that explain the limited work on such designs, and suggest a new direction in this area.

[1]Fan Li is doctoral candidate and Constantine Frangakis is Associate Professor, Department of Biostatistics, Johns Hopkins University, USA. Responsible author: Fan Li, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205, USA, email: fli@jhsph.edu, Fax: 410-955-0958

1. Introduction

In order to evaluate the effects of factors on an outcome, interest has been increasing for a broad class of studies that can control the mechanism of assignment of some but not all of the factors. Such "partially controlled studies" have their origin in economics, where they have been evaluated traditionally with the method of instrumental variables (IV) (1). However, as the use of partially controlled studies has expanded in medicine, public health, and social sciences, it has created the need also for broader methods than IV, in order to address different target quantities for estimation, to allow more plausible underlying assumptions, and also to allow different modes of inference.

Partially controlled studies presently are important in the following three main categories. First, a prospectively designed study respecting ethical concerns may be able to only control the encouragement of participants to receive specific treatments, and leave to the participants the choice of the treatment they will actually select. In this category belong, for example, randomized studies with noncompliance to treatment (e.g., (2)-(11)). Second, even if there is sufficient ethical basis, the study may not have a priori a known mechanism by which to change directly the factors of interest. In this second category belong, for example, studies examining pathways that mediate the effect that a controlled treatment has on a clinical outcome (e.g., (12)-(16)). Finally, for studies that are not prospectively designed as controlled, it is still beneficial if they can be framed in a template of a partially controlled study, in order to provide more accurate evaluations of the factors of interest. This requires that a factor be controlled by some system, and the mechanism of assignment (24) of that factor be ignorable by the investigator (24), since, then, a distinction between the investigator and the system that actually controls the factor is inessential for valid likelihood inference. In this category belong, for example, some studies in economics that observe human behavior without an attempt to explicitly control it (e.g., (17)-(18)) but where it is plausible to assume control of other factors related to that behav-

ior. Such studies become more demanding when they need to handle a multitude of partially controlled factors, such as mechanisms creating missing or undefined data (e.g., (19)-(22)).

The extended use of partially controlled studies, and the improvement in methods of their evaluation given data, would imply that there should also exist better designs, for example, in order to better focus on more informative subsets of the data, or to increase cost-effectiveness in conducting the studies. However, study on more flexible designs in partially controlled studies has been limited.

To help understand why, we review main designs that have been used for such studies. We focus on designs that can be combined with relatively general measurements of each of the factors at a single time, as the issues in these designs can then extend to longitudinal factors. In the review in the next section, we examine the main problem each design attempts to address, and we summarize briefly examples. For each example, we examine the reason for using a template of a partially controlled study, the reason for the corresponding design, and the statistical analysis. The examples reviewed are not all meant to be chronologically the first ones using the corresponding design, but are chosen for their suitability to highlight the main arguments, and because they share some assumptions commonly made with partially controlled studies. The review confirms the limited work on designs even when making such assumptions. Section 3 gives two main explanations for this limitation, and uses these explanations to suggest a new direction of work in designs in this area.

2. Designs for partially controlled studies

2.1 Cohort design with completely randomized assignment

We discuss the design of a simple cohort in terms of a template introduced in Angrist, Imbens and Rubin (7) for noncompliance, and later generalized for partially controlled studies (13). This template is based on potential outcomes, (23), (24), that is, outcomes of a person that,

before the study starts, are observable under different levels of a controlled factor. We use this template in the context of a hypothetical study, in order to establish some terminology and notation to make connections with the other designs.

The study controls the assignment of each of a group of consenting participants i to either an experimental treatment $(z = 1)$ or the standard treatment $(z = 0)$. Suppose that the experimental treatment has been developed to improve a clinical outcome Y by first trying to increase levels of a variable E, for example, to regulate the expression level of a particular gene by targeting its enhancer or promoter region. Because the levels of E are not controlled directly, but are affected by the controlled factor z – the treatment, we call E a "partially controlled" factor. Let $E_i(1) = 1(\text{ or } 0)$ indicate if the participant's gene expression level is high (or low) two months after assignment, if that participant is originally assigned to experimental treatment; and let $E_i(0) = 1(\text{or } 0)$ indicate the participant's level of E two months after assignment, if that participant is originally assigned to standard treatment. Also, let $Y_i(z)$ be the outcome of interest (e.g., 1 year survival) if the participant is assigned to treatment $z, z = 0, 1$.

The goal is to examine the role that gene expression E has in the effect of treatment on outcomes. To address this, it is first important to consider the principal stratification (13) of participants defined by the joint values of the gene's expression under the two treatments, $S_i = (E_i(0), E_i(1))$: participants whose expression would be low if assigned standard and also if assigned experimental treatment, $\{i : E_i(0) = E_i(1) = 0\}$, whom we call "low-expressed"; participants that would have a low expression level if assigned standard treatment and that would have high expression level if assigned experimental treatment, $\{i : E_i(0) = 0 \text{ and } E_i(1) = 1\}$, whom we call "responders" to the experimental treatment; participants that would have a high expression level if assigned the standard and also if assigned the experimental treatment; $\{i : E_i(0) = E_i(1) = 1\}$, called "fully-expressed"; and participants who would have a high expression if assigned standard but a low expression if assigned experimental treat-

ment, $\{i : E_i(0) = 1 \text{ and } E_i(1) = 0\}$, whom we call "special".

Although a participant's gene expression level can be affected by treatment, a participant's membership S_i to the above principal strata is not affected by treatment. For this reason, the substantive goal in studies with such structure for a partially controlled factor can often be to estimate the causal effects that the treatment z has on the outcomes Y conditionally on some of the principal strata S.

Addressing this goal is complicated by the fact that memberships to the principal strata cannot all be directly observed, so it is important to consider simplifying assumptions on the above template, and it is necessary to also describe the design mechanism for collecting the observed data. A large part of work cited in Section 1 using this template makes some variants of the following assumptions of instrumental variables (7).

Monotonicity. There are no "special" participants.

Exclusion Restriction. If for a participant, treatment does not affect the factor E, then treatment does not affect the outcome Y, i.e., if $E_i(0) = E_i(1)$ then $Y_i(0) = Y_i(1)$.

Monotonicity is posited when assignment to the experimental treatment is designed to keep the levels of the partially controlled factor E to be at least those that would be achieved under standard treatment. Exclusion is posited when the experimental treatment has been designed so as to have no other plausible way to affect the outcome Y except if it affects the factor E.

In the example where gene expression is the partially controlled factor, exclusion and monotonicity can be made plausible by targeting, respectively, a specific mechanism and specific direction of that mechanism, as when a treatment targets the promoter or enhancer of a specific gene in studies of prostate cancer (e.g., (25)). Note that such properties do not assume that, if the treatment does affect expression for an individual, then it will have no side effect, and side effects should also be studied as outcomes. Of course, depending on the application,

it may be important to consider deviations from the above assumptions.

Figure 1 here

Under monotonicity, the "responders" are the only principal stratum for which the experimental comparison between different treatments is a comparison between different gene expression levels. Also, under exclusion, the "responders" are the only principal stratum whose outcome can be affected by treatment. For these reasons and under these assumptions, the effect of treatment on outcome for "responders" in the population, $E(Y_i(1) - Y_i(0) \mid S_i = \text{responder})$, denoted here by $\delta Y_{\text{responder}}$, often can quantify the target of estimation. In the gene expression example, the magnitude of $\delta Y_{\text{responder}}$ will indicate how experimentally increasing expression levels E of the targeted gene associates to achieving better clinical outcomes Y, and thus will provide guidance on whether a more focused research is needed on that gene's properties.

For the mechanism of observing data, a simple cohort randomized design assumes the following.

Completely randomized assignment. Assignment to the experimental or standard treatment is completely randomized, $Z_i \perp\!\!\!\perp (Y_i(z), S_i)$

Simple random sampling of participation. Study participants are a simple random sample from a population of consenting participants.

The first assumption essentially implies that assignment is ignorable in the sense of Rubin (24) without conditioning on other variables; the second assumption allows the results from the study participants to be generalized in a simple way to a larger population. Figure 1 shows the relation of observed data to potential outcomes and principal strata under these assumptions, where $E_i^{obs} = E_i(Z_i)$ and $Y_i^{obs} = Y_i(Z_i)$ are the observed expression levels and outcomes. Responders are not observed alone in either treatment arm, and estimation of $\delta Y_{\text{responder}}$ needs to rely on its indirect relation to the observable data.

Such a relation is obtained by decomposing the intention-to-treat effect of assignmed treatment, $\delta Y = E(Y_i(1) - Y_i(0))$, across principal strata as $\sum_p E(Y_i(1) - Y_i(0) \mid S_i = s)\text{pr}(S_i = s)$, which reduces to $\delta Y_{\text{responder}}\text{pr}(S_i = \text{responder})$, giving

$$\delta Y_{\text{responder}} = \delta Y / \text{pr}(S_i = \text{responder}). \tag{2.1}$$

by exclusion and monotonicity, as in Angrist et al. (7). The assumptions imply that the probability of being a responder equals the difference in the probabilities of having a high gene expression when assigned experimental versus when assigned standard treatment, or $\text{pr}(E_i^{obs} = 1 \mid Z_i = 1) - \text{pr}(E_i^{obs} = 1 \mid Z_i = 0)$. The latter difference is estimable consistently by the corresponding sample averages. Also, the effect of assigned treatment on all participants, δY, is estimable directly from the difference in sample averages in the outcomes of the participants assigned experimental vs. standard treatment. So, in this design, the effect on responders, $\delta Y_{\text{responder}}$, can be estimated consistently by substituting simple sample analogues of the quantities in the right side of (2.1).

The above template is often useful in studies that do not explicitly control treatment either, but where using the template through some modifications can be justified and can lead to more reliable inferences on the effect associated with the partially controlled factor.

2.2 Paired Availability Design

Often, a common concern is that the assignment mechanism of the controlled factor varies systematically across strata of known variables that are also associated with the outcome and the principal strata. This can be addressed by planning for a design that measures the known variables conditionally on which the assignment can then be assumed ignorable. From the standpoint of analysis, this design can then be addressed with methods of varying degrees of

adjustments.

A simple such design that demonstrates the point is the paired availability design (PAD) proposed in a study by Baker and Lindeman (26), prior to the more explicit template with potential outcomes cited in Section 2.1 (see also Cuzick, Edwards, and Segnan (27) in the setting of a multi-center randomized trial). The study's goal was to evaluate the effect that receiving epidural analgesia (EA) during women's labor has on getting a Caesarean section (CS). In this case, randomized clinical trials were difficult to be implemented. Alternatively, the study could have compared the rate of CS among women who received EA to those who did not. However, these two groups of women expectedly differed in baseline characteristics in both observed and unobserved ways that related to the likelihood of having CS (26).

The study addressed the above issue by measuring availability of EA, in addition to actually receiving it. In terms of the template of Section 2.1, the factor z is whether or not EA was available to a woman at her hospital at the time she was giving birth; $E_i(z)$ indicates whether or not the woman actually would have received EA as a function of availability; and $Y_i(z)$ is the indicator of getting a CS, for $z = 0, 1$. Because a woman could not receive EA if it was not available, there are two principal strata S: women who would receive EA if it were available; and women who would not receive EA no matter its availability.

The main problem here with the assumptions of the previous section focuses on complete randomization. The availability of EA varied with the timing that hospitals started offering the procedure. Hospitals, however, can differ systematically in both the relative frequency in which they made EA available to women, and also in the characteristics associated with the likelihood of the women receiving EA and having CS. This makes the assignment unconditionally not ignorable.

The PAD addressed this problem by planning for and measuring the information on the hospitals to which the women in the study gave birth. Availability of EA then was treated as a

controlled factor, in the sense here that, conditionally on the hospital information, the rule used by the hospitals for the actual timing of the availability is assumed ignorable by the investigator. In this design, a "pair" is specific to each hospital and its two "members" are the two groups of assignment, women when EA was less available and women when EA was more available within each hospital.

The study analysed the design by first carrying out completely separate analyses within each hospital, each analysis being analogous to that described at the end of the template in the previous section. These analyses for hospital h produce estimates, say $dY_{\text{responder},h}$, of the effect $\delta Y_{\text{responder},h}$, which is now the effect that availability of EA has on the probability of CS, for the women who would receive EA if it were available. To combine the results from each hospital, the study used an average of $dY_{\text{responder},h}$, weighted by the inverse estimated variance of the estimated effects within each hospital. A generalized version of this analysis of PAD has also been given (28), which allows different types of availability and multiple time periods.

2.3 Matched pairs of assignment design

The full stratification and corresponding analyses in PAD is feasible when the strata are relatively few. However, a finer stratification based on more than one covariate is often desirable. Although this can be addressed with modelling, it can also be addressed with a design that matches on covariates two participants assigned different treatments. This design is important in partially controlled studies, as in other types of studies, for two reasons. First, the investigator may want to rely less on model assumptions of how the covariates relate to principal strata and potential outcomes. Second, the investigator may need to limit cost by selecting for follow-up a subset of the participants, and matching can help such selection. For each reason, we review below a partially controlled study with the matched pairs of assignment design, the

first using randomization inference and the second using likelihood inference.

Randomization inference.

Rosenbaum (29) provided an example of a matched pairs design from an earlier study by Card and Krueger (30) on the effect that increase in minimum wage has on changes in employment in the fast food industry. The original study had been designed around April 1992 when New Jersey increased its minimum hourly wage from \$4.25 to \$5.05. The original study had collected wage and employment data on fast food restaurants in New Jersey (where the law changed) and in neighbouring Pennsylvania (where the law did not change), before and after the change in New Jersey's law.

To estimate the effect of wage increases on employment, the study of Rosenbaum (29) used the change in law as an instrumental variable. In terms of the template of Section 2.1, the "participants" are restaurant units; the factor z for restaurant i is the indicator for whether the new law would be applicable (in New Jersey) or not applicable (in Pennsylvania) to that restaurant after April 1992; and $E_i(z)$ and $Y_i(z)$ are, respectively, the log wage and log employment at the restaurant after April 1992 if the new law were applicable $(z = 1)$ or not applicable $(z = 0)$ to that restaurant. Although the study did not explicitly control the change of the law, it first treated the change as controlled under the assumption of ignorability, but it also considered nonignorable assumptions. Ignorability here means comparability between the restaurants to which the new law was applicable and those to which the new law was not applicable.

The matched pairs approach in this study had a design, a model, and an analysis component. The design created 66 matched pairs of restaurants, each pair having one from New Jersey and one from the neighbouring area of eastern Pennsylvania. Within pairs, restaurants were matched for restaurant chain and starting wages before the increase in New Jersey's law. The analysis with these data was based on a model on the potential outcomes and wages, and a permutation inference on a robust estimate of the parameter characterizing the effect in that

model. We examine the model and inference separately, to highlight their different roles for this design.

The model is specific to each pair of the design. Let $(E_{p,1}(0), E_{p,1}(1))$, be the principal stratum in (log) wages of unit "1" in pair p, and $(E_{p,2}(0), E_{p,2}(1))$ be that of unit "2" in pair p, where "1" and "2" are randomly given labels. Similarly let $(Y_{p,i}(0), Y_{p,i}(1))$ be the potential (log) employment values for unit i in pair p. The study's model is:

$$\begin{aligned} &Y_{p,i}(z) = r_{p,i} + \beta E_{p,i}(z), \text{ which implies} \\ &Y_{p,i}(1) - Y_{p,i}(0) = \beta(E_{p,i}(1) - E_{p,i}(0)). \end{aligned} \tag{2.2}$$

In analogy to the template of Section 2.1, the effect of interest β by the bottom expression of (2.2) is defined through the principal strata $(E_{p,i}(0), E_{p,i}(1))$, as the ratio of the effect that the law has on the increase in employment Y, per unit of effect that the law has in increasing actual starting wage E. The bottom expression also clarifies that exclusion is assumed here as defined in Section 2.1, although here E is continuous.

To estimate β, the study considers the within-pair p observed contrasts in outcome, dY_p, and in wage, dE_p,

$$dY_p = \begin{cases} Y_{p,1}(1) - Y_{p,2}(0) \\ Y_{p,2}(1) - Y_{p,1}(0) \end{cases} \quad \text{and} \quad dE_p = \begin{cases} E_{p,1}(1) - E_{p,2}(0) & \text{if} \quad Z_p = 1 \\ E_{p,2}(1) - E_{p,1}(0) & \text{if} \quad Z_p = 0 \end{cases}$$

where Z_p is the indicator for unit "1" of the pair being subject to the new law. Model (2.2) then implies that in matched pair p, dY_p and dE_p are related through

$$dY_p - \beta dE_p = (2Z_p - 1)(r_{p,1} - r_{p,2}) \tag{2.3}$$

Based on the above, the study proposed a permutation based inference by inverting a function $g(\{dY_p - \beta dE_p\})$, considering both ignorable and non-ignorable assignments within pairs; we focus here on inference under ignorability. Such inference hypothesizes a value β^*; then calculates $(r_{p1} - r_{p2})$ for each pair using in (2.3) the hypothesized β^* and the observed values of dY_p, dE_p and Z_p; calculates the distribution of $g(\{dY_p - \beta^* dE_p\})$ as induced by the distribution in the right hand side of (2.3) and using, based on ignorability of the assignment Z_p within pairs, the probabilities pr$(Z_p = 1) = 0.5$; and tests the hypothesized β^* by testing, against that latter distribution, the value of $g(\{dY_p - \beta^* dE_p\})$ evaluated at the observed differences $\{dY_p, dE_p\}$. For the function g, the study proposed a rank based statistic placing low weight on outliers.

Remark. While the above analysis is useful, it is not the only one that can address this design. That study argued ((29), p. 75) that permutation as a mode of inference allows one to avoid making assumptions such as monotonicity of Section 2.1, while such assumptions are required by other modes of inference. However, a more careful examination of the above arguments shows that it is the deterministic (and thus often implausible) aspect of the model (2.2), rather than the permutation inference, that avoids making, for example, the monotonicity assumption. To see this, note that when taking expectations of both sides of (2.3), first with respect to the assignment given pair, and then over the larger population of pairs, we have that

$$E(dY_p) = \beta \; E(dE_p)$$

since the expectation of the right hand side of (2.3) is 0. So using the averages, over pairs, of dY_p and dE_p to replace the expectations in the above expression gives an estimator that is consistent for β and whose large sample properties are easy to derive with the delta method on the ratio of averages. Thus inference with the model (2.2) can also be done without permutation

arguments and without using a monotonicity assumption.

Likelihood inference.

The matched pairs of assignment design is also useful for cost reduction, and an example is described in detail by Barnard et al. (11). The goal in that study was to assess the effect that using school choice vouchers to attend private schools can have on children's school performance. For this question, directly comparing children who attend private choice schools versus those who attend public schools is problematic due to the multitude of socioeconomic and motivation factors that can be different between these two groups and that are difficult to measure. To better evaluate the effect of school choice, the study designed and conducted a randomized experiment in New York City. Funding for the experiment was provided for 1300 scholarships to low-income families to attend private schools. A randomized lottery determined the 1300 winning families whose children, among over 20 000 applicants, were offered the scholarships. Children were then followed and evaluated with standardized tests at yearly intervals; see Peterson and Howell (31) for an update.

Given the large number of children who did not win the scholarship, and the fixed budget for the study, it was not feasible to follow all the children. For this reason, a matched pairs of assignment design was considered for a subset of the children. Because the number of covariates was fairly large, exact matching on all covariates was not possible. To address this, the study estimated a propensity score (32) for the assignment to private schools (winning the lottery) and used that to match each child who had been assigned to private school to a child who not been so assigned. With this propensity score matched pairs design (33), although, by randomization, the true propensity score is known and constant, matching in the estimated propensity score generally increases efficiency compared to having selected a random subset (34).

Two main issues make this study different from that of Rosenbaum (29). First, this study

had more than one factors that were partially controlled: compliance of children with the assigned vouchers, missingness in outcomes and missingness in covariates. For such cases, it has been shown that using a standard instrumental variables approach is not appropriate to estimate the effect of using vouchers on the outcomes (19). Details for how this can be addressed using principal stratification are given in Barnard et al. (11).

The second main distinction is the role played by the matched pais of assignment in the analysis. In the likelihood approach taken by Barnard et al. (11), as long as the variable used to create the matched pairs – the estimated propensity score – is used in the models for the potential outcomes and principal strata, the pairing of the design becomes ignorable in the sense of Rubin (24). So, from this perspective, the main role of the design of the assignment is to generate better data, i.e., that are more likely to produce more efficient estimates; but after the data have been obtained, the ignorable design is not relevant in the inferential method.

The contrast between the above two examples for the design on matched pairs of assignment emphasizes a trade-off in approaches to its analysis. First, data from this design can be analysed with a likelihood model, although its adequancy is a concern and needs checking as in other model settings. On the other hand, with a permutation approach, a deterministic model between the potential outcomes and the principal strata as in model (2.2) seems central for allowing exact calculation of the distribution of a pivotal quantity for the effect of interest. Thus, although in partially controlled studies permutation inference could be advocated as a means for avoiding model assumptions, it needs such and possibly even stronger assumptions to address nuisance unknown quantities such as the values of the principal strata.

2.4 Unbalanced assignment and compliance subsampling designs

In randomized trials that a priori expect noncompliance to the treatment, conventional calculation of the sample size and the proportion of participants to be assigned to the experimental

arm is based on a specified target value for the intention-to-treat effect, and on the properties of intention-to-treat analyses. When the target estimand and analysis are different, such as those described in Section 2.1, then the above calculations can be suboptimal. To address this, one can obtain the relation of cost to precision for estimating the new target, and based on this relation find a better design.

Jo ((35),(36)) explored this issue of cost in randomized trials where noncompliance is expected and is to be addressed with models for the template described in Section 2.1. The study crossed different cost scenarios, depending on what cost is incurred by actually receiving the assigned treatment, with scenarios of different fraction of participants in the experimental assignment and different expected conditions on compliance rate and outcome distributions. For each scenario, the study calculated the total cost, over simulations, to have fixed power to detect a specified target value for the effect of receiving treatment, $\delta_{\text{responder}}$. Estimation was based on maximum likelihood of the model parameter corresponding to the estimand. The results showed that to attain the same statistical power for estimating the effect of receiving treatment, an appropriate unbalanced assignment design has usually lower cost than a balanced design. The gains are increasing with increasing difference in the costs of treating individuals with the experimental versus with the standard treatment.

In the above work, compliance behavior was considered to be measured for all units. For situations when measuring compliance can be costly and optional, a general class of "compliance sub-sampling"(CSS) designs has been proposed (37) that allows compliance information to be measured for only subgroups of participants. Specifically, CSS allows the investigator to choose on the total number of participants, the fractions assigned to experimental versus standard treatment, and the fractions of participants for whom compliance will be measured in the two assignment arms. These choices are allowed to depend on the cost parameters of the study and the anticipated distributions of principal strata and potential outcomes. For each CSS

design, that study developed expressions of the maximul likelihood estimator for the effect of receiving treatment, $\delta_{\text{responder}}$, and derived the minimal-cost CSS design that achieves the required precision for estimation. The study showed that the optimal design to balance total cost with precision for estimating the effect of receiving treatment, is often a design that measures the compliance behavior for only representative subgroups in two unbalanced assignment arms.

2.5 Clustered assignment design with individual noncompliance

When potential participants are part of clusters structure, a practical design can often be to assign treatments by the clusters. Then, if participants comply differentially within and across clusters, statistical methods need to address the combination of individual noncompliance with the design effect of the clustered assignment.

An example of this clustered assignment with individual noncompliance was reported by West et al. (38) and addressed as a partially controlled study by Korhonen et al. (39). The original study's goal was to examine if oral supplements of vitamin A could reduce infant mortality in Nepal. A complete randomization of assignment of participants, as in Section 2.1, and even of families, could be problematic in this case, for example, if families assigned to pills of vitamin A would share pills with neighbouring families assigned placebo, leading to contamination. To avoid such problems, the study randomized the assignment at the level of wards, where all families within a ward were assigned the same treatment. Because subsequently the treatment actually received was unknown for 13% of the children, Korhonen et al. (39) discussed estimation of the effect of taking vitamin A on survival under different hypotheses about compliance.

The study defined the effect of taking vitamin A on survival by (a) assuming for all subjects a survival time, P_i, that would have been observed had child i received no vitamin A; and

(b) assuming that P_i equals to $Y_i^{obs} \exp(\psi_0 E_i^{obs})$ in distribution, where Z_i, E_i^{obs}, and Y_i^{obs} are, respectively, child i's observed assignment, observed treatment received and observed survival time after enrollment. Estimation was actually implemented using the stronger model:

$$\log(Y_i^{obs}) = \log(P_i) - \psi\, E_i^{obs}, \tag{2.4}$$

for all individuals i, and inverting a test of independence between the treatment-free survival $\{\log(P_i)\}$ and the randomized assignments $\{Z_i\}$ of treatment, adjusting for age at enrollment. The design effect of clustering was addressed by constructing a robust estimate of the variance of the estimate of the effect ψ_0.

This approach can be connected to potential outcomes and principal strata, if we note that model (2.4) can be viewed as arising from the following model,

$$\log(Y_i(z)) = \log(Y_i(0)) - \psi\, E_i(z). \tag{2.5}$$

for $z = 0, 1$ and all individuals i. Note also that it would be hard to justify assuming model (2.4) for the observed data without the mechanism (2.5). This is because, model (2.4) does not, in principle, preclude a study with just two participants ($i = 1, 2$) with exactly the same potential outcomes and principal strata, i.e., $Y_1(z) = Y_2(z)$ and $E_1(z) = E_2(z)$ for $z = 0, 1$. But then, making assumption (2.4) for both individuals, and simultaneously believing that (2.5) is wrong would lead to a contradiction.

In this sense, then, models (2.4) and (2.5) are equivalent and the approach of estimation by test inversion was, in the same spirit as Rosenbaum's (2.2), except that here, estimation was large-sample rather than permutation-based. An additional complication addressed in Korhonen et al. (39) was the censoring of the survival time for some participants. Clustered assignment with individual noncompliance has also been addressed with random effects (40),

(41).

3. Discussion and proposal for a broader class of designs

The reviewed designs allow deviations from the cohort completely randomized assignment design, in order to either limit model dependency, such as with the matched pairs of assignment, or to increase cost-effectiveness and practicality, such as with the unbalanced and clustered designs.

These designs are, in most part, limited to exploration of different plans only of the assignment of the controlled factor. In many instances, however, it would be important to be able to use alternative designs that would explore aspects of the partially controlled factors or of the outcome. For example, consider a study that has a relatively low variability in the observed outcome, such as with low number of cases, i.e., participants who experience an event indicated by $Y^{obs} = 1$. Then, to limit influence of covariate models in outlying regions of the data, we may want to consider a case-control design that would select data on all cases and only a subset of controls $(Y^{obs} = 0)$ who would match the cases based on some rule using covariates and possibly also the controlled and partially controlled factors. Although such "reduced" designs are frequently used in simpler frameworks (e.g., case-control matching on covariates with an ignorable treatment), they do not appear to have been used in any systematic way with partially controlled studies, even among those that make the assumptions of Section 2.1. The correspondence of analyses to the designs reviewed in the previous sections helps point at two related reasons for this.

First, addressing a "reduced" design using the traditional equations of instrumental variables as a starting framework is complicated because these equations are posited in terms of the observed controlled, partially controlled factor and outcomes. As Rubin (16) also observes, these equations already tie together assumptions specific to each participant (e.g., exclusion)

with assumptions of the design (e.g., some variant of the randomized assignment in the cohort). As a consequence, such equations cannot easily represent how assumptions on the observed data are changed when changing to a more general design but keeping other participant-specific assumptions the same. Weighting these estimating equations could theoretically lead to estimation if the inverse weights reflect formally probabilities of selection to the "reduced" design conditionally on the cohort. However, most often the selection rule conditionally on the cohort sample is closer to a deterministic rather than to a probabilistic rule, for example when choosing the control that most closely matches a case on some metric of other variables. Then, probabilities of selection to the reduced sample are at the boundary and the weighting approach is not useful.

On the other hand, using the framework of principal stratification as a starting point makes clear separation of the design from the participant-specific assumptions. This separation makes it relatively straightforward to calculate the likelihood of data from a "reduced" design, such as a conditional likelihood for the above case-control design. But because causal effects in this framework are defined based on the partially unobserved principal stratification, they are generally not fully identifiable from the likelihood of "reduced" designs.

This suggests that a way to make use of "reduced" designs in partially controlled studies is through a "polydesign", that is, a combination of a reduced and the cohort designs in two conceptual stages. In the first stage, a reduced design can be chosen so that its (conditional) likelihood better focuses on a subset of the cohort data. This "reduced" design can then be used to estimate the part of the causal effect that is identifiable from the "reduced" likelihood. If the causal effect is fully identifiable from that likelihood, then estimation stops here. Otherwise, in a second stage the cohort design is used to estimate the part of the causal effect that was not identifiable from the reduced likelihood.

Estimation based on a polydesign is expected to have two general properties. First, esti-

mation of the causal effect is expected to be valid if the model for the cohort is valid. Second, such estimation based on the polydesign is expected to be more robust than that using the full cohort, for misspecification of aspects of the cohort model that are not fitted when using the reduced likelihood in the polydesign. These two properties of polydesigns can be important also in more demanding problems.

Although most partially controlled studies, including those reviewed above, make at least some of the assumptions considered in Section 2.1, these assumptions may not always be plausible. In such cases, it is important to also consider sensitivity to these assumptions for example, in the sense of (43), (44) and (39). If such different assumptions are formulated in terms of potential outcomes and principal strata, rather than in terms of the observed data, their interpretation will be preserved across different designs or polydesigns.

Also, here we have formulated potential outcomes following the reasoning of Rubin (1978), as explicit functions of only the factor for which there is sufficient understanding to treat it as controlled. If similar understanding exists also for the other factors, which we treated here as partially controlled, then, by the same reasoning, potential outcomes should be extended to functions of both factors. For example, focus here has been on studies with single-time administration of the controlled and partially controlled factors. However, many studies are better formulated as longitudinal, which can be an especially fertile ground for polydesigns. Continued research is needed in this area, but preliminary work (42) suggests that such an approach can help better evaluate causal effects in partially controlled studies.

Acknowledgements

We thank Stuart Baker, Els Goetghebeur, Donald Rubin, and two referees for constructive comments.

References

[1] Bowden RJ, Turkington DA. " *Instrumental variables*." Cambridge: Cambridge University Press, 1984.

[2] Zelen M. "A new design for randomized clinical trials." *New England Journal of Medicine* 1979; **300:** 1242–1245.

[3] Sommer A, Zeger S. "On estimating efficacy from clinical trials." *Statistics in Medicine* 1991; **10:** 45–52.

[4] Connor R, Prorok PC, Weed DL. The case-control design and the assessment of the efficacy of cancer screening. *Jounral of Clinical Epidemiology* 1991; **44:** 1215–1221.

[5] Imbens GW, Rubin DB. "Causal inference with instrumental variables". Discussion paper # 1676. Cambridge, MA: Harvard Institute of Economic Research, 1994.

[6] Robins JM, Greenland S. "Adjusting for differential rates of prophylaxis therapy for PCP in high-versus low-dose AZT treatment arms in an AIDS randomized trial." *Journal of the American Statistical Association* 1994; **89:** 737–479.

[7] Angrist J, Imbens GW, Rubin DB."Identification of causal effects using instrumental variables (with discussion)" *Journal of the American Statistical Association* 1996; **91:** 444-472.

[8] Imbens GW, Rubin DB. "Bayesian inference for causal effects in randomized experiments with noncompliance." *Annals of Statistics* 1997; **25:** 305–327.

[9] Goetghebeur E, Molenberghs G. "Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance." *Journal of the American Statistical Association* 1996; **91:** 928–934.

[10] Baker SG. “Analysis of survival data from a randomized trial with all-or-none compliance: estimating the cost-effectiveness of a cancer screening program.” *Journal of the American Statistical Association* 1998; **93:** 929–934.

[11] Barnard J, Frangakis CE, Hill JL, Rubin, DB. “APrincipal Stratification approach to broken randomized experiments: a case study of School Choice vouchers in New York City.” *Journal of the American Statistical Association* (with discussion) 2003; **98:** 299-323.

[12] Joffe MM, Colditz GA. “Restriction as a method for reducing bias in the estimation of direct effects.” *Statistics in Medicine* 1998; **17:** 2233–2249.

[13] Frangakis CE, Rubin DB. “Principal stratification in causal inference.” *Biometrics* 2002; **58:** 21–29.

[14] Palmgren J. “Models for direct and indirect effects of exposure on response.” NORDSTAT 2002 (abstract).

[15] Gilbert PB, Bosch RJ, Hudgens MG. “ Sensitivity analysis for the assessment of causal vaccine effects on viral load in AIDS vaccine trials.” *Biometrics* 2003; **59:** 531–541.

[16] Rubin DB. “Direct and indirect causal effects via potential outcomes.” *Scandinavian Journal of Statistics* (with discussion) 2004; **31:** 161–170.

[17] Card D. “Using geographic variation in college proximity to estimate the return to schooling.” *National Bureau of Economic Research* 1993; Paper No. 4483.

[18] McClellan M, McNeil BJ, Newhouse JP. “Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables.” *Journal of the American Medical Association* 1994; **272:** 859–866.

[19] Frangakis CE, Rubin DB. “Addressing complications of intention-to-treat analysis in the presence of all-or-none treatment-noncompliance and subsequent missing outcomes.” *Biometrika* 1999; **86:** 365–379.

[20] Zhang JN, Rubin DB. “Estimation of Causal Effects via Principal Stratification When Some Outcomes Are Truncated By ’Death’.” *Journal of Educational and Behavioral Statistics* 2003, **28:** 353-368.

[21] Loeys T, Goetghebeur E.“A Causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance.” *Biometrics* 2003, **59**, 100–105.

[22] O’Malley J, Normand S-L T. “Likelihood methods for accounting for all-or-nothing treatment non-compliance and subsequent non-response in randomized clinical trials.” In: Proceedings of the Section on Bayesian Statistical Science, American Statistical Association 2003, (in press).

[23] Rubin DB. “Estimating causal effects of treatments in randomized and nonrandomized studies.” *Journal of Educational Psychology* 1974; **66:** 688–701.

[24] Rubin DB. “Bayesian inference for causal effects.” *Annals of Statistics* 1978; **6:** 34–58.

[25] Lee S-J, Zhang Y, Lee SD, Jung C, Li X, et al. “Targeting prostate cancer with conditionally replicative adenovirus using PSMA enhancer.” *Molecular Therapy* 2004; **10:** 1051–1058.

[26] Baker SG, Lindeman KS. “The paired availability design: a proposal for evaluating epidural analgesia during labor.” *Statistics in Medicine* 1994; **13:** 2269–2278.

[27] Cuzick J, Edwards R, Segnan N. “Adjusting for non-compliance and contamination in randomized clinical trials.” *Statistics in Medicine* 1997; **16:** 1017–1029.

[28] Baker SG, Lindeman KS. “Rethinking historical controls.” *Biostatistics* 2001; **2:** 383–396.

[29] Rosenbaum PR. “Using quantile averages in matched observational studies.” *Applied Statistics* 1999; **48:** 63–78.

[30] Card D, Krueger AB. “Minimum wage and employment: a case study of the fast-food industry in New Jersey and Pennsylvania.” *American Economic Review* 1994; **84:** 772–793.

[31] Peterson PE, Howell W. “Voucher reserch controversy.” *Education next* 2004; 73–78.

[32] Rosenbaum PR, Rubin DB. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 1983; **70**(1): 41–55.

[33] Hill JL, Rubin DB, Thomas N. “The Design of the New York School Choice Scholarship Program Evaluation.” In *Donald Campbell’s Legacy*, ed. L. Bickman, Sage Publications, 2000.

[34] Rubin DB, Thomas N. “Matching using estimated propensity scores: relating theory to practice.” *Biometrics* 1996; **52:** 249–264.

[35] Jo B. “Power to detect intervention effects in randomized trials with noncompliance.” Technical Report, Graduate school of education and information studies, UCLA, 1999.

[36] Jo B. “Statistical power in randomized intervention studies with noncompliance.” *Psycological Methods* 2002; **7:** 178–193.

[37] Frangakis CE, Baker SG. “Compliance sub-sampling designs for comparative research: estimation and optimal planning.” *Biometrics* 2001; **57:** 899–908.

[38] West KP Jr, Katz J, Shrestha SR, LeClerq SC, Khatry SK, Pradhan EK, et al. “Mortality of infants < 6 mo of age supplemented with vitamin A: a randomized, double-masked trial in Nepal.” *American Journal of Clinical Nutrition* 1995; **62:** 143–148.

[39] Korhonen P, Loeys T, Goetghebeur E, Palmgren J. “Vitamin A and infant mortality: beyond intention-to-treat in a randomized trial.” *Lifetime data analysis* 2000; **6:** 107–121.

[40] Loeys T, Vansteelandt S, Goetghebeur E. “Accounting for correlation and compliance in cluster randomized trials.” *Statistics in Medicine* 2001; **20:** 3753–3767.

[41] Frangakis CE, Rubin DB, Zhou XH. “Clustered encouragement design with individual noncompliance: Bayesian inference and application to Advance Directive Forms.” *Biostatistics* (with discussion) 2002; **3:** 147–164.

[42] Li F, Frangakis CE, and Varadhan, R. “Polydesigns in causal inference: definition and motivation.” In *Proceedings of the Biopharmaceutical Section of the American Statistical Association* 2003 (in press).

[43] Rosenbaum PR and Rubin DB. (1983). “Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome.” *Journal of the Royal Statistical Society* B **45:** 212–218.

[44] Scharfstein DO, Rotnitzky A, and Robins JM. (1999). “Adjusting for Nonignorable Drop-out Using Semiparametric Nonresponse Models (with discussion). *Journal of the American Statistical Association* **94:” 1096–1146.**

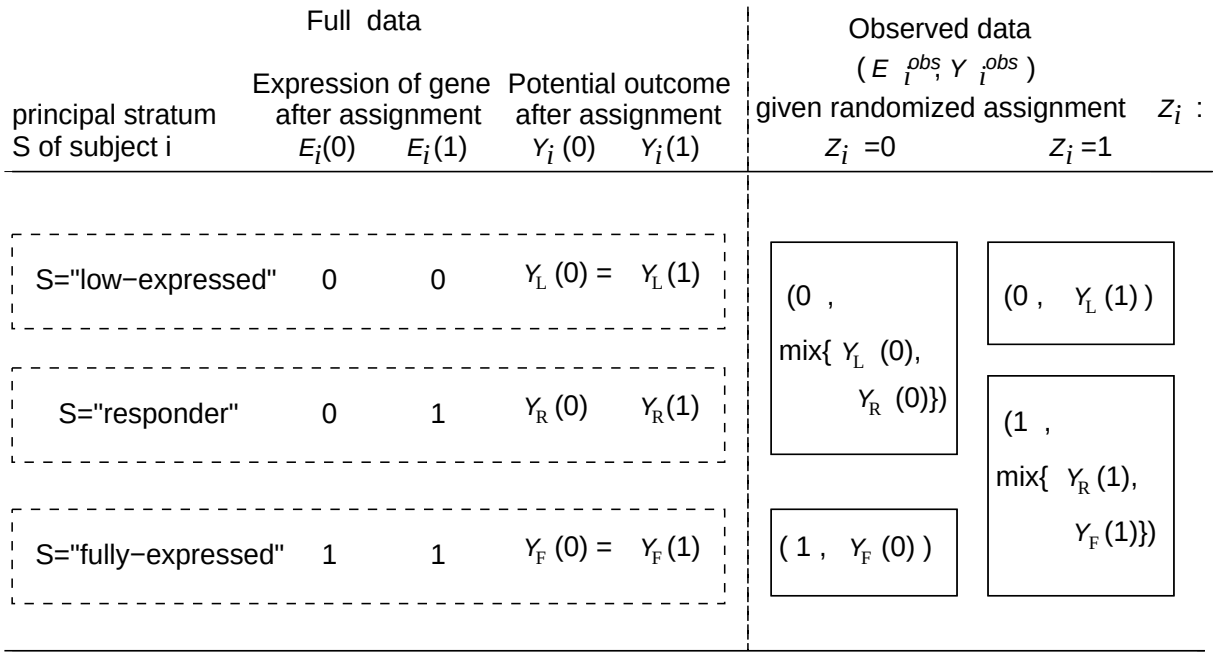

Figure 1. Template of a treatment trial using principal stratification on gene expression and clinical outcome. In the right side, “mix” indicates that observed outcome distributions, stratified by observed treatment and gene expression level, are mixtures of potential outcome distributions across principal strata of gene expression shown in the left side.