NIH Collaboratory *Rethinking Clinical Trials®*

Health Care Systems Research Collaboratory

Analyses of Randomized Controlled Trials in the Presence of Noncompliance and Study Dropout

A working document from the NIH Collaboratory Biostatistics and Study Design Core Working Group. This work is supported within the National Institutes of Health (NIH) Health Care Systems Research Collaboratory by the NIH Common Fund through cooperative agreement U24AT009676 from the Office of Strategic Coordination within the Office of the NIH Director. The views presented here are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Prepared by: Rui Wang, PhD; Fan Li, PhD; Andrea J. Cook, PhD; and Elizabeth DeLong, PhD; for the NIH Health Care Systems Research Collaboratory Biostatistics and Study Design Core
Version: 1.0, last updated July 29, 2019

Analyses of randomized controlled trials in the presence of noncompliance and study drop out

1 Background

Randomized controlled trials (RCTs) achieves the highest standard of evidence to inform decision making. The primary analysis of a RCT is often based on the "intention-to-treat" (ITT) principle, where analysis proceeds based on the treatment assignment "as randomized", as opposed to the actual treatment received. In our experience working with investigators, it has become apparent that there is confusion regarding the use of the ITT analysis, the "as-treated" and the "per-protocol" analysis, and how they relate to the noncompliance and/or missing outcome data in a RCT. One common misconception is that the intention-to-treat analysis produces unbiased estimates of treatment effect regardless of missing outcome data. "As-treated" or "per-protocol" analysis that aims to estimate the average causal treatment effect is often done without carefully considering the assumptions required. In this paper, we aim to clarify on the following issues: 1) difference between an *intention-to-treat effect* and an *average causal treatment effect* within a formal counterfactual framework [Rubin, 1978]; 2) in the presence of noncompliance and/or missing outcome data, the validity of analytic methods depends on assumptions about the missingness process.

To solidify the ideas, we use the Childhood Adenotonsillectomy Trial (CHAT) [Marcus et al., 2013] for illustration. The CHAT study was designed to evaluate the efficacy of early adenotonsillectomy versus watchful waiting with supportive care, with respect to cognitive, behavioral, quality-of-life, and sleep factors at 7 months of follow-up, in children with the obstructive sleep apnea syndrome. The primary outcome was the change in a neuro-behavioral measure of attention and executive function (NEPSY). The NEPSY scores range from 50 to 150, with 100 representing the population mean and higher scores indicating better functioning. Four hundred sixty-four children were randomly assigned to early adenotonsillectomy (surgery within 4 weeks after randomization) or a strategy of watchful waiting (control, no surgery).

2 Notation and Definitions: an ITT effect and an average causal effect

Let Y denote the change in the NEPSY score from baseline to 7 months of follow-up. Let $R = 1$ or 0 denote that a child is randomized to the surgery arm or the control arm and $A = 1$ or 0 denote that a child actually underwent surgery or not, respectively.

The counterfactual outcomes $Y^{a=1}$ and $Y^{a=0}$ are the potential changes in the NEPSY score if a child had surgery ($a = 1$) or not ($a = 0$). Similarly, the counterfactual outcomes $Y^{r=1}$ and $Y^{r=0}$ are the potential changes in the NEPSY score if a child was randomized to surgery ($r = 1$) or to control arm ($r = 0$). Clearly, not all of the counterfactual outcomes can be observed. If a child was randomized to the surgical arm, then $Y^{r=1}$ was observed and $Y^{r=0}$ was missing; If a child actually

underwent surgery, then $Y^{a=1}$ was observed and $Y^{a=0}$ was missing.

Under the counterfactual framework, the *ITT* effect can be defined in the following way:

$$\mathrm{E}[Y^{r=1}] - \mathrm{E}[Y^{r=0}] \tag{1}$$

That is, the ITT effect represents the expected difference in changes in NEPSY scores if a child had been randomized to the surgery arm compared to if the same child had been randomized to the control arm. Here we choose a mean difference scale but other measures of effect (e.g. relative measures) could have been analogously defined.

In the similar spirit, the *average causal effect* is the difference between the expected value of $Y^{a=1}$ and the expected value of $Y^{a=0}$ with respect to the study population, formally defined as the following contrast:

$$\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}] \tag{2}$$

where $\mathrm{E}[\cdot]$ denotes the mean with respect to the study population. This counterfactual contrast reflects the expected difference in changes in NEPSY scores if a child had undergone surgery compared to if the same child had not.

An equivalent way to write the average causal treatment effect (2) is

$$\mathrm{E}[Y^{a=1}|R=1] - \mathrm{E}[Y^{a=0}|R=0] \tag{3}$$

This equivalence follows because the assigned treatment R is independent of a subject's (counterfactural) outcomes due to randomization.

In the ideal situation with perfect compliance ($R = A$) and no missing outcome data, the ITT effect (1) is the same as the average causal effect (2). In practice, noncompliance and study drop outs (either loss to follow-up or study withdrawal) occur frequently. In CHAT, among the 226 subjects randomized to early adenotonsillectomy arm, 210 received assigned intervention, and 30 subjects were lost to follow-up or withdrew from study; among 227 assigned to the control arm, 16 crossed over to the surgery arm, and 23 were lost to follow-up or withdrew from study.

In what follows, we discuss the impact of noncompliance and study drops out on estimating the ITT effect and the average causal effect. We begin with the simplifying assumption that no subject drops out of the study, that is, changes in NEPSY were measured for all subjects enrolled in the study, and first consider the noncompliance issue where some subjects fail to comply with their assigned treatment; that is, there are some subjects with $R \neq A$. For simplicity, we consider a specific type of noncompliance where subjects cross-over to the alternative treatment arm. We do not consider other types of noncompliance such as the settings where subjects receive lower dosage than actually assigned, although similar issues apply. In Section 4, we consider the real-world settings where both noncompliance and study drops out occur.

3 Estimating the ITT effect and the average causal effect

3.1 Under noncompliance but no study drop out

In the presence of noncompliance, i.e., the assigned treatment R and the actual treatment A differ for some subjects, the ITT effect and the average causal effect will not generally be the same. As we now explain: In the absence of study drop out, we can validly estimate the ITT effect using standard methods ignoring noncompliance but the proper estimation of the average causal effect will require untestable assumptions about noncompliance.

3.1.1 Estimating the ITT effect

Because of randomization, the treatment indicator R is independent of a subject's potential outcomes $Y^{r=0}$ and $Y^{r=1}$:

$$Y^r \coprod R \tag{4}$$

Here we use $\coprod$ to denote the notion of independence [Dawid, 1979]. The assumption (4) can be interpreted as there is no confounding for the ITT (randomized) effect. It equivalently encodes an assumption about missingness. In our example, for a subject randomized to the surgery arm ($R = 1$), her counterfactual outcome had she been, instead, randomized to the control arm ($Y^{r=0}$) is missing. In other words, we can only observe one counterfactual outcome for a given subject – the one corresponding to the arm to which the subject was actually randomized. The assumption (4) encodes the assumption that whether or not a subject is *missing* her counterfactual outcome is not associated with the *value* of that outcome.

A key consequence of assumption (4) is that it allows us to use outcome data amongst only those subjects *not missing* Y^r to estimate the mean of Y^r in the whole study population (including those subjects for which it *is* missing). That is, it allows us to equate the ITT effect (1), which is a difference in population counterfactual means, to the following difference in observed conditional means:

$$\mathrm{E}[Y|R = 1] - \mathrm{E}[Y|R = 0] \tag{5}$$

where $\mathrm{E}[Y|R = r]$ is the mean of the outcome at 6 month follow-up amongst those randomized to arm r. It follows that, given (4), we can estimate the ITT effect (1) via an estimate of the observed conditional mean difference (5). This observed conditional mean difference can be very simply estimated by computing the difference in *sample* means of the outcome amongst those with $R = 1$ versus $R = 0$ in the data, or fitting a standard linear regression model:

$$E(Y|R) = \alpha + \beta R, \tag{6}$$

and the coefficient associated with the treatment indicator R, β, would correspond to the ITT effect. Because of randomization, the treatment assignment is guaranteed "independent" of a subject's (counterfactual) outcomes and the equivalence between the ITT effect and the data function (5) holds regardless of whether or not there is noncompliance in the study. We note that while the

maginal model (6) provides an unbiased estimate for the ITT effect, methods based on semiparametric efficiency theory ([Van Der Laan and Rubin, 2006, Tsiatis et al., 2008, Zhang et al., 2008] have been proposed to improve efficiency and also to reduce the potential baseline imbalance in covariates due to chance. These methods overcome the limitations of traditional covarites adjustment approaches that directly including covariates in the marginal model (6).

3.2 Estimating the average causal effect

In addition to the ITT effect, it is often of interest to estimate the average causal effect because this represents the effect most directly associated with the intervention under study. When there is noncompliance, $R \neq A$ for some subjects, the regression coefficient β in (6) does not represent the average causal effect.

Two commonly-used analyses that attempt to estimate the average causal treatment effects are the as-treated analysis (AT) and the per-protocol analysis (PP). The AT analysis entails fitting a model relating the outcome Y to the actual treatment received indicator A. The PP analysis restricts the comparison to those who comply with the protocol, that is, those whose actually received their randomly assigned treatment ($R = A$).

Unlike the randomized treatment indicator R, the actual treatment received A in general cannot be randomized; the investigators can not force a child to undergo surgery. Therefore, the coefficient β in the following regression model:

$$E(Y|A) = \alpha + \beta A \tag{7}$$

does not necessarily provide an unbiased estimate for the average causal treatment effect (2) unless

$$Y^a \coprod A \tag{8}$$

That is, whether or not a subject receives the surgery is independent of the values of his/her potential outcomes. The assumption (8) does not hold by design because the investigators have no control over the actual treatment received. It is an untestable assumption because, as Y^a is missing for some subjects, we cannot test for independence between Y^a and A. Under failure of this assumption, an estimate of the average causal effect based on an estimate of β will be biased – confounded by factors that contributing to whether or not a subject receives the treatment.

While we cannot avoid the untestable assumption (8), it is possible to weaken it. In particular, suppose that, the study has collected a set of baseline risk factors for the outcome (e.g., age, sex, race, etc). In what follows, we use L to denote a vector of these baseline factors. Proper analysis can proceed based on a weaker assumption as the following:

$$Y^a \coprod A|L, \tag{9}$$

That is, there is no confounding for the treatment effect after controlling for L. This is a weaker assumption than (8) because the independence of Y^a and A is only required to hold within each

level of the covariates in L rather than in the overall study population. Informally, this means that, the measured baseline factors L captures the difference between children who underwent surgery versus those who did not have surgery.

Similar to previous arguments, a key consequence of (9) is that it allows us to use the observed outcome data among those who received the treatment to estimate the mean of the study population within levels of the measured baseline confounders L. In this case, it allows us to equate the average causal effect (2), which again is a difference in counterfactual means (without regard to level of L), to a *weighted average* of differences in observed conditional means (additionally conditioned on L) with weights defined by the distribution of L, specifically:

$$\sum_l \{\mathrm{E}[Y|A=1, L=l] - \mathrm{E}[Y|A=0, L=l]\} f(l) \tag{10}$$

Here the notation $\sum_l$ reads "the sum over all possible levels of L" and $f(l)$ is the "probability that L equals one possible value l". Note the sum can more generally be written as an integral when L is a continuous variable. In this case, $f(l)$ would correspond to a density function. It follows that, given assumption (9), we can estimate the treatment effect via a sample estimate of (10).

Similarly, the PP analysis is to fit the model (7) restricting to those who comply with treatment as prescribed in the study protocol. In the PP analysis, the coefficient β measures the difference in observed conditional means:

$$\mathrm{E}[Y|A=1, R=1] - \mathrm{E}[Y|A=0, R=0] \tag{11}$$

The equivalence of the average causal effect (2) and the difference in observed conditional means (11) requires the following assumption:

$$Y^a \coprod A|R \tag{12}$$

That is, no confounding for the average causal effect within each randomized treatment arm. In other words, for a subject who was randomized to the surgery arm ($R = 1$), if he/she did not have surgery, then his/her outcome $Y^{a=1}$ is missing. But the missing mechanism does not depend on his/her outcome so that we can use the observed data from those who comply to estimate the population quantity from all.

Again, the assumption (12) does not hold by design because the investigators can not randomize the actual treatment received within each assigned treatment arm. Under failure of this assumption, an estimate of the average causal effect based on an estimate of (11) will be biased – confounded by factors that contributing to whether or not a subject complies and receives the assigned treatment. Proper analysis may proceed based on a similar weaker condition:

$$Y^a \coprod A|R, L \tag{13}$$

The assumption (13) allows us to use outcome data amongst only those subjects who comply within each arm to estimate the mean had everyone complied – but now within levels of the measured baseline confounders L. In this case, it allows us to equate the treatment effect (2), which again is a difference in counterfactual means (without regard to level of L), to a *weighted average* of a difference in observed conditional means (additionally conditioned on L) with weights defined by the distribution of L, specifically:

$$\sum_l \{\mathrm{E}[Y|R=1,A=1,L=l] - \mathrm{E}[Y|R=0,A=0,L=l]\}f(l) \tag{14}$$

As we can see, the price for the weaker untestable assumption (9) for the AT analysis (or (13) for the PP analysis) relative to (8) (or 12) is that it implies a less straightforward approach to estimating the treatment effect. That is, it is computationally more involved to estimate (14) especially when L can take on more than a few levels (which will usually be the case). To handle the realistic setting of high-dimensional L, several approaches may be used including standardization, inverse probability weighted methods or g-estimation. These methods differ only in terms of what aspects of the distribution of the measured variables are modeled. The approaches are equivalent in the special case where saturated models can always be fit (i.e. such that all models perfectly fit the data). See [Hernan and Robins, 2010], as well as [Van Der Laan and Rubin, 2006] for details on these various estimation approaches. An analysis based on any approach to estimating (14) might be called an *adjusted per-protocol analysis*, adjusted for measured baseline confounding.

Finally, we note that the need for a weighted average in (10) or (14) is due to the fact that our original question/estimand under consideration was the population level average causal effect (2) – without regard to a particular level of L – and *not* the treatment effect *within levels of* L. Were we to change our interest from (2) to

$$\mathrm{E}[Y^{a=1}|L=l] - \mathrm{E}[Y^{a=0}|L=l] \tag{15}$$

then, given the weaker assumption (9), we can estimate the treatment effect within levels of L (15) by an estimate of

$$\mathrm{E}[Y|A=1,L=l] - \mathrm{E}[Y|A=0,L=l]] \tag{16}$$

Or, given the weaker assumption (13), we can estimate the treatment effect within levels of L by an estimate of

$$\mathrm{E}[Y|R=1,A=1,L=l] - \mathrm{E}[Y|R=0,A=0,L=l]] \tag{17}$$

When L is continuous or can otherwise take many levels, the quantity $\mathrm{E}[Y|A=a,L=l]$ could be estimated via a familiar outcome regression model. However, whether to change the question of interest based on necessary untestable assumptions on confounding/missingness as a result of study flaws (here, noncompliance) should always be considered carefully in light of the original goals of the study.

In summary, when there is noncompliance and no study drop out, we can obtain an unbiased estimate of the ITT effect via standard computationally simple methods. By contrast, under the

same conditions, unbiased estimation of the average causal effect will require untestable assumptions. Under less restrictive (weaker) untestable assumptions, unbiased estimation of the treatment effect requires proper confounding adjustment.

4 Study drop-out as a challenge to estimating either effect

Thus far, we have considered how noncompliance can be understood as a particular form of missingness in the sense that it precludes our ability to observe a subject's *counterfactual* outcome had, contrary to fact, she complied (for subjects who did not comply). However, if the only source of missingness is noncompliance, we will still obtain a measurement of the observed outcome Y (even for those who fail to comply), which would be included in the AT analysis but not the PP analysis. By contrast, when there is study drop out (with or without noncompliance) we will be missing the outcome Y entirely (and, in turn, all counterfactual outcomes), for subjects who drop out. In the presence of study drop out, untestable assumptions will be required to estimate *both* the ITT effect and the average causal effect. As above, stronger versions of these assumptions will result in simpler estimation procedures while weaker/less restrictive versions will required more complex estimation procedures.

4.1 Estimating the ITT effect in the presence of both noncompliance and study drop out

With the additional complication of study drop out, estimating the ITT effect requires untestable assumptions because study drop out may induce *selection bias*. Formally, denote C as an indicator of study drop out such that, if $C = 1$ for a given subject, Y is not measured for that subject. Suppose that, in addition to (4), we made the following assumption

$$Y \coprod C|R \tag{18}$$

We can say that assumption (18) encodes the assumption that there is "no selection bias by study drop out for the ITT effect". This assumption would hold by design if we had physically randomized C (e.g. we flipped a fair coin to determine who would drop out of the study in each study arm). However, as only R (and not C) is randomized in an RCT, (18) is an untestable assumption in practice. This assumption will fail to hold, for example, if those who drop out of the study are "sicker" (or "healthier") than those who do not, within each arm of the trial.

A key consequence of the joint assumptions (18) and (4) is that they allow us to write the ITT effect as a function of only measurements from subjects with no missingness, in particular

$$\mathrm{E}[Y|C = 0, R = 1] - \mathrm{E}[Y|C = 0, R = 0] \tag{19}$$

It follows that, given (18) and (4), we can estimate the ITT effect (1) via an estimate of (19). Such an estimate is simply obtained by taking a difference in sample means amongst those who did not

drop out of the study. An analysis based on any approach to estimating (19) is called a *complete case ITT analysis*.

Again, we cannot avoid untestable assumptions in the face of missing data but we might weaken them. Specifically, we might assume the following weakened (less restrictive) version of (18)

$$Y \coprod C|R, L \tag{20}$$

This encodes the assumption that there is "no selection bias by study drop out for the ITT effect within levels of L". Hence, here, we can call L the "measured baseline selection factors". Informally, we can understand (20) as a less restrictive version of (18) because, if those subjects who do not drop out of the study are systematically "sicker' (or "healthier") than those who do drop out, (18) would fail to hold but (20) might still hold – as long as L successfully captured what makes a subject "sicker" or "healthier". Given (4) and the weaker untestable assumption on drop out (20), we can write the ITT effect as the following (more complex) function of only measurements from subjects with no missingness, in particular

$$\sum_l \{\mathrm{E}[Y|C=0, R=1, L=l] - \mathrm{E}[Y|C=0, R=0, L=l]\} f(l) \tag{21}$$

Just like the weighted average (14), the expression (21) can be estimated via standardization, IPW or g-estimation. An analysis based on any approach to estimating (21) might be called an *adjusted ITT analysis*, adjusted for measured baseline selection bias due to study drop out.

4.2 Estimating the average causal effect in the presence of both noncompliance and study drop out

Unbiased estimators for the average causal effect in the presence of both noncompliance and study drop out follow immediately from arguments in Section 3.2 for the case of only noncompliance. There we considered stronger and weaker versions of untestable assumptions about confounding in the presence of noncompliance. With the addition of study drop out we will require additional untestable assumptions about selection bias (here we will only explicitly consider weaker/less restrictive versions of such assumptions). Formally, suppose that, in addition, to (13) we assumed

$$Y \coprod C|A, R, L \tag{22}$$

The assumption (22) encodes the assumption that there is "no selection bias by study drop out for the per protocol effect within levels of L". Under both assumptions (13) and (22), we can say that L are the "measured baseline confounders *and* selection factors for drop out". Given these assumptions, we can write the treatment effect as the weighted average

$$\sum_l \{\mathrm{E}[Y|C=0, A=1, R=1L=l] - \mathrm{E}[Y|C=0, A=0, R=0, L=l]\} f(l) \tag{23}$$

Given (13) and (22), it then follows that an estimate of the treatment effect in a study with noncompliance and drop out can be obtained by any estimate of (23). Methods discussed above (standardization, IPW, g-estimation) apply here as well. An analysis based on any approach to estimating (23) might be called an *adjusted per-protocol analysis*, now adjusted for both measured baseline confounding and selection bias due to study drop out.

5 Summary

In this paper, we have reviewed the difference between an ITT treatment effect and an average causal effect which can be formally defined in terms of different counterfactual contrasts. Choice of treatment effect (the ITT or the average causal effect) should depend on the scientific goals [Shrier et al., 2014]. For randomized clinical trials with substantial noncompliance and/or loss to follow-up, it would be useful to provide estimates for both effects to understand both the population effect of the assigned treatment and the effect of actually receiving the treatment.

We reviewed untestable assumptions required for unbiased estimation of each effect. We discussed noncompliance and study drop out as two examples of missing outcome data. Under only noncompliance, assumptions required to obtain an unbiased estimate of the ITT effect will hold by design in an RCT while assumptions required for the average causal effect in general does not hold and are untestable. We discussed stronger and weaker untestable assumptions in this case and appropriate estimation procedures under each case. Finally, when study drop out is present, with or without noncompliance, unbiased estimation of both the ITT effect and the treatment effect is not guaranteed by design and requires untestable assumptions. Here we also considered stronger and weaker versions of untestable assumptions and implications for analytic approaches. Inverse probability weighting, g-estimation, and instrumental variable (IV) estimation can reduce the bias introduced [Hernán and Hernández-Díaz, 2012].

Our presentation was limited to questions about so-called time-fixed treatment effects (i.e. effects of applying one versus another level of the treatment of interest at a single point in time) and, further, untestable assumptions that give unbiased estimation under adjustment for only *baseline* confounders and selection factors. When the treatment effect of interest is time-varying (i.e. we would like to compare the effect of following one versus another treatment strategy over a period of time) unbiased estimation of the treatment effect may require appropriate adjustment for time-varying confounders. Even for a time-fixed treatment effect, if the outcome is measured "long" after baseline, unbiased estimation of either the ITT or the treatment effect may require appropriate adjustment for time-varying selection factors for study drop-out. For extensions of the above ideas to these more complex settings see for examples [Toh and Hernán, 2008], [Hernan and Robins, 2010] and [Hernán et al., 2017].

References

A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31, 1979.

Miguel A Hernán and Sonia Hernández-Díaz. Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials*, 9(1):48–55, 2012.

Miguel A Hernan and James M Robins. *Causal inference*. CRC Boca Raton, FL:, 2010.

Miguel A Hernán, James M Robins, et al. Per-protocol analyses of pragmatic trials. *N Engl J Med*, 377(14):1391–1398, 2017.

Carole L Marcus, Reneé H Moore, Carol L Rosen, Bruno Giordani, Susan L Garetz, H Gerry Taylor, Ron B Mitchell, Raouf Amin, Eliot S Katz, Raanan Arens, et al. A randomized trial of adenotonsillectomy for childhood sleep apnea. *New England Journal of Medicine*, 368(25): 2366–2376, 2013.

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

Ian Shrier, Russell J Steele, Evert Verhagen, Rob Herbert, Corinne A Riddell, and Jay S Kaufman. Beyond intention to treat: what is the right question? *Clinical trials*, 11(1):28–37, 2014.

Sengwee Toh and Miguel A Hernán. Causal inference from longitudinal studies with baseline randomization. *The international journal of biostatistics*, 4(1), 2008.

Anastasios A Tsiatis, Marie Davidian, Min Zhang, and Xiaomin Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in medicine*, 27(23):4658–4677, 2008.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Min Zhang, Anastasios A Tsiatis, and Marie Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715, 2008.