Is intent-to-treat analysis always (ever) enough?

Lewis B. Sheiner

Departments of Laboratory Medicine, Biopharmaceutical Sciences, Medicine; Schools of Medicine and Pharmacy; University of California San Francisco, Box 0626, UCSF, San Francisco, CA, 94143, USA

Keywords: hierarchical models, intention-to-treat, missing data, noncompliance, pharmacokinetic-pharmacodynamic modelling

Introduction

In a commentary written several years ago [1], Professor D. Rubin and I pointed out that one is usually advised to analyse clinical trials – the preferred modern strategy for empirical evaluation of medical therapies [2, 3] – for significance levels and estimates using only, or at least primarily, the intention-to-treat principle (see, e.g. [4]). We noted, however, that the intention-to-treat estimator (see below) estimates so-called 'use-effectiveness', the causal effect on outcome of prescribing the drug, rather than the medically more important 'method-effectiveness', the causal effect on outcome of actually taking the drug. We argued that studies should be designed so that they can yield valid estimates of method-effectiveness, although so long as protocols are not followed exactly, such estimates may depend heavily on additional assumptions (scientific models – see section on longitudinal data with dropout, below); ones that go beyond the data at hand. In this paper, I present a conceptual framework for thinking about causal estimands and estimators for both use and method effectiveness, with emphasis on the latter, focusing on the problems posed by deviations from protocol, notably noncompliance and dropout.

The paper is a review and exposition, rather than a presentation of original material. Section 2 presents a conceptual framework, setting context, defining causal estimands (population quantities describing causal effects of treatments) for use effectiveness and method effectiveness, how these change in the face of noncompliance, and the problem created by dropout; section 3 discusses estimators of the causal estimands defined in section 2; section 4 considers the effect of dropout on the estimators; and section 5 discusses implications for study design. Although different in some details, the ideas, point of view, and notation presented here are essentially those presented in the statistical literature by D. Rubin and colleagues, based on the Rubin Causal Model. Rather than provide extensive references throughout this work, I limit further citations to important additional works which discuss specific points or present examples, and offer here a minimal set of references which provide background and an entree to the relevant literature [5, 6], and a more complete and technical discussion of noncompliance [1, 7–9], dropout [10, 11], and scientific modelling of drug dose-concentration-response [12, 13].

Correspondence: Professor L. B. Sheiner, Departments of Laboratory Medicine, Biopharmaceutical Sciences, Medicine; Schools of Medicine and Pharmacy; University of California San Francisco, Box 0626, UCSF, San Francisco, CA, 94143, USA.

Presented at the 3rd Meeting on Clinical Measurement and New Drug Development, Royal Society of Edinburgh, 13–14 June 2001.

Conceptual framework

Data and design

The context is a clinical trial in which n individuals are nominally assigned either to a control treatment, coded 0, or a test treatment, coded 1, and a univariate outcome is observed. The control and test treatments are assumed to differ. The random variable indicating nominal assignment for individual i is Z_i, and the random variable denoting the individual's outcome is Y_i.

A key design feature of randomized clinical trials is 'random assignment' to nominal treatment. Since the validity of every estimator discussed in section 3 depends on this assumption, it deserves to be stated formally: it is assumed henceforth that the probability of being assigned to test (control) treatment is independent of pretrial prognosis and all other baseline patient characteristics.

Two types of deviation from protocol will be of concern to this paper, noncompliance and dropout, and I introduce next the notation required to deal with them.

Non-compliance

Non-compliance recognizes that nominal assignment, the treatment assigned by the experimenter, may differ from actual assignment, the treatment taken by the patient. The random variable indicating actual assignment is D_i; it can be written $D_i(Z_i)$ to explicitly recognize that it is a function of assignment. D_i can take the value 0, 1 or 'other', where the first two values correspond to values of Z – control (0) and test (1) – and the last indicates no particular treatment, but simply that actual treatment is unknown; some alternative is pursued. Thus, while recognizing only one 'version' of control and test treatments, I acknowledge the possibility of multiple versions of the alternative. In general, Y will be a function of D (and hence Z). Let $Y_i(d)$ denote the *potential outcome* (see below) for individual i actually receiving treatment $D_i = d$. To maintain parallelism with dropout (below), I introduce the (redundant) random variable $U_i(z) := I(D_i(z) = z)$, where $I(\cdot)$ is the indicator function, taking the value 1 if its logical argument is true, and 0 if it is not, to designate whether the intended outcome is observed ($U_i = 1$) or not ($U_i = 0$).

I assume that the compliance tendency of each individual in the circumstances of the trial is a (unknown) baseline variable, V_i, which determines the values of $U_i(z)$ according to Table 1.

Dropout

Dropout, as a problem for analysis and interpretation of a clinical trial herein refers to a data set that is incomplete due to the loss of the outcome datum from one or more subjects. The random variable indicating responder status is R_i. It takes the value 1 if the outcome of individual i is observed, and 0 if it is not. Just as with D_i, R_i can be a function of assignment: $R_i(z)$ is the potentially observable variable indicating whether or not individual i will drop out on assignment to treatment z. Similarly to V_i, the responder tendency of each individual in the circumstances of the trial is a (unknown) baseline variable, W_i, which determines the values of $R_i(z)$, as in Table 2.

Table 1 Compliance behaviour V.

V	*Type*	*U(0)*	*U(1)*
a	Always-complier (complies with any assignment)	1	1
t	Test-only complier (only complies when $z_i = 1$)	0	1
c	Control-only complier (only complies when $z_i = 0$)	1	0
n	Never-complier (never complies)	0	0

Table 2 Responder Behaviour W.

W	*Type*	*R(0)*	*R(1)*
a	Always-respond (never drops out)	1	1
t	Test-only responder (drops out only if $z_i = 0$)	0	1
c	Control-only responder (drops out only if $z_i = 1$)	1	0
n	Never-responder (always drops out)	0	0

Potential data

For each individual in the trial, there are 11 potentially observable quantities, Z_i, V_i, W_i, and $D_i(z)$, $U_i(z)$, $R_i(z)$, and $Y_i(z)$, for $z = 0,1$, although there are only 7 independent choices, as $D_i(z)$ fully determines both $U_i(z)$ and V_i, and $R_i(z)$ fully determines W_i. I refer to the $N \times 11$ matrix of potentially observable data as the potential data matrix, or just the potential data. In any actual trial, we observe, for example, only one of $Y_i(D_i(0))$ or $Y_i(D_i(1))$ (these might or might not be different see Table 1), depending on whether z_i is 0 or 1. The Rubin Causal Model claims that even if the control treatment was assigned and we observed $Y_i(D_i(0))$, nonetheless the test treatment *could have* been assigned, and in that case we *would have* observed $Y_i(D_i(1))$. Counterfactuals, that is unobserved potential outcomes, such as outcomes on unassigned treatments, are central in the Rubin causal model, as the model defines causal estimands (population quantities describing causal effects) using them.

Although the idea of potential outcomes seems natural and intuitive to most persons, caution must be taken when they are used as a basis for causal reasoning. The only kinds of potential outcomes that may meaningfully feature in the definition of the causal relationships that a clinical trial might establish are those that would result from decisions made *after* study inception and before study termination[1]. Clearly, the 'decision' to assign test or control treatment to an individual could be altered after enrolment, so that the potential outcome on the control treatment for a type a complier assigned to the test treatment is certainly a meaningful concept. A bit less certainly so are the potential outcomes on both treatments for a type *n* complier: if through added inducements and reminders incorporated into a modified study design, one could imagine altering his 'decision' not to comply with any assigned treatment, then they may be meaningful, but if not, not. However, there is no possibility of altering a possible pre-existing correlation, perhaps a genetic-cultural link, between a propensity to behave as a type *n* complier and more rapid drug metabolism, leading to reduced sensitivity to the test treatment if it were taken. Thus, the type *n* complier's potential outcome on the test treatment if he were a type a complier is not a meaningful potential outcome, and hence causality in the sense of the Rubin causal model cannot be associated with complier type.

Observed data

Of the 11 potentially observable variables for any one individual, we always observe three, $z_i = Z_i$, $d_i = D_i(z_i)$, and $r_i = R_i(z_i)$. If $r_i = 1$, then we also observe $y_i = Y_i(d_i)$.

[1] A strict view would limit potential outcomes to only those outcomes consequent on actions assigned *randomly* by the experimenter, but I will not so limit the discussion herein.

The $N \times 3$ or 4 matrix of observed data is called the actual data matrix or the actual data.

Note in particular that, contrary to usual practice, d_i is assumed to be observed: modern methods of monitoring compliance make this assumption realistic [14]. The characteristics of such methods, in fact, determine the definition of 'actual assignment' I have used in the face of noncompliance (and in this way, this work differs from that of the other authors cited previously): modern compliance monitoring methods record medicine container openings. They can therefore indicate that an individual is not taking the prescribed medication as instructed, but if they do so indicate, they do not indicate what other treatment he is taking. Thus, as noted above, the actual assignment category – 'other' denotes some version of alternative treatment, but does not specify what it is. This view is particularly applicable, for example, when $D_i(z) = z$ is defined as $>80\%$ of prescribed doses taken correctly, as is commonly done. In that case an individual who takes less than 80% (but more than 0%) of the test treatment when $Z_i = 1$ and less than 80% (but more than 0%) of the control treatment when $Z_i = 0$ is of compliance type n, and for him by definition both $D_i(1) =$ 'other' and $D_i(0) =$ 'other'; yet the two treatments are clearly not the same, nor, in general, would they be expected to lead to the same outcome.

Missing data

Both noncompliance and dropout can be viewed as particular instances of missing data; data that were scheduled to be observed but were not in fact observed. Here, those data are $Y_i(z_i)$ for those i designated as having missing outcomes by an observation indicator as follows. If noncompliance, but not dropout is present, i such that $u_i^1 \neq 1$ are missing; if dropout, but not noncompliance is present, i such that $u_i^1 \neq 1$ are missing; and if both are present, i such that $r_i\ u_i^1 \neq 1$ are missing.

Missing data may be classified by the relationship between the observation indicator and the rest of the potential data. Depending on the class, missing data present minor (loss of precision) or major (validity) problems. Data are Missing Completely At Random (MCAR) if the missingness is independent of all other data. Data are Missing At Random (MAR) if the missingness depends only on other *observed* data and not at all on unobserved or *potential* data, including the missing data themselves. So-called nonignorable[2] missing data are neither MAR nor MCAR, and this is the class that presents serious problems. Note that it may be the case that missingness depends unconditionally on unobserved data, but not on such data after conditioning on the observed data. In that case the missingness is MAR, not MCAR. In such a case, of course, for the method of analysis of the MAR data to be valid, it must correctly condition on the observed data. Thus, although MAR data can be validly analysed, such analyses are more sensitive to modelling assumptions than are analyses based on complete data or data that are only MCAR.

Both noncompliance and dropout can separately cause nonignorable missingness, and this is why they present a problem for clinical trial interpretation. A simple example makes the problem with nonignorable missingness obvious: Consider a randomized trial of a new drug *vs* placebo for a chronic disease, with outcome being a measure of disease severity after a fixed time on treatment. If patients drop out (fail to comply) on the test treatment because of toxic side-effects, and if greater toxicity is associated with greater efficacy, then the best potential outcomes on the new drug in the group assigned to take it will be missing from the actual outcomes in that group due to the dropout (noncompliance). Therefore a natural estimator of drug effect, the difference in average disease severity between treated and controls *among those providing outcome data (among those on the assigned treatment)* will be downwardly biased relative to the value it would have had if dropout (noncompliance) had not been present. The missing severity measures (outcomes) are nonignorable because their probability of being missed is related to their value (through their association with toxicity) *within treatment groups*; that is, after controlling for the observed data (assignment).

Causal estimands

This discussion focuses on population-level causal estimands, that is, expected values of population characteristics. However, for simplicity, and because it emphasizes ideas rather than technique, I assume henceforth that the trial is large enough so that I can focus on valid estimation without worrying about small-sample variability. Given this assumption, trial-level causal estimands defined on averages over *all* N rows of the potential data matrix are valid (unbiased) estimators of the corresponding population estimands, and for this reason, in an abuse of notation, no distinction is made between the two.

Before proceeding further, I define some functions that will prove useful. Let A and B be the vectors $(a_1, a_2, \ldots, a_i, \ldots, a_N)$ and $(b_1, b_2, \ldots, b_i, \ldots, b_N)$,

[2] This term has a technical origin. Strictly speaking ignorable missing data is a fourth type of missing data. Such data are MAR, but also the Distinct Parameters (DP) assumption holds: the missingness mechanism (probability distribution of the observation indicator) shares no parameters with the data mechanism. The term 'ignorable' stems from the fact that with probability model-based estimators and ignorable missingness, the model for the missingness mechanism need not appear in the full probability model for the estimator to be valid and efficient; the missingness can thus be 'ignored'. We will not consider this detail further: MAR alone is sufficient to justify estimator validity; the DP assumption affects only precision.

respectively. The elements of the former are logical, in which case the symbol '$=$' means 'is equal to'; the latter is algebraic, in which case the symbol '$=$' means 'is assigned the value.' Let:

- $p(A) := N^{-1}\sum_{i=1,N} I(a_i)$ denote the fraction of the study population for which a_i is true; and
- $\bar{Y}(A,B) := \frac{\sum_{i=1,N} I(a_i)Y_i(b_i)}{\sum_{i=1,N} I(a_i)}$ denote the average outcome with actual assignment b_i over those individuals in the study population for whom a_i is true.

For convenience, I define the logical condition A that includes all study participants ($A=\{i\in 1{:}N\}$) in the expressions above to be the default condition, signified by the symbol '*all*'. Further, where it will cause no confusion as an argument in the expressions above, in an abuse of notation, I let a_i denote A and b_i denote B, so that for example, $\bar{Y}=(all,D_i(1))=\sum_{i=1,N} Y_i(D_i(1))$.

Use effectiveness

The intention-to-treat estimand, Average Causal Effect

$$ACE := \bar{Y}(all,D_i(1)) - \bar{Y}(all,D_i(0)) \tag{2.1}$$

is the expected value of the difference between outcome with nominal assignment to the test treatment and with nominal assignment to control. It describes use effectiveness in the population of interest, the one from which the trial participants are drawn.

Method effectiveness

Because estimating method effectiveness is less often an explicit goal of clinical trials than use effectiveness, and because the main purpose of this paper is to emphasize its importance and discuss how it can be validly estimated, I take the opportunity here to discuss at greater length why it is an important causal estimand.

Method effectiveness – the expected outcome difference due to treatment with a new drug *vs* a standard alternative (often placebo) – is almost certainly a more important population pharmacological characteristic for purposes of treating individual patients than is use effectiveness – the expected outcome difference due to prescribing the new drug *vs* the standard – for a number of reasons.

Extrapolation. Method effectiveness is a biological quantity, rather than a combination of biological and behavioural quantities. As such, one can more confidently extrapolate conclusions regarding its magnitude from current trial participants to future patients – the sole justification for insisting on clinical trials before approving the use of a drug – than conclusions that mix pharmacology and behaviour. Behaviour is far more dependent on conditions specific to a particular study, conditions that may not characterize future use.

Understanding. As a biological estimand, method effectiveness provides both qualitative and quantitative insight. Qualitatively, for example, if a trial fails (i.e. the data do not support a conclusion that use effectiveness is necessarily greater than zero), the data may still be compatible with method effectiveness, indicating that the trial may have failed because of noncompliance or dropout rather than because of lack of intrinsic efficacy. Quantitatively, method effectiveness for a specific pharmacological endpoint that is quantitatively compatible with *in vitro* values for that endpoint (e.g. antimicrobial MIC) adds a 'proof-of-principle' component to trail interpretation.

Sufficiency for drug approval. The unequivocal demonstration of method effectiveness is sufficient to satisfy the efficacy requirement for drug approval. There is neither a theoretical nor a legal requirement to demonstrate use effectiveness.

Necessity for dosing. A use-effectiveness dose–response relationship describes the expected efficacy and toxicity of a drug averaged over rates of compliance in a particular trial. It is hard to see how this can be used to choose a dose for a new patient. In contrast, an argument can be made that a rational starting dose for a new patient is one that is likely to yield desirable efficacy without excessive toxicity if taken as prescribed. A method-effectiveness dose–response relationship provides the basis for choosing such a dose.

Exclusive emphasis on designing studies solely so that ITT analyses will be valid, unfortunately still the norm, precludes attention to design features that allow valid estimates of method effectiveness. I return to this point later.

Returning now to method effectiveness estimands, note first that if all individuals were always-compliers, the obvious method effectiveness estimand would be ACE. Because, however, the pair (D(1), D(0)) of actual treatments taken in response to the two nominal assignments can be different for all 4 types of compliers, 4 distinct method effectiveness estimands may be considered, one for each compliance group. Note that similar logic does not apply to the 4 types of responders: (pure) nonresponse does not affect actual treatment, so that this subsection need only concern itself with the distinct method effectiveness estimands induced by noncompliance.

Within a group of like-compliers, the obvious method effectiveness estimand is ACE restricted to that group, $ACE^v := \bar{Y}(v_i=v,D_i(1)) - \bar{Y}(v_i=v,D_i(0))$. ACE as defined in (2.1) can be expressed as a function of these group-specific ACE values:

$$\begin{aligned}ACE = {} & p(v_i=a)ACE^a + p(v_i=t)ACE^t \\ & + p(v_i=c)ACE^c + p(v_i=n)ACE^n\end{aligned}$$

Of least interest are ACE^n and ACE^c: the former is uninteresting because it does not necessarily involve outcomes on either actual treatment chosen for study; the latter is of interest only if the control treatment itself is of intrinsic interest, and if the alternative treatment chosen by c-type individuals when assigned to the test treatment in fact differs from the control treatment.

The two remaining method effectiveness estimands ACE^a and ACE^t (hereafter $ACACE$ and $TACE$) are of interest: Under the reasonable assumption that the alternative treatment taken by type t compliers when assigned to control is never the test treatment (likely to be true, for example, when access to the test treatment is limited to those assigned to it, as is common in new drug testing – see below), both estimands compare expected outcomes under the test treatment to some nontest alternative. $ACACE$ is arguably of greater interest than $TACE$ as it compares the test treatment to the control treatment, which presumably was chosen for the trial for a good reason.

Perhaps the practically most relevant method effectiveness estimand is a weighted average of $ACACE$ and $TACE$ called the Compliance Average Causal Effect

$$CACE := \frac{p(v_i=a)ACACE + p(v_i=t)TAC}{p(v_i=a) + p(v_i=t)}. \qquad (2.2)$$

It is the expected difference in outcome associated with a difference in nominal assignment in that subpopulation which, under the conditions of the trial, will take the test treatment when assigned to it. If, as above, test-only compliers never take the test treatment when they are assigned to the control then the subpopulation that $CACE$ describes is composed of all those individuals who will take the test treatment under trial conditions *if and only if* they are assigned to it. Henceforth I regard $ACACE$ and $CACE$ as the sole method effectiveness estimands of interest.

Valid causal estimators

Estimands are defined on potential data, whereas estimators are limited to actual (observed) data. For simplicity, I extend the 'large sample' assumption for the potential data made in the previous section to apply to (the number of outcomes observed in each assignment group of) the actual data. Having done so, discussion henceforth can be limited to finding estimators that are unbiased (valid) for their estimands.

For simplicity, I consider first valid estimators of ACE, $CACE$, and $ACACE$ in the absence of missing data, and then reconsider the same estimators in its presence.

Use effectiveness in the absence of dropout

The actual data counterpart of ACE,

$$\widehat{ACE} := \bar{Y}(z_i=1,d_i) - \bar{Y}(z_i= -0,d_i),$$

is an unbiased estimator of ACE (the 'hat' notation will be used to denote estimators henceforth) because of the random assignment – the sample average outcome of a subgroup is an unbiased estimator of the group average outcome if the subgroup is chosen independently of outcome.

Method effectiveness in the absence of dropout

Just as for $\widehat{ACE}$ itself, randomized assignment assures that (if v_i were known) the counterparts of its components, e.g. $AC\widehat{ACE} = \bar{Y}(v_i=a \vee z_i=1,d_i) - \bar{Y}(v_i=a \vee z_i= 0,d_i)$, and $\hat{p}(v_i=a)$, would be valid estimators of the corresponding causal estimands, and hence $CACE$ would be validly estimated by substituting estimators into formula (2.2). The problem, however, is that v_i is not known. The remainder of this subsection considers how $CACE$ and $ACACE$ might validly be estimated despite this.

CACE

Substituting the definition of ACE into (2.2) yields

$$CACE = \frac{ACE}{p_{at}} - \frac{ACE^c + AC}{p_{at}}, \qquad (2.3)$$

where $p_{at} := p(v_i=a \vee v_i=t) = p(v_i=a) + p(v_i=t)$. Because of random assignment and the fact that test-compliance is observed in test-assigned individuals, p_{at} is validly estimated by

$$\hat{p}_{at} := \frac{p(z_i=1 \wedge d_i=1)}{p(z_i=1)}.$$

Hence, if $(ACE^c + ACE^n)$ were known, substituting $\widehat{ACE}$ and $\hat{p}_{at}$ into the first term in (2.3) and subtracting the known value of $(ACE^c + ACE^n)/\hat{p}_{at}$ would yield a valid estimator of $CACE$.

Instrumental variables (IV) estimator

The IV estimator [7]

$$IV := \widehat{ACE}/\hat{p}_{at}$$

validly estimates $CACE$ under the Instrumental Variables assumption

(IV): $ACE^c = ACE^n = 0$.

Note that the IV assumption implies neither $\bar{Y}(v_i=c, D_i(1)) = \bar{Y}(v_i=n, D_i(1))$, nor $\bar{Y}(v_i=c, D_i(0)) = \bar{Y}(v_i=n, D_i(0))$; it does imply $\bar{Y}(v_i=v, D_i(1)) = \bar{Y}(v_i, D_i(0))$ for both $v=c$ and $v=n$. These will hold in the special type of clinical trial discussed by Zelen [15, 16] that meets the following conditions (henceforth denoted Zelen's conditions):

1. Compliance with test treatment is all or none;
2. Everyone who does not take the test treatment takes the control treatment;
3. The test treatment is not available to control-assigned individuals.

Zelen's conditions are likely to be met in practice if, corresponding to each similarly numbered condition, all of the following hold: (1) Treatment is one-off; for example a single vaccine injection, or a one-time surgical procedure; (2) The control treatment is a broadly defined 'standard of care', including whatever individuals generally do for the condition defining study eligibility; and (3) The test treatment is difficult to imitate, and cannot be obtained without authorization. Sommer & Zeger [17] provide an instructive example.

The effect of Zelen's conditions is to make all alternative treatments the same and identical to the control treatment, and hence to make complier types *c* and *n* (and, separately, *a* and *t*) indistinguishable even on potential outcomes. Thus only two types of compliers remain, types *a* and *n*; the latter always take control, and the former always take control unless assigned to test. Since the never-compliers so defined always take the same treatment regardless of assignment, it is hard to imagine how their outcome could possibly depend on that (blinded) assignment. If it does not, then the i.v. assumption holds[3]. Note also that when the i.v. assumption holds, strictly speaking, there is no need to observe compliance, so long as a valid estimate of p_{at} is available, perhaps from some other study.

Using a covariate that predicts compliance

Observation of u_i identifies test treatment compliers in the group assigned to test (i.e. types *a* and *t*). If somehow these same complier types could also be identified in the group assigned to the control treatment, *ACE* limited to just such individuals, i.e. $\bar{Y}(z_i=1 \wedge u_i=1, d_i) - \bar{Y}(z_i=0 \wedge (v_i=a \vee v_i=t),$ would validly estimate *CACE*.

Of course, the key data – compliance behaviours under unassigned treatments – are missing, so that there is no way to distinguish *a* and *t* compliers from *n* and *c* compliers in the control-assigned group (u_i sorts them instead into *a* and *c vs n* and *t*), but a covariate, X_i may be available that is observable in all and, if not an infallible indicator of compliance type (or at least *a vs c*), is at least highly correlated with it. One covariate that is commonly used in this way is u_i itself, despite the fact that it does not do the job without further assumptions. (Efron & Feldman [18] provide an example of such an assumption and an associated procedure for identifying the appropriate control subjects to compare with the compliers among the test-assigned; unfortunately, the clinical trial data cannot provide evidence that the assumption is correct).

If u_i is used (without additional assumptions) to select control-assigned subjects to compare to complying treatment-assigned subjects, this defines the per-protocol estimator, to be discussed in the next section. Using u_i in this way is justified for estimating method effectiveness if one can assume that the average outcome on control is the same for *t* and *c* type compliers. But doing so invokes the causally dubious estimand $\bar{Y}(v_i=t,0)$, which imagines that control noncompliers could somehow be induced to take the control treatment. Of course, under Zelen's conditions, they do, but then also, and without any additional assumptions, the i.v. estimator can be used. Perhaps a better way to use d_i when Zelen's conditions are not met is to compare $\hat{p}_{at}$ to $\hat{p}_{ac} := p(z_i=0 \wedge u_i=1)/p(z_i=0)$: if the two are very close, it can be argued that since it is unlikely that compliance groups *c* and *t* would be exactly the same non-null size by coincidence, perhaps neither group exists. If not, the IV estimator can be used. Less usefully, but more justifiably perhaps, a large discrepancy between the two estimated probabilities might be used to rule out use of the IV estimator. Another possibility is to expose all subjects to the test treatment for a short pretrial period and to use the observed compliance in that period as the compliance-predicting covariate (of course, it would be more efficient to use the pretrial compliance as an eligibility criterion for the subsequent study, rather than as a covariate in the final analysis [19]).

ACACE

Non-confounding compliance

A natural estimator of method effectiveness, mentioned in the last section, is the Per-Protocol estimator,

$$PP := \bar{Y}(u_i=1 \wedge z_i=1, d_i) - \bar{Y}(u_i=1 \wedge z_i=0, d_i),$$

which deletes all noncompliers with the test treatment from the test-assigned group and all noncompliers with the control treatment from the control-assigned

[3] Under the Zelen conditions, the assumption is known as the 'exclusion restriction' assumption. This name comes from econometrics, the field in which the IV estimator was first used. It is so named because the assumption 'excludes' an effect of compliance type on outcome except as it determines the treatment actually received.

group, and then contrasts the mean outcome in the two assignment groups over the remaining individuals.

Given random assignment, *PP* validly estimates

$$PP := \frac{p(v_i=a)\bar{Y}(v_i=a,1) + p(v_i=t)\bar{Y}(v_i=t,1)}{p_{at}} - \frac{p(v_i=a)\bar{Y}(v_i=a,0) + p(v_i=c)\bar{Y}(v_i=c,0)}{p_{ac}}$$

where $p_{ac}=p(v_i=a)+p(v_i=c)$. As it stands, *PP* is an estimand with no particular method effectiveness credentials. The reason is that although the *PP* definition correctly contrasts average outcomes between groups of individuals receiving the two treatments being evaluated, it does so in two different patient groups, *a* plus *t*-type compliers taking the test treatment *vs a* plus *c*-type compliers taking the control treatment. As long as the average outcomes are not the same for each pair of complier types taking the same treatment, a nonzero value of *PP* can be caused by the treatment difference *or* a type difference. That is, it is possible that $ACACE = TACE = 0$ and hence $CACE=0$, and yet $PP \neq 0$ because, for example, $\bar{Y}(v_i=a,1) \neq \bar{Y}(v_i=t,1)$.

Compliance type is acting here as a 'confounder', a term from epidemiology describing a covariate linked to outcome via (at least) two causal paths (see, for example [20]), one indirectly, mediated by its effect on the 'level' of a (observed) causal variable of primary interest (*D*), and at least one other, either direct or indirect, not involving that variable. Comparing groups with different values of *D* rather than *Z* ensures that the groups also differ in their distributions of *V*, and then one cannot distinguish which difference causes the observed difference in outcome. However, given the Non-Confounding Compliance assumption

$$(\mathrm{NCC}): \bar{Y}(v_i=a, D_i(1)) = \bar{Y}(v_i=t, D_i(1))$$

and

$$\bar{Y}(v_i=a, D_i(0)) = \bar{Y}(v_i=c, D_i(0)),$$

that compliance type does not affect average outcome given identical actual treatments, then deleting the noncompliers from each treatment group before analysis creates treatment groups that behave just as if they were composed of always-compliers only, and it is easy to see that *PP* estimates $ACACE$ (and, under NCC, $CACE$ as well, as then the two are identical).

Note that the deleted data (the noncompliers' responses) are nonignorably missing if the average outcome when assigned to control differs between *t* and *c* compliers and/or if $ACE^n \neq 0$: *PP* is invalid for ACE even under NCC! This is simply to say that use effectiveness and method effectiveness are not necessarily identical, even under the NCC assumption.

Non-confounding within strata of a prognostic covariate

The NCC assumption may be difficult to justify globally. However, it may be less difficult to justify within strata of a prognostic covariate that correlates with Y_i independently of V_i: V_i can affect outcome given identical treatment only through a causal path involving some other nontreatment prognostic factor; to the extent that results are stratified on such prognostic factors, the confounding will be diminished. Thus, given a prognostic covariate *PP* is computed within strata to estimate the within-strata $ACACE$. A regression setting can be used to accomplish the same thing with a continuous-valued prognostic covariate. Prognostic covariates are central to the methods discussed in the next section for dealing with dropout.

Longitudinal data with dropout

Under certain restricted conditions it is possible to make progress in the presence of dropout without relying on covariates or modelling. For example, if dropout and noncompliance are the same (that is, subjects either comply and complete, or don't comply and dropout), and if NCC holds or a covariate inducing nonconfounding can be found, then *PP* validly estimates $CACE = ACACE$. Alternatively, Frangakis & Rubin [9] showed that under Zelen conditions, the IV assumption, and the additional assumption that dropout for never-compliers is not related to assignment, a valid estimator for ACE can be found.

Restrictive conditions are not always met; this section discusses a more general approach than either of the above. It relies, as already mentioned, on adding a prognostic 'covariate' to transform the nonignorable data into MAR data, just as a prognostic covariate is used to transform confounded data into nonconfounded data in the presence of noncompliance. Intuitively, the covariate unbiasedly predicts outcome so that the residual missing data (the difference between the predictions and the missing values) are almost MAR, and hence, one can proceed as above using the observed data augmented by the predictions[4]. Unfortunately, under standard 'endpoint-only' designs one cannot generally count on having a baseline prognostic covariate powerful enough to render the missingness MCAR within strata (and hence MAR in the data set as a whole). In contrast, the serial response data gathered in longitudinal studies can often serve in this role, and the rest of this discussion therefore focuses on such data.

[4] This 'single imputation' approach is adopted here for heuristic purposes; a full probability-model based analysis uses the probability model for the missing data in a more principled way, and avoids the inferential problems with single imputation, discussed, for example, by Rubin [21].

Longitudinal data

A longitudinal study is one in which serial observations of response are scheduled to be observed on each subject, and the 'outcome' to be compared between assignment groups is a function of those serial responses. In the simplest case, and the one considered here, the outcome variable is simply the last measured value of the serial response. For example, if a symptom severity index (SSI) is filled out and scored at monthly visits after randomization, the outcome might be the score at a predesignated endpoint visit $T^\star$ (say 1-year). Dropout then means that there is a 'last' visit time T_i for each individual such that that all responses are available prior to and including that time[5], but none thereafter. The value of R_i then depends on the relationship of T_i to $T^\star$: if $T_i = T^\star$ the individual did not drop out; if $T_i < T^\star$, he did.

The longitudinal design means that the outcome Y_i is now the last entry in an individual longitudinal response vector $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ...,)$. Let Y^{obs}denote the $Y_{iT^\star}$ actually observed, and Y^{miss} denote those that are missing.

Achieving ignorability by modelling the missing outcomes

One way to render ML estimation valid in the face of nonignorable missingness is to introduce, if possible, a good model for the missing outcomes. What is meant by good, and how such a model can help, is easily understood if we imagine, as at the beginning of this section, that the missingness will be dealt with by filling in (but see footnote 4) the missing values $Y^{miss} = (Y_1^{miss}, Y_2^{miss}, ..., Y_N^{miss})$ with imputed values $Y^{imp} = (Y_1^{imp}, Y_2^{imp}, ..., Y_N^{imp})$ and then proceeding with the analysis using (Y^{obs}, Y^{imp}) as if it were the actual (complete) data vector Y. Clearly, if each Y_i^{imp} provides an unbiased estimate of Y_i^{miss}, then the analysis using (Y^{obs}, Y^{imp}) instead of Y, whilst it may be less efficient than that originally planned, will be valid. The predictions will almost surely depend on the 'covariate' Y_c^{obs}, that is the serial response actually observed, and may require as well, other covariates.

Truly unbiased imputation can in principle be obtained if missingness depends on the modelled response, e.g. on a true physiological state, and not on its possibly imprecisely measured observed value. If it depends on the observed value itself, e.g. on the patient's stated mood, not on an imprecise model for it, then the imputation model will still be somewhat biased. Even if the latter condition holds, however, if predictions are reasonably accurate, the degree of damage the missingness can now do is much less than before, as the only missing data is the difference between the imputation and the observation, not the full value of the observation itself.

Last Observation Carried Forward

Last Observation Carried Forward (LOCF) is a very popular imputation strategy. It imputes the missing Y_i at time $T^\star$ for (dropped out) individual i as the last observed value at time T_i. Unfortunately this 'model' does not often make the missingness ignorable. Most responses that evolve over time follow a smooth (say increasing) trajectory, so that the last observed response before T is almost surely a (downwardly) biased estimate of the true last value. If the response were a SSI for a progressive chronic disease, say arthritis, the LOCF 'model' would say that once a patient drops out, his SSI stays fixed at the last observed value despite the fact that it usually rises monotonically, in accord with the natural history of the disease. For most chronic conditions there is generally no reason to believe that SSI and disease would miraculously cease to progress the day a patient drops out of a study (indeed, were it so, dropping out would be a valuable treatment!). The nonignorability here almost certainly conservatively biases any causal estimator computed from the imputed data: if the test treatment slows the rate of progression, and if progression past a certain point induces dropout, then more dropout will occur in the placebo group because their rate of progression is unaltered. LOCF will impute these unobserved high SSI values to be the (lower) values last observed. Whilst this conservatism may make an analysis using LOCF more acceptable to consumers than if it induced anticonservative estimates (which would lead to apparent efficacy of a possibly worthless treatment), neither the magnitude nor the direction of the bias can always be guaranteed, and in any event, the estimate is not valid for method effectiveness.

Scientific models

Models that incorporate scientific knowledge are the only ones that can consistently produce unbiased imputations. They might be as simple as using the knowledge that SSI progression is approximately linear to justify estimating missing values from a linear regression of the observed data on time for each subject who drops out, or they might be as complex as a full physiological pharmacokinetic/pharmacodynamic model, cast in an hierarchical statistical model framework. The key point is that, obviously, the adequacy of the model for the missing data cannot be tested on the observed data: credibility for predictions must therefore rest *entirely* on the external (scientific) knowledge that justifies them.

[5] Allowing for a more general 'nonmonotone' missingness pattern in which occasional visits are missed prior to time T_i does not require any new concepts, although it does complicate certain methods of data analysis (but we do not plan to discuss these in detail anyway); we avoid sporadic missingness here simply because it complicates notation.

Conclusions

This review began with the observation that method effectiveness – average outcome difference due to taking a test treatment *vs* a control – is an important causal estimand on which to base therapeutic decisions (e.g. whether to use the treatment, and if so, at what intensity). The intention to treat estimator of treatment effect validly estimates use effectiveness (but only if any missing data are ignorable), an estimand primarily of value for public policy decisions (e.g. regulatory decisions and whether to publicly finance the use of the treatment), not method effectiveness.

Clinical trials, if perfectly executed answer both questions with the same ($\widehat{ITT}$) estimator. However, trials may be marred by deviations from protocol, notably some patients failing to comply with the prescribed treatment, and the same or other patients dropping out before the study endpoint can be observed. Both of these deviations mean that the intention-to-treat use-effectiveness estimator no longer validly estimates method effectiveness (and in the case of dropout, it may not even validly estimate use effectiveness). Depending on study circumstances, other estimators may be available that validly estimate important method effectiveness estimands. It should therefore be a priority to design studies so that the conditions permitting such estimators to be valid are met.

In general, other than the obvious features of randomized and blinded assignment, and encouragement of complete follow-up, the following additional design features increase the likelihood that one of the method effectiveness estimators discussed herein will be valid.

Zelen's conditions: (1) treatment is one-off, (2) the control treatment is 'standard of care', and (3) the test treatment is available only to those for whom it is prescribed.

Compliance is measured in all patients for the duration of the trial.

The study is longitudinal, and the 'endpoint' is a function of the serially measured response(s).

Prognostic covariates (and, if available, those predicting compliance with test treatment) are measured at baseline and serially.

In addition, for data seriously marred by deviations from protocol, scientific model-based analyses rather than simple comparisons of summary statistics may be required to achieve unbiased estimates of causal estimands. Such estimates will necessarily be sensitive to scientific assumptions that cannot be verified on the study data themselves. Models and assumptions should be prespecified and discussed in the study protocol to avoid accusations of *posthoc* 'data dredging'.

References

1 Sheiner LB, Rubin DB. *Intention to treat analysis and the goals of clinical trials. Clin Pharmacol Ther* 1995; **57**: p. 6–15.

2 Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials*, 2nd edn. 1985. Boston: PSG Inc.

3 Pocock SJ. *Clinical Trials: A Practical Approach* 1983. New York: John Wiley & Sons.

4 Lee YJ, *et al*. Analysis of clinical trials by treatment actually received: is it really an option? *Stat Med* **10**: 1595–1605.

5 Holland PW, Rubin DB. Causal inference in retrospective studies. *Evaluation Review* 1988; **12**: 203–231.

6 Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health* 2000; **21**: 121–145.

7 Imbens GB, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* 1997; **25**: 305–327.

8 Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Amer Stat Assoc* 1996; **91**: 444–472.

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

10 Little RJA, Rubin DB. *Statistical analysis with missing data* 1987. New York: Wiley & Sons.

11 Diggle P, Kenward MG. Informative dropout in longitudinal data analysis. *Appl Statis* 1994; **43**: 49–93.

12 Sheiner L, Wakefield J. Population modelling in drug development. *Statist Meth Med Res* 1999; **8**: 183–193.

13 Sheiner LB, Steimer J-L. Pharmacokinetic/pharmacodynamic modelling in drug development. *Annu Rev Pharmacol Toxicol* 2000; **40**: 67–96.

14 Urquhart J. Role of patient compliance in clinical pharmacokinetics. A review of recent research. *Clin Pharmacokin* 1994; **27**: 202–215.

15 Zelen M. A new design for randomized clinical trials. *New Engl J Med* 1979; **300**: 1242–1245.

16 Zelen M. Randomized consent designs for clinical trials: an update. *Statistics in Medicine* 1990; **9**: 645–656.

17 Sommer A, Zeger SL. On estimating efficacy from clinical trials. *Stat Medical* 1991; **10**: 45–52.

18 Efron B, Feldman D. Compliance as an explanatory variable in clinical trials (with discussion). *J Amer Stat Assoc* 1991; **86**: 9–26.

19 Sheiner LB, Rubin DB. Intention-to-treat analysis and the goals of clinical trials. *Clin Pharmacol Ther* 1995; **57**: 6–15.

20 Greenland S, Pearl J, Robins J. Causal diagrams for epidemiological research. *Epidemiology* 1999; **10**: 37–48.

21 Rubin DB. Multiple imputation after 18+years. *J Amer Stat Assoc* 1996; **91**: 473–489.