Published in final edited form as:

Clin Trials. 2009 April ; 6(2): 136–140. doi:10.1177/1740774509103868.

The calibration of treatment effects from clinical trials to target populations

Constantine Frangakis, PhD

Department of Biostatistics, Johns Hopkins University, 615 N Wolfe St, Office E3642 Baltimore MD 21205, USA, cfrangak@jhsph.edu

Standard practice teaches that randomized clinical trials should be preferred to observational studies if the aim is to obtain unbiased estimates of efficacy. It is often acknowledged that treatment efficacy might differ in the general population than in the clinical trial, but the source of those differences is often poorly understood or not regarded as important. Substantive differences between RCT and observational study results are typically ascribed to confounding of the latter.

Recent literature [1,2] suggests that differences between the results seen in an RCT and in an observational study can be larger than expected not because of confounding, but because of differences between their respective patient populations. These efficacy estimates can differ enough to change treatment recommendations. The paper by Weisberg *et al.* in this issue [3] is a very clear and innovative example of such emerging literature. But there is another important message that one should derive from this work. That is, RCT results are useful only if we can calibrate their results to predict treatment efficacy in the target population of interest. This editorial makes the case for why calibration requires both clinical knowledge from observational studies, and new statistical insights.

Calibration based on pre-treatment exclusion criteria

To do such a calibration we need to first record pre-treatment factors that we know can differ between the RCT and the target patient population. These factors can exist by design, as with exclusion criteria discussed by Weisberg *et al.* or exist unintentionally, as with factors driving patients to volunteer for the RCT. We can use these factors to predict the effect that would be experienced in the target population. For example, suppose the target population has a 50 : 50 mix of patients with and without a family history that increases the risk of suicide under standard treatment. Suppose further that in order to reduce the number of patients with high risk of suicide in a clinical trial, we restrict enrollment based on family history, producing an RCT with a 20% prevalence of the risk factor. We can then predict the effect of the treatment on the target population by applying the efficacy results we see in the 20% subgroup with family history in the RCT to the 50% with family history in the target population, and by applying the results we see in the 80% without family history in the RCT to the 50% without family history in the target population. This is a direct way to address the problem raised by Weisberg *et al.* for differences due to exclusion criteria used by physicians, because such criteria are based on factors that are measured.

The need to calibrate post-treatment measurements and mediators

Differences between the RCT participants and the target population can also manifest in variables measured after treatment. An aim of this editorial is to show why calibrating such differences is important, and why this calibration is substantially more challenging and requires new methods.

To calibrate such differences, clinical knowledge needs to be used to determine which RCT components are generalizable to the patient population. When it comes to post-treatment measures, such as adherence, the problem is that standard methods are deficient because they can only generalize components that are not causal effects of treatment. Yet causal effects are of most interest for estimation because they are arguably most generalizable. Methods should thus be developed to predict the outcomes in a target population, by calibrating it to the RCT on certain key causal effects. The meaning of this can be clarified only if we become more specific about certain fundamental concepts, as we do below. Such predictions can be markedly different and more plausible than standard predictions, as also shown below.

To describe what such methods must do, we will use the example of an RCT with noncompliance (or more generally, any mediator of the treatment effects). This RCT compares assignment to old and new treatments, labeled with 'Z' and the outcome Y is survival at 1 year (Figure 1(a)). However, after assignment to a treatment, some patients do not comply, and actually receive either the old or the new treatment. To be concrete, we will describe the RCT in terms of potential outcomes [4,5] and let:

- i. $D_i(z)$ be the treatment that patient i will receive if assigned treatment z .
- ii. $Y_i(z)$ be the survival outcome (0/1) if patient i is assigned treatment z .
- iii. Z_i be the treatment a patient gets actually assigned;
- iv. D_i^{obs} be the treatment a patient actually receives, i.e., $D_i(Z_i)$; and
- v. Y_i^{obs} be the observed survival, i.e., $Y_i(Z_i)$.

As we can see, some of the above characteristics cannot be observed, namely the treatment received and the outcome that would have been measured if patient i had been assigned to a treatment different than the one (namely, Z_i) they were actually assigned. Consequently, the RCT result can be represented based either on the observed data, or on the more fundamental, partly unobserved data. This difference can impact how one does calibration to a new study, so we briefly summarize here these two ways of representation.

(a) Standard (observed) conditional distributions to represent the experiment

$$\text{pr}(Y^{\text{obs}}, D^{\text{obs}}|Z) = \text{pr}(D^{\text{obs}}|Z) \times \text{pr}(Y|D^{\text{obs}}, Z) \quad (1)$$

This equation says that the joint probability of the observed survival and received treatment, given the assigned treatment, is the product of the probability of the received treatment

given the assigned treatment and the probability of the survival, given the received treatment and the assigned treatment (which may not be the same).

(b) Principal causal effect distributions to represent the experiment

To describe this, we first classify patients into three strata denoted by S :

- S = ‘Never-takers’: patients who would not take the new treatment in the RCT, no matter the assignment ($D_i(\text{old}) = D_i(\text{new}) = \text{old}$)
- S = ‘Always-takers’: patients who would take the new treatment in the RCT, no matter the assignment ($D_i(\text{old}) = D_i(\text{new}) = \text{new}$)
- S = ‘Compliers’: patients whose taking of treatment would agree with their assignment no matter what that would be, ($D_i(\text{old}) = \text{old}$ and $D_i(\text{new}) = \text{new}$).

We assume there are no ‘defiers’, i.e., no patients who would do the opposite no matter the assignment, an assumption called monotonicity [5].

These groups parallel Weisberg *et al.*’s four groups, in the sense that the membership of a patient to each such group does not change with actual assignment [6,7]. Here, however, these groups are critical also because they form strata in which to define true causal effects on survival. For this reason, these strata are known as ‘principal strata’ [8]. Principal strata allow us to define causal effects that account for compliance. For example, the comparison between:

$$\Pr(Y(z=\text{old})|S=\text{'complier'}) \text{ vs } \Pr(Y(z=\text{new})|S=\text{'complier'}) \quad (2)$$

is the only experimental comparison for which the contrast between *assignment* to $z = \text{old}$ and $z = \text{new}$ is precisely the same as the contrast between actually *receiving* the old versus the new treatment.

The observed data and the principal strata are connected, in the sense that:

$$\Pr(Y^{\text{obs}}, D^{\text{obs}}|Z) = \text{a function of } \Pr(S) \text{ and } \Pr(Y(z)|S) \quad (3)$$

This equation says that the probability of the observed outcome and received treatment, as a function of assigned treatment, is a function of the distribution of principal strata and of the survival probability within principal strata. Figure 1(a) shows this relation. The right side of the figure denotes the observed data: among patients assigned the new treatment ($z = \text{new}$), 95% complied (i.e., $\Pr(D^{\text{obs}} = \text{new} | Z = \text{new}) = 95\%$), and had 34% survival ($\Pr(Y^{\text{obs}} = \text{alive} | D^{\text{obs}} = \text{new}, Z = \text{new})$). Among patients assigned the old treatment ($z = \text{old}$), 65% complied (i.e., $\Pr(D^{\text{obs}} = \text{new} | Z = \text{old}) = 35\%$), and had 15% survival ($\Pr(Y^{\text{obs}} = \text{alive} | D^{\text{obs}} = \text{old}, Z = \text{old})$). Assuming the exclusion restriction, the distributions of $\Pr(S)$ and $\Pr(Y(z) | S)$ (left side of Figure 1(a)) can be deduced from the observed data (right side of Figure 1(a)). The figure shows that 60% of the original population were compliers, and that in this subgroup the probability of survival is 15% if assigned to the old treatment versus 25% if assigned the new treatment. (i.e., $\Pr(Y(\text{old}) = 1 | S = \text{complier}) = 15\%$, whereas $\Pr(Y(\text{new}) = 1 | S = \text{complier}) = 25\%$).

Methods that can allow better calibration of post-treatment variables (mediators)

The preceding framework now allows us to explain why the distinction between representations (1) (standard conditional distribution) and (3) (distributions based on principal strata) implies that there are different ways of generalizing results from the RCT to a study in the target population. In this target study, we assume we have an arm that assigns a large number of people to the new treatment, and an arm that originally assigns a smaller number of people to the standard treatment, to measure compliance in the target population. Here, distinguish the distributions of principal strata and of outcomes given principal strata between the RCT in which we observe all data (Figure 1 (a)) and the target study (Figure 1(b)), respectively:

$$\text{RCT: } \begin{array}{cc} \text{Principal strata} & \text{Outcome distributions} \\ \text{pr}^{\text{RCT}}\{D(\text{old}), D(\text{new})\} & \text{pr}^{\text{RCT}}\{Y(z)|D(\text{old}), D(\text{new})\} \end{array} \quad (4)$$

$$\text{Target: } \text{pr}^{\text{target}}\{D(\text{old}), D(\text{new})\} \quad \text{pr}^{\text{target}}\{Y(z)|D(\text{old}), D(\text{new})\} \quad (5)$$

The left sides of (4) and (5) are principal strata, because principal strata as defined in (b) are simply another way to describe a value of both $D(\text{old})$ and $D(\text{new})$.

Because we see all the data in the RCT, and only compliance data but not the outcome data in the target study, all distributions above are estimable empirically except $\text{pr}^{\text{target}}\{Y(z) | D(\text{old}), D(\text{new})\}$.

There is often interest to predict the outcomes Y_i^{obs} in the target study before waiting to measure them. This could be done based on their predictive distribution, denoted by $\text{pr}^{\text{target}}\{Y^{\text{obs}} | D^{\text{obs}}, Z\}$. Because the distributions (5) determine the distributions of all observable data in that study, we have (under randomization):

$$\text{pr}^{\text{target}}(Y^{\text{obs}} | D^{\text{obs}}, Z=z) = \frac{\int \text{pr}^{\text{target}}\{Y(z) | D(\text{old}), D(\text{new})\} \times \text{pr}^{\text{target}}\{D(\text{old}), D(\text{new})\} dD^{\text{mis}}}{\int \text{pr}^{\text{target}}\{D(\text{old}), D(\text{new})\} dD^{\text{mis}}} \quad (6)$$

where D^{mis} are the missing potential values of $D(z)$. Readers not interested in the integral can take (6) to simply mean that we could predict the outcome from its correct distribution if we knew the proportions of principal strata, and the distribution of the outcome within principal strata. Without waiting for any outcome Y^{obs} , however, the correct predictive distribution in (6) is not available because $\text{pr}^{\text{target}}\{Y(z) | D(\text{old}), D(\text{new})\}$ is not available. To address this, the standard approach predicts the outcomes Y^{obs} in the target study using the predictive distribution from the RCT, $\text{pr}^{\text{RCT}}\{Y^{\text{obs}} | D^{\text{obs}}, Z\}$, effectively replacing in (6) both distributions of (5) with those of (4). This is represented in Figure 1(a), and the right panel of (b), by the lines connecting the observed data of the two figures.

The problem with this approach is that the target study can differ from the RCT in *either* the distribution of principal strata or the potential outcomes given principal strata, in which case

the RCT predictive distribution will be incorrect for the target study. This may help to explain empirical evidence that regressions $\text{pr}^{\text{RCT}}\{Y^{\text{obs}} | D^{\text{obs}}, Z\}$ in one RCT can be quite different in another study with the same type of outcome and assigned and received treatment [9].

To address this, consider, alternatively, replacing only the outcome component in the right side of (6) with that of the RCT, to obtain the synthetic predictive distribution defined as,

$$\text{pr}^{\text{SYNTHESIS}}(Y^{\text{obs}} | D^{\text{obs}}, Z=z) = \frac{\int \text{pr}^{\text{RCT}}\{Y(z) | D(\text{old}), D(\text{new})\} \times \text{pr}^{\text{target}}\{D(\text{old}), D(\text{new})\} dD^{\text{mis}}}{\int \text{pr}^{\text{target}}\{D(\text{old}), D(\text{new})\} dD^{\text{mis}}} \quad (7)$$

By any measure, it is more likely that the left side of (5) equals the left side of (4) (i.e., the principal strata distributions are equal) than it is that both the right side *and* the left side of (5) equal, respectively, those in (4) (i.e., both the principal strata and outcome distributions are equal). This suggests that using the synthetic predictive distribution (7) should be a more plausible approximation to the correct predictive distribution in the target study, than the predictive distribution from the RCT.

The connections between Figure. 1(a) and (b) demonstrate this new calibration method and compare them to the standard one. The standard method assumes the survival $\text{pr}^{\text{RCT}}\{Y^{\text{obs}} | D^{\text{obs}}, Z\}$ in the observed compliance strata generalizes from the RCT to the target study. Accounting for the differences in compliance distributions, this method predicts that if the new treatment is implemented in the target population, the overall resulting survival would be 26%. The new method, in contrast, generalizes the distributions $\text{pr}^{\text{target}}\{Y(z) | D(\text{old}), D(\text{new})\}$, whose contrasts between assignments z are causal effects. The new method predicts that if the new treatment is implemented in the target population, the overall resulting survival would be 21%, and not 26%.

In practice, a study in a target population is most often an observational one. Of course, issues arise about how to choose reasonable clinical and statistical assumptions for such calibration. But the alternative – to just assume full generalizability of the RCT to the clinical practice in the target population – is also an assumption, which as we see, can be wrong and lead to bad decisions if observational data are not used for calibration.

RCTs should continue – like a rhythm in a musical piece – to serve as a model of rigor. And observational studies should continue – like melodies – to be free to explore and make conjectures about generalizations or explanations, to be further subject to testing by experimentation. And as with the whole of a musical piece, complete knowledge of clinical effects needs to draw on both experimental and observational information.

References

1. Hernan MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. 2008; 19:766–779. [PubMed: 18854702]
2. Dehejia HR, Wahba S. Causal effects in nonexperimental studies: reevaluating the evaluation of training Programs. *Journal of the American Statistical Association*. 1999; 94:1053–1062.

3. Weisberg HI, Hayden V, Pontes V. Selection criteria and generalizability within the counterfactual framework: explaining the paradox of antidepressant-induced suicidality? *Clin Trials*. 2009; 6:109–118. [PubMed: 19342462]
4. Rubin DB. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*. 1974; 66:688–701.
5. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*. 1978; 6:34–58.
6. Imbens G, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*. 1996; 25:305–327.
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. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002; 58:21–29. [PubMed: 11890317]
9. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Annals of Internal Medicine*. 1996; 125:605–613. [PubMed: 8815760]

Figure 1.

An example of calibrating treatment effects using post-treatment measures from an RCT (a) to a study in the target population (b) using two methods. The first method, represented in the left panel of (b), borrows from (a) the survival probabilities in the principal strata (thus generalizing causal effects); the second method, represented in the right panel of (b), borrows from (a) the survival probabilities on the strata of observed assignment arm and treatment received. Dashed boxes represent principal strata; solid boxes represent strata defined by assignment arm (old/new in *italics*) and treatment actually received (old/new in **bold**)