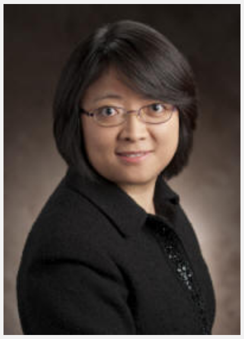
RESEARCH ARTICLE

HETEROGENEITY IN TREATMENT EFFECT AND COMPARATIVE EFFECTIVENESS RESEARCH

Zhehui Luo, PhD, Michigan State University

ABSTRACT

The ultimate goal of comparative effectiveness research (CER) is to develop and disseminate evidence-based information about which interventions are most effective for which patients under what circumstances. To achieve this goal it is crucial that researchers in methodology development find appropriate methods for detecting the presence and sources of heterogeneity in treatment effect (HTE). Comparing with the typically reported average treatment effect (ATE) in randomized controlled trials and non-experimental (i.e., observational) studies, identifying and reporting HTE better reflect the nature and purposes of CER. Methodologies of CER include meta-analysis, systematic review, design of experiments that encompasses HTE, and statistical correction of various types of estimation bias, which is the focus of this review.



Zhehui Luo, M.S., Ph.D.
Assistant Professor of Epidemiology, Division
of Biostatistics, Michigan State University,
Department of Epidemiology
B601 West Fee Hall,
East Lansing, Michigan 48824,
Tel: 517.353.8623x161 Email: zluo@msu.edu

INTRODUCTION—THE RELEVANCE OF CER IN HEALTH CARE REFORM

The American Reinvestment and Recovery Act boosts the funding for Comparative effectiveness research (CER), which is deemed to hold significant promise to improve health care quality. Nevertheless, it faces some significant methodological challenges in fulfilling this promise. By definition, CER is “the conduct and synthesis of research comparing the benefits and harms of different interventions and strategies to prevent, diagnose, treat and monitor health conditions in ‘real world’ settings” (Federal Coordinating Council 2009). The objective of CER is to aid decision-makers to make evidence-based decisions to improve health care at both the *population* and *individual* levels. Thus, as a tool for CER, traditional randomized controlled trials (RCTs) are limited by poor generalizability. Observational studies in “real world” settings suffer from unmeasured confounding bias. Meta-analyses are complicated by heterogeneity of study populations and interventions. Some efforts have been made to overcome these methodological limitations of CER, such as innovative trial designs, propensity scores or instrumental variables methods for observational data analyses, and Bayesian meta-analyses.

The ultimate goal of CER is to predict “which interventions are most effective for which patients under specific circumstances”, which coincides with the goal of the federal initiative for personalized health care. A broad definition of personalized medicine includes “all factors that distinguish an individual’s health characteristics and risks, including family history, clinical data, behavioral factors, and genomics when applicable” (Snyderman and Dinan 2010). In the era of personalized medicine, to directly address the goal of CER, there is a call for CER methodologies that can estimate the treatment effect at the individual patient level (Basu 2009).

The need for individualized CER (*i*CER) is most prominent for diseases with heterogeneous treatment effect (HTE). The development of *i*CER methodology will help to clarify which forms of evidence are sufficiently informative for a specific patient (Garber and Tunis 2009).

CURRENT CER METHODS AND FRAMEWORKS RELATED TO ICER

A Framework for Causal Inference

The main problem with empirical analyses of treatment response, in RCTs or observational studies, is the non-observability of counterfactual outcomes. Studies of CER aim to predict outcomes that would occur if different treatment strategies were applied to a population. The Neyman-Rubin-Holland tradition, or the Rubin Causal Model (RCM), is becoming the mainstay of causal inference in both RCTs and observational studies. Denote the potential outcome by $Y_i(A)$ if individual i is exposed to treatment alterative A. In a two-arm RCT, an intervention (A=1) is compared with placebo or control condition (A=0). A person is randomly assigned to one of the alternatives so that one of the $Y_i(A)$ is not observed. Individual level causal effects are defined by $Y_i(1) - Y_i(0)$. Random assignment leads to the direct estimate of $E[Y_i(0)]$ for the treated group using observed outcomes in the control group. The RCM allows for general heterogeneity in treatment responses and for direct handling of complications, such as noncompliance with assigned treatment. The treatment assignment mechanism is a stochastic rule governing the actual receipt of an alternative. It can be a known or an unknown function independent of potential outcomes, or a function dependent on potential outcomes. Different treatment assignment mechanisms require different analytical methods for estimating treatment effects.

In observational studies, treatments are rarely randomly assigned. Thus corrections for overt and hidden biases have to be made under certain assumptions to produce valid inference on causal effects. Much of the recent development in the statistics literature for estimating average treatment effects (ATE) and bias correction builds on the work by Rubin, Rosenbaum and others (Rubin 1973; Rosenbaum and Rubin 1983; Heckman, Ichimura et al. 1998). Applicable in a special case, variously referred to as unconfoundedness, exogeneity, ignorability, or selection on observables, these methods can be grouped into five categories: (i) estimating the unknown regression functions of the outcome; (ii) matching on covariates; (iii) propensity score methods; (iv) combination of the above; and (v) Bayesian methods (Imbens 2004).

A parallel literature in econometrics for program evaluation was developed, with a focus on issues of endogeneity, or self-selection, when the unconfoundedness assumption is violated. Individuals who choose to enroll in a program are, by definition, different from those who choose not to enroll. Without unconfoundedness, five general categories of estimation methods have been summarized by Imbens and Wooldridge (2009). These methods are: (a) sensitivity analyses; (b) bound analyses; (c) instrumental variables (IV) with exogeneity and exclusion restrictions; (d) regression discontinuity designs; and (e) a set of methods referred to as difference-in-differences.

Estimand and Hypothesis Testing

In any evaluation study, the primary objective of estimation must be made explicit. Table 1 summarizes the estimands of traditional CER in the simple case with two treatment options. Well-conducted RCTs estimate average treatment effects (ATEs) between treated and control groups. If ATEs can be generalized to the population it becomes the population ATEs (PATEs, τ_{PATE}). The PATE provides answers to policy questions such as "what would happen if entire population was subject to the intervention as compared to what would happen if the entire population was not subject to the intervention?" Such global estimates may not always be of interest. A second quantity of interest that has received much recent attention is the average treatment effect on the treated (ATT), which is the mean effect for those who actually were treated. Population ATTs (PATTs, τ_{PATT}) are useful when an intervention is not applicable to the entire population. When estimates are adjusted by persons' characteristics (X_i), such as women aged 65 and above with histories of depression, ATEs and ATTs become conditional ATEs (CATEs) and conditional ATTs (CATTs). When a pre-specified subgroup is the target of interest, estimates are CATEs in a subpopulation (CATE-S). Marginal q-th quantile treatment effect (MQTE-q, τ_q) is the difference between $Y_i(1)$ and $Y_i(0)$ at the q-th quantile of the marginal distributions $F^{-1}_{Y(1)}(q)$ and $F^{-1}_{Y(0)}(q)$. This quantity is particularly useful for continuous measures with varying treatment effects by the level of the outcome. Imbens and Woodridge (2009) point out issues of interpretation for this estimate. First, unless there is perfect rank correction between the potential outcomes, the two marginal distributions will not be the same. Second, the MQTE-q will not be the same as unit level q-th quantile treatment effect (UQTE-q, $\tilde{\tau}_q$), which is based on the joint distribution of the potential outcomes $F^{-1}_{Y(1)-Y(0)}(q)$. The problem with UQTE-q is that $\tilde{\tau}_q$ is not point identified without assumption of the rank correlation between the potential outcomes even in RCTs. Imbens and Angrist (1994) define a local ATE (LATE, τ_{LATE}) that is estimable using IV under very weak conditions. For example, in the case where the treatment indicator, A_i, is binary and a valid instrument, Z_i, is binary, they show that, under the assumption of monotonicity, the estimated treatment effect, τ_{LATE}, is for the subpopulation that is affected by the instrument. However, this subpopulation is generally not identifiable. In this example, potential treatment options received, $A_i(0)$ and $A_i(1)$, depends on the particular instruments used for estimation ("compliers" in their nomenclature, whose treatment choice can be "manipulated" based on the values of the instrument).

Hypothesis testing in CER concentrates on the null hypothesis that the estimand of interest is zero. In cases where τ_{PATE} is zero, it may be of interest to test whether there is positive treatment effect in a subpopulation, by observable covariates, or by the quantiles of the potential outcomes. This is highly relevant to *i*CER because, to make a treatment decision for a specific patient, a clinician needs to have evidence of the efficacy of alternative options based on a sample of other individuals for whom both outcome and covariate information are available before s/he can apply it to a new patient. Such information is not summarized by τ_{PATE}. Addressing the uncertainty of such decision rules is a new and growing literature.

Table 1. Estimands of Interest in CER

Estimand	**Definition**
Population Average Treatment Effect (PATE)	$\tau_{\text{PATE}} = E[Y_i(1) - Y_i(0)]$
PATE on the Treated (PATT)	$\tau_{\text{PATT}} = E[Y_i(1) - Y_i(0) \mid A_i = 1]$
Conditional Average Treatment Effect in the sample (CATE)	$\tau_{\text{CATE}} = \frac{1}{N}\sum_{i=1}^{N} E[Y_i(1) - Y_i(0) \mid X_i]$
CATE on the Treated in the sample (CATT)	$\tau_{\text{CATT}} = \frac{1}{N_1}\sum_{i \mid A_i = 1} E[Y_i(1) - Y_i(0) \mid A_i = 1, X_i]$
CATE in a subpopulation (CATE-S)	$\tau_{\text{CATE-S}} = \frac{1}{N_S}\sum_{i : X_i \in S} E[Y_i(1) - Y_i(0) \mid X_i]$
CATE on the Treated in a subpopulation (CATT-S)	$\tau_{\text{CATT-S}} = \frac{1}{N_S}\sum_{i \mid A_i = 1 : X_i \in S} E[Y_i(1) - Y_i(0) \mid A_i = 1, X_i]$
Marginal q-th quantile treatment effect	$\tau_q = F^{-1}_{Y(1)}(q) - F^{-1}_{Y(0)}(q)$
Unit level q-th quantile treatment effect	$\tilde{\tau}_q = F^{-1}_{Y(1)-Y(0)}(q)$
Local Average Treatment Effect (LATE)	$\tau_{\text{LATE}} = E[Y_i(1) - Y_i(0) \mid A_i(0) = 0, A_i(1) = 1]$

Sources of Heterogeneity in Treatment Effect (HTE)

Responses to treatment for an individual or a group may depart from the population average due to differences in susceptibility, vulnerability to adverse side effects and utilities for different outcomes (Kravitz et al, 2004). Heterogeneity in treatment responses is the magnitude of the variation of individual or group treatment effects across a population. By estimating this variation, a clear distinction can be drawn on whether one treatment is superior to the other for everybody in the population or whether one treatment is superior to the other for an *average* or typical person in the population.

Adaptive Treatment Strategy and Time-varying Covariate

Robins pioneered estimation of causal effect of a time-varying treatment in observational studies (Robins 1986). He and his colleagues introduced three methods under sequential ignorability: the inverse-probability-of-treatment weighted (IPTW) estimator, g-estimation of structural nested models, and the iterative conditional expectations estimator (Hernán, Brumback et al. 2001). Extensions of these estimators to cases with unmeasured confounder or censoring have been proposed (Brumback, Hernán et al. 2004; Bryan, Yu et al. 2004). In recent years, attention has shifted to identifying optimal treatment sequence in an adaptive experiment. Murphy developed an iterative minimization method based on the dynamic programming or backwards induction principle (Murphy 2003). Robins does so using g-estimation in structural nested mean models (Robins 2004). Moodie et al. demonstrate that these two methods are closely related (Moodie, Richardson et al. 2007) and Dawson and Lavori argue that a sequential Bayesian predictive inference method is more efficient than the marginal mean approach (Dawson and Lavori 2008).

Counterfactual Choice Probabilities and Decision Theories

McFadden initiated the parametric random utility models to describe observed discrete-choice behavior and to predict the choices that a person would make in counterfactual choice settings (McFadden 1974). Since then much work was done to advance the field to include more tractable models with less severe preference assumptions (Train 2003). A distinct and innovative literature builds on the basic structure of discrete choice models using observed choice probabilities to partially infer the distribution of types of persons in the population (Manski 2007; Hirano and Porter 2009). The results are then used to predict behaviors in unrealized choice settings. This literature is very much in progress and relevant to *i*CER because it studies the problem faced by a clinician or an experimental designer: a clinician or patient has to make a choice given limited information derived from another population.

Pragmatic Clinical Trial

Clinical trials for which "the hypothesis and study design are developed specifically to answer the questions faced by decision makers are called pragmatic or practical clinical trials (PCTs)" (Tunis, Stryer et al. 2003). Different from RCTs, PCTs compare practical interventions, include diverse populations in diverse practice settings, and collect data on a broad range of health outcomes. The former is explanatory to aid our understanding of an intervention; and the latter is pragmatic to aid our decision on preferred intervention for a patient, i.e., PCTs are designed to meet the needs of decision makers. NIMH funded several large-scale PCTs, including STAR*D for depression, STEP-BD for bipolar disorder, and CATIE for schizophrenia and Alzheimer's disease (Insel 2006).

After two international meetings on PCT design issues, a new tool, PRECIS, to assist trialists in making design decisions that are consistent with their stated purpose has been developed (Thorpe, Zwarenstein et al. 2009). PRECIS evaluates 10 domains of trial design with graphics to determine how generalizable results of a trial are. These 10 domains are: participants, flexibility of the intervention under evaluation, intervention practitioner expertise, comparison intervention, comparison intervention provider expertise, follow-up intensity, primary outcome, participant compliance/adherence, practice fidelity, and analysis of outcome. We believe the tool can also be used to evaluate the generalizability of an observational study.

DISCUSSION

Not all methodologies in CER are relevant to *i*CER because the two frameworks take on different perspectives: the former of a population and the latter of an individual. However, the methods reviewed above are related to *i*CER. Many challenges for CER are germane to *i*CER because many of the threats to external validity in CER are due to individual unobserved confounders. In the future, researchers need to build on these methods and address (a) how to predict treatment responses while taking into consideration the patient's preference and choice; and (b) what the empirical use of the *i*CER framework is in real world practice.

More fundamental, perhaps, is the issue of appropriate theoretical framework for *i*CER. Most medical research on treatment response has been focused on testing the null hypothesis of zero average treatment effect (even in pre-specified subgroup analysis, the idea is the same). If researchers, clinicians and policy makers wish to inform treatment choice for a given individual or group, they should not view statistical insignificance as a reason to refrain from studying heterogeneity in treatment response. In other words, they should not treat the null and alternative hypothesis asymmetrically, fixing the probability of a type I error and seeking to minimize the probability of a type II error. Instead, they must be concerned with the quantitative variation of outcomes with treatments and covariates. Hypothesis testing simply does not address this problem.

REFERENCES

Basu, A. (2009). "Individualization at the Heart of Comparative Effectiveness Research: The Time for i-CER Has Come." Medical Decision Making 29(6): NP9-NP11.

Brumback, B. A., M. A. Hernán, et al. (2004). "Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures." Statistics in Medicine 23(5): 749-767.

Bryan, J., Z. Yu, et al. (2004). "Analysis of longitudinal marginal structural models." Biostatistics 5(3): 361-380.

Dawson, R. and P. W. Lavori (2008). "Sequential causal inference: Application to randomized trials of adaptive treatment strategies." Statistics in Medicine 27(10): 1626-1645.

Federal Coordinating Council (2009). Report to the President and the Congress on Comparative Effectiveness Research. Washiongton DC.

Garber, A. M. and S. R. Tunis (2009). "Does comparative-effectiveness research threaten personalized medicine?" New England Journal of Medicine 360(19): 1925-1927.

Heckman, J. J., H. Ichimura, et al. (1998). "Matching as an econometric evaluation estimator." Review of Economic Studies 65(2): 261-294.

Hernán, M. A., B. Brumback, et al. (2001). "Marginal structural models to estimate the joint causal effect of nonrandomized treatments." Journal of the American Statistical Association 96(454): 440-448.

Hirano, K. and J. R. Porter (2009). "Asymptotics for statistical treatment rules." Econometrica 77(5): 1683-1701.

Imbens, G. W. (2004). "Nonparametric estimation of average treatment effects under exogeneity: A review." Review of Economics and Statistics 86(1): 4-29.

Imbens, G. W. and J. D. Angrist (1994). "Identification and estimation of local average treatment effects." Econometrica 62(2): 467-475.

Imbens, G. W. and J. M. Wooldridge (2009). "Recent Developments in the Econometrics of Program Evaluation." Journal of Economic Literature 47(1): 5-86.

Insel, T. R. (2006). "Beyond efficacy: the STAR*D trial." Am J Psychiatry 163(1): 5-7.

Kravitz, R. L., Duan, N. and J. Braslow (2004). "Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages." The Milbank Quarterly 82(4): 661-687.

Manski, C. F. (2007). "Partial identification of counterfactual choice probabilities." International Economic Review 48(4): 1393-1410.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. Frontiers in Econometrics. P. Zarembka. New York, NY, Academic Press.

Moodie, E. E. M., T. S. Richardson, et al. (2007). "Demystifying optimal dynamic treatment regimes." Biometrics 63(2): 447-455.

Murphy, S. A. (2003). "Optimal dynamic treatment regimes." Journal of the Royal Statistical Society Series B-Statistical Methodology 65: 331-355.

Robins, J. M. (1986). "A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect." Mathematical Modelling 7(9-12): 1393-1512.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. Proceedings of the Second Seattle Symposium on Biostatistics. K. Y. Lin and P. Haeagerty. New York, NY, Springer.

Rosenbaum, P. R. and D. B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." Biometrika 70(1): 41-55.

Rubin, D. B. (1973). "Matching to Remove Bias in Observational Studies." Biometrics 29(1): 159-183.

Snyderman, R. and M. A. Dinan (2010). "Improving health by taking it personally." JAMA 303(4): 363-364.

Thorpe, K. E., M. Zwarenstein, et al. (2009). "A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers." J Clin Epidemiol 62(5): 464-475.

Train, K. (2003). Discrete Choice Methods with Simulation. Cambridge, UK, Cambridge University Press.

Tunis, S. R., D. B. Stryer, et al. (2003). "Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy." JAMA 290(12): 1624-1632.