Combining randomized trial data to estimate heterogeneous treatment effects

Laura A. Hatfield[1], Daniel B. Kramer[2,3,4], and Sharon-Lise T. Normand[1,5]

March 23, 2015

Abstract

Heart failure arises, progresses, and responds to therapy differently in different people. Yet clinical trials often lack power to estimate treatment effects for subgroups, or enforce eligibility criteria that exclude some patients entirely. Combining information across trials increases power for subgroup estimates and expands generalizibility. However, naively pooling patient-level data sacrifices the benefits of randomization, and pooling study-level estimates must consider trial heterogeneity.

We develop and illustrate approaches for combining information across trials to estimate effects in men and women with heart failure who are treated with implantable cardioverter-defibrilliator (ICD) alone or in combination with cardiac resynchronization therapy (CRT-D). We consider individual- and trial-level factors that may confound or mediate subgroup treatment effects. For example, ischemic disease is more common in men; could this explain why women appear to benefit more from CRT-D than men?

Our Bayesian models estimate sex-specific treatment effects across trials, accounting for uncertainty, confounding, and mediation. We find that with a very small number of heterogeneous studies, hierarchical modeling offers few benefits over conventional effect pooling, producing wider credible intervals but little shrinkage. We also find little evidence for residual confounding within subgroups, but some evidence of interactions between left bundle branch blockage and ischemic etiology in the sex-specific treatment effects, suggesting further study.

Acknowledgments

LAH and SLN are supported by contract DHHS/FDA-223201110172C and grant 1U01FD004493-01 from the Center for Devices and Radiological Health, US Food and Drug Administration. DBK is supported by a Paul B. Beeson Career Development Award (NIA K23AG045963).

Keywords

Bayesian methods; cardiac resynchronization therapy; implantable cardioverter-defibrillator; meta-analysis; treatment effect heterogeneity

[1]Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA 02115
[2]Beth Israel Deaconess Medical Center, Boston, MA
[3]Hebrew SeniorLife Institute for Aging Research, Boston, MA
[4]Harvard Medical School, Boston, MA
[5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

1 Background

1.1 Treatment effect heterogeneity and meta-analysis

Patients may respond differently to treatment due to differences in biology, psychology, study conduct, treatment setting, operating physician, and other factors. If this variability is consistently associated with observable features, i.e., as a systematic treatment-covariate interaction, we say that the treatment effects are heterogeneous. For example, older patients may have a diminished response to treatment compared to younger patients. Motivations for investigating heterogeneous treatment effects (HTEs) include understanding the biological mechanisms of the intervention and targeting treatment to the patients who will benefit most. This is particularly true in the setting we consider, which entails risks of surgically implanting a complex cardiac electric device.

Exploratory analyses to *discover* subgroups across which treatment effects vary differs from confirmatory analyses to estimate effects in *pre-defined* subgroups. The literature contains many examples of exploratory analysis, especially in studies of genetic markers as potential subgroup identifiers – see Yang et al. (2009) for example. In this paper, we focus on confirmatory analysis, namely differences in responses to implantable cardiac device therapy in men and women.

Randomized controlled trials are an ideal setting for studying HTE, thanks to the well-known benefits of randomization, chiefly freedom from bias due to confounding. If trial populations comprise different mixtures of subgroup populations (in the extreme, separate trials for men and women), we must separate between-trial variability due to subgroup differences from variability due to other trial-level factors. Thus we estimate and combine subgroup-specific treatment effects from each trial. To preserve the benefits of randomization, assignment should be stratified by subgroup within each trial. Regression strategies may be used to correct for residual confounding due to imbalance across subgroup-treatment arm combinations. Sample sizes should be large to obtain precise estimates within study subgroups, but this is rarely the case, especially in device trials. Combining across trials increases sample size and allows estimation of an interaction parameter between the treatment effect and covariates of interest.

When only aggregate, study-level data are available, meta-regression is the usual technique for computing treatment-covariate interactions. For treatment effects that vary with a continuous modifier such as age, we can regress study-level treatment effects on the study-level mean age. However, when effects vary among discrete subgroups, such as men and women, regressing study-level treatment effects on the trial-level proportion of each sex is not powerful; differences *within* studies usually dwarf differences *between* studies. Worse, meta-regression is vulnerable to ecological bias (Berlin et al., 2002).

When individual patient data (IPD) are available, more rigorous analyses are possible (Lambert et al., 2002, Schmid et al., 2004). Methods to combine individual patient data in meta-analysis exist for numerous outcome types (Higgins et al., 2001, Goldstein et al., 2000, Turner et al., 2000, Smith et al., 2005, Whitehead et al., 2001). A nascent literature on methods for combining across data sources that have some combination of study- and individual-level data (Riley et al., 2010, Debray et al., 2012, Sutton et al., 2008) exists, with the applied literature lagging behind the statistical methods in terms of flexibility and sophis-

tication (Riley et al., 2007). Other work has established the circumstances under which IPD analysis should be considered (Simmonds and Higgins, 2007), between the extremes of zero within-study variation, when IPD offers no advantage, and zero between-study variation, when meta-regression fails.

Men and women have different distributions of important clinical variables that predict outcomes. Sex obviously predates the development of heart disease, so the direction of a causal arrow between biological sex and heart failure indicators and outcomes is not controversial. However, we do not wish to "adjust away" the differences between men and women, rather to obtain treatment effect estimates that are precise and unconfounded within subgroups. There are, however, different approaches for estimating causal effects in the presence of effect modification.

Assuming no threats to internal validity, subgroup-specific treatment effect estimates derived from randomized controlled trials are free of confounding. However, the combination of treatment arm and subgroup membership may not be randomized to patients, thus testing for differences among subgroup-specific estimates does not benefit from randomization in the same way. VanderWeele and Robins (2007) distinguish among four types of effect modification (Figure 1). In our analyses of treatment effect sex heterogeneity in implantable cardiac devices, we consider treatment T, outcome Y, effect modifier (i.e., sex subgroup) $\tilde{X}$, and observed prognostic factors U_1 and U_2. The simplest scenarios are direct effect modification by sex as a cause of the outcome (top left panel), or indirect modification by sex as a cause of observed prognostic factors U_1 (top center panel). These two scenarios represent a *causal* effect of subgroup on the outcome.

Using these causal directed acyclic graphics (DAGs) and the rules from Pearl (1995), we can establish the conditions for causal identification of subgroup-specific treatment effects. That paper presents three rules for manipulating a causal diagram to establish identification criteria. We do not review these here, only note that the probability distribution of interest is $Pr(Y|T,\tilde{X})$, where trial protocols assign the value of the treatment variable T but not subgroup $\tilde{X}$ (sex). We seek conditions that allow the subgroup variable $\tilde{X}$ value to provide the same causal identification as if it had been assigned.

Rule 2 of Theorem 3 in Pearl (1995) allows the exchange of observation for assignment if we can show that after removing the arrows emerging from $\tilde{X}$ from the applicable DAG, we can obtain conditional independence between the outcome Y and subgroup $\tilde{X}$ given the observable quantities. This is obviously true in the direct and indirect cases of Fig. 1, because removing the arrows out of $\tilde{X}$ isolates this node.

The bottom two panels of Figure 1 illustrate more complicated scenarios of confounding by a second observed variable U_2. Distinguishing between variables that lie on the causal pathway between the subgroup and outcome from those that are "merely" confounders is a key analysis issue to which we return below. We frame our methods in the context of these effect modification scenarios.

1.2 Heart failure

'Heart failure' means cardiac function that cannot meet the metabolic demands of the body at normal pressures, leading to shortness of breath, chest pain, fatigue, edema, or heart rhythm disorders (Libby and Braunwald, 2008). As the disease progresses, these symptoms

expand to include exercise intolerance, poor quality of life, and risk of hospitalization and death. The New York Heart Association (NYHA) classifies function ranging from Class I (asymptomatic) to Class IV, in which patients experience symptoms at rest, worsening with activity (The Criteria Committee for the New York Heart Association, 1994). In the United States, HF affects 5.7 million people, resulting in a million hospitalizations each year, 10% annual mortality among those with advanced disease, and lifetime health care costs of more than $100,000 per patient (Dunlay et al., 2010, National Heart Lung and Blood Institute, 2012).

Heart failure therapy depends in large part on whether the pumping function of the left ventricle is diminished, also referred to as "systolic heart failure." Therapy for systolic heart failure includes medications including ACE-inhibitors and beta-blockers. Such patients are also at heightened risk for ventricular arrhythmias, and may be eligible for an implantable cardioverter-defibrillator (ICD) (Epstein et al., 2008). This device alone does not improve the symptoms of HF or alter its physiology, but will reduce the likelihood of a sudden arrhythmic death. Impaired electrical signal transmission in the heart can lead to dyssynchrony of ventricular contractions and further functional impairment. When cardiac output is poor and patients are symptomatic, biventricular pacing may indicated. This is called cardiac resynchronization (CRT) and is abbreviated CRT-D when combined with an ICD (Jarcho, 2006). Figure 2 shows the chest x-ray of a patient following implantation of a CRT-D.

Recent evidence indicates that men and women with heart failure respond differently to biventricular pacing (Mooyaart et al., 2011, Arshad et al., 2011, Cheng et al., 2012, Kirubakaran et al., 2011). Three contemporary trials comparing CRT-D to ICD had planned male and female subgroup analyses. The earliest, RAFT, found a suggestive but non-significant treatment effect difference in the primary endpoint (hazard ratio for death or HF hospitalization) (Tang et al., 2010). Women appeared to benefit more, but sex-specific hazard ratios were not reported. A second study, REVERSE, found no treatment effect difference between men and women in the primary endpoint (odds ratio for "worsening HF") and the secondary endpoint (change in left ventricular end-systolic volume) (Linde et al., 2008). The most compelling evidence of sex heterogeneity in CRT-D effectiveness arises from the MADIT-CRT study, which estimated the hazard ratio for death or nonfatal heart failure for CRT-D versus ICD as 0.37 (95% CI: $0.22-0.61$) in women and 0.75 (95% CI: $0.59-0.97$) in men (Moss et al., 2009).

More recent work (Loring et al., 2012, 2013, Cheng et al., 2012) demonstrated differences in the etiology of heart failure that may explain these sex differences. Left bundle branch block (LBBB) and non-ischemic disease are both more common in women, and CRT-D benefits patients with LBBB more than those without (Perrin et al., 2012, Bilchick et al., 2010). Moreover, non-ischemic disease is strongly correlated with LBBB because this form of heart failure is rarely caused by a prior myocardial infarction (Strauss et al., 2013).

Guidelines to target CRT therapy require clarification of the following issue: do women appear to benefit more because they are more likely to have non-ischemic etiologies and LBBB or does sex heterogeneity persist after controlling for those factors? Refining patient selection is a critical unmet need for the field because despite the physiologic appeal of CRT, many recipients do not respond and may even be harmed by it. In all of the landmark trials, the number of women was small compared to men, resulting in much wider confidence intervals for the treatment effect estimates in women. Thus we consider pooling across

studies for more precise estimates of the treatment effects in men and women.

2 Methods

We develop a framework for estimating heterogeneous treatment effects using data from multiple studies, first in the absence of covariates, and then in the presence of potential confounders and mediators.

2.1 Subgroup-specific treatment effects with no covariates

We begin by estimating treatment effect within subgroups defined by $\tilde{X}$ alone, consistent with direct effect modification without confounding (recall Figure 1). With two treatments and two subgroups, $\tilde{X} \times T$ defines 4 groups, enabling direct specification of the outcome probabilities in each treatment-subgroup combination.

Indexing the trials by $j = 1, \ldots, J$, assume that $n_{t,\tilde{x},j}$ individuals are assigned to treatment t with subgroup identifier $\tilde{x} \in \tilde{X}$. Of these, $Y_{t,\tilde{x},j}$ experience the binary outcome, and we write a simple Binomial model

$$Y_{t,\tilde{x},j} \mid T, \tilde{X} \sim \text{Binomial}\left(n_{t,\tilde{x},j}, p_{t,\tilde{x},j}\right) \ . \tag{1}$$

We express the treatment effect in each subgroup $\tilde{x}$ and trial j as the probability difference

$$\delta^{unadj}_{\tilde{x},j} = p_{1,\tilde{x},j} - p_{0,\tilde{x},j} \ . \tag{2}$$

In this paper, we will use the probability difference scale for the treatment effect, but this is not required; we could also use relative risks, odds ratios, etc.

If either the treatment or effect modifier is continuous (instead of categorical), we require a link function to put a linear combination of coefficients and covariates on the $[0, 1]$ support of a probability parameter. The generalized linear model (GLM) extension of Eq. (1) defines the outcome probability as

$$g(p_{t,\tilde{x},j}) = \beta_{0j} + \beta_{Tj} t + \beta_{Xj} \tilde{x} + \beta_{TXj} t\tilde{x} \ , \tag{3}$$

where g is a link function such as logit, probit, or complementary log-log. We invert the link to get probabilities and again take differences to form treatment effects as in Eq. (2). The link functions mean that these probability differences are non-linear functions of the β_j coefficients,

$$p_{t,\tilde{x},j} = g^{-1}\left(\beta_{0j} + \beta_{Tj} t + \beta_{Xj} \tilde{x} + \beta_{TXj} t\tilde{x}\right) \tag{4}$$

$$\delta^{unadj}_{\tilde{x},j} = p_{1,\tilde{x},j} - p_{0,\tilde{x},j} \ . \tag{5}$$

2.1.1 Pooling unadjusted treatment effects

Combining information across trial-level treatment effects to obtain global effects requires exchangeability, which may require conditioning on individual- or trial-level covariates. First, we describe approaches without conditioning, then address handling covariates in Section 2.2.

A conventional estimate of the global treatment effect estimate uses a weighted liner combination of trial-level treatment effects,

$$\delta_{\tilde{x}}^{pool} = \frac{\sum_{j=1}^{J} w_j^* \delta_{\tilde{x},j}^{unadj}}{\sum_{j=1}^{J} w_j^*} \ . \tag{6}$$

The weights, w_j^*, are based on inverse variances of the estimates and include between-trial heterogeneity for the fairest comparison to hierarchical models (DerSimonian and Laird, 1986). Omitting the $\tilde{x}$ subscript for the moment, let σ_j^2 be the variance of the j^{th} study's treatment effect estimate δ_j. A model that ignores heterogeneity uses weights $w_j = 1/\sigma_j^2$. We construct an estimate of the between-study variance

$$\tau^2 = \frac{\sum w_j \delta_j^2 - \frac{(\sum \delta_j w_j)^2}{\sum w_j} - 4}{\sum w_j - \frac{\sum w_j^2}{\sum w_j}} \ . \tag{7}$$

When this is positive, it indicates "significant" heterogeneity, and thus we use the modified weights $w_j^* = (1/w_j + \tau^2)^{-1}$. The variance of this pooled estimate is $(\sum w_j^*)^{-1}$.

Hierarchical models borrow information across parameters, and produce posterior distributions for both trial- and global-level treatment effects, facilitating inference at both levels. The generalized linear *mixed* model (GLMM) extension of Eq. (3) adds shrinkage distributions on the study-level parameters, $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{Tj}, \beta_{Xj}, \beta_{TXj})'$,

$$\boldsymbol{\beta}_j \overset{iid}{\sim} \text{Normal}\left(\boldsymbol{\beta}^{Bayes}, \Sigma\right) \ , \tag{8}$$

where the global parameter is $\boldsymbol{\beta}^{Bayes} = \left(\beta_0^{Bayes}, \beta_T^{Bayes}, \beta_X^{Bayes}, \beta_{TX}^{Bayes}\right)'$. To form global treatment effects on the probability difference scale, we take linear combinations of these parameters and invert the link to obtain global outcome probabilities and thence treatment effects,

$$p_{t,\tilde{x}}^{Bayes} = g^{-1}\left(\beta_0^{Bayes} + \beta_T^{Bayes} t + \beta_X^{Bayes} \tilde{x} + \beta_{TX}^{Bayes} t\tilde{x}\right) \tag{9}$$

$$\delta_{\tilde{x}}^{Bayes} = p_{1,\tilde{x}}^{Bayes} - p_{0,\tilde{x}}^{Bayes} \ . \tag{10}$$

2.2 Subgroup-specific treatment effects with covariate adjustment

When randomization is not stratified by subgroup, imbalanced variables can bias subgroup-specific treatment effect estimates. Standard methods for assessing confounding bias exist for continuous (Senn, 1989), binary (Robinson and Jewell, 1991), and survival outcomes (Ford et al., 1995). The strength of the relationship between confounder and outcome dominates across-arm imbalance in determining whether adjustment is needed. As Pocock et al. (2002) note, "...if the correlation is weak... even a statistically significant covariate imbalance is unimportant. ... On the other hand, if a covariate is strongly related to outcome... then it is important to adjust for it regardless of the extent (or lack) of covariate imbalance." The outcome variable measured at baseline is often the most strongly prognostic (Frison and Pocock, 1992).

In the simplest case, a single, fully observed binary variable U further divides the population into $T \times \tilde{X} \times U$, or 8 groups when the treatment and subgroup are also binary. We can extend model (1) by adding another subscript for the covariate,

$$Y_{t,\tilde{x},j,u} \mid T, \tilde{X}, U \sim \text{Binomial}\left(n_{t,\tilde{x},j,u}, p_{t,\tilde{x},j,u}\right) \ . \tag{11}$$

Similarly, we can extend the GLM model (3) with another regression coefficient,

$$g(p_{t,\tilde{x},j,u}) = \beta_{0j} + \beta_{Tj}t + \beta_{Xj}\tilde{x} + \beta_{Uj}u + \beta_{TXj}t\tilde{x} \ . \tag{12}$$

From either of these, we form a *conditional* global treatment effect estimate by taking a linear combination of probability differences as before,

$$\delta^{cond}_{\tilde{x},j,u} = p_{1,\tilde{x},j,u} - p_{0,\tilde{x},j,u} \tag{13}$$

$$\delta^{cpool}_{\tilde{x},u} = \frac{\sum_{j=1}^{J} w^*_{j,u}\delta^{cond}_{\tilde{x},j,u}}{\sum_{j=1}^{J} w^*_{j,u}} \ . \tag{14}$$

The weights have the same form as above, using the conditional treatment differences $\delta^{cond}_{\tilde{x},j,u}$ and their variances in the weights and between-study variance estimate of Eq. (7).

Conditioning on the covariate u is a nuisance necessary to correct for confounding, but our goal is unconfounded *marginal* treatment effects. Thus we integrate the outcome probabilities over the distribution of the confounder, take differences to form trial-level treatment effects, and pool these effects using a linear combination,

$$p^{marg}_{t,\tilde{x},j} = \int p_{t,\tilde{x},j,u}p(u|t,\tilde{x},j)du \tag{15}$$

$$\delta^{marg}_{\tilde{x},j} = p^{marg}_{1,\tilde{x},j} - p^{marg}_{0,\tilde{x},j} \tag{16}$$

$$\delta^{mpool}_{\tilde{x}} = \frac{\sum_{j=1}^{J} w^*_{j}\delta^{marg}_{\tilde{x},j}}{\sum_{j=1}^{J} w^*_{j}} \ . \tag{17}$$

This approach does not borrow strength across covariate effects, but does allow trial-level effects to be adjusted for a *different* sets of confounders. Here, we use the marginalized treatment differences $\delta^{marg}_{\tilde{x},j}$ and their inverse variances in the weights and across-study variance estimate of Eq. (7).

The extension of the hierarchical model in Eq. (8) to include a confounder is straightforward, we just use the trial-level parameters from the expanded GLM (12) and analogously expand the global parameter, $\boldsymbol{\beta}^{Bayes} = (\beta_0, \beta_T, \beta_X, \beta_U, \beta_{TX})'$. Then we construct global adjusted probabilities and use these to form conditional treatment differences,

$$p^{cBayes}_{t,\tilde{x},u} = g^{-1}(\beta_0 + \beta_T t + \beta_X\tilde{x} + \beta_U u + \beta_{TX}t\tilde{x}) \tag{18}$$

$$\delta^{cBayes}_{\tilde{x},u} = p^{cBayes}_{1,\tilde{x},u} - p^{cBayes}_{0,\tilde{x},u} \ . \tag{19}$$

As before, we can integrate the probabilities over the distribution of u and convert to the probability difference scale to obtain marginal global treatment differences,

$$p^{mBayes}_{t,\tilde{x}} = \int p^{cBayes}_{t,\tilde{x},u}p(u|t,\tilde{x})du \tag{20}$$

$$\delta^{mBayes}_{\tilde{x}} = p^{mBayes}_{1,\tilde{x}} - p^{mBayes}_{0,\tilde{x}} \ . \tag{21}$$

2.2.1 Meta-regression on trial-level variables

We may require conditioning on trial-level variables to justify the exchangeability of trial-level parameters in the above hierarchical models. The extension to meta-regression is straightforward; we simply incorporate trial-level predictors z_j into the shrinkage distribution's mean. Without conditioning on any covariates, we have

$$\boldsymbol{\beta}_j \sim \text{Normal}\left(\boldsymbol{\beta}^{meta} + \boldsymbol{\alpha}\mathbf{z}_j, \Sigma_{meta}\right) \tag{22}$$

$$p^{meta}_{t,\tilde{x}} = g^{-1}\left(\beta^{meta}_0 + \beta^{meta}_T t + \beta^{meta}_X \tilde{x} + \beta^{meta}_{TX} t\tilde{x}\right) \tag{23}$$

$$\delta^{meta}_{\tilde{x}} = p^{meta}_{1,\tilde{x}} - p^{meta}_{0,\tilde{x}} \tag{24}$$

and similarly with adjustment for a covariate,

$$\boldsymbol{\beta}_j \sim \text{Normal}\left(\boldsymbol{\beta}^{cmeta} + \boldsymbol{\alpha}\mathbf{z}_j, \Sigma_{meta}\right) \tag{25}$$

$$p^{meta}_{t,\tilde{x}} = g^{-1}\left(\beta^{cmeta}_0 + \beta^{cmeta}_T t + \beta^{cmeta}_X \tilde{x} + \beta^{cmeta}_U u + \beta^{cmeta}_{TX} t\tilde{x}\right) \tag{26}$$

$$\delta^{cmeta}_{\tilde{x}} = p^{cmeta}_{1,\tilde{x},u} - p^{cmeta}_{0,\tilde{x},u} \tag{27}$$

and in the marginal setting,

$$p^{mmeta}_{t,\tilde{x}} = \int p^{cmeta}_{t,\tilde{x},u} p(u|t,\tilde{x}) du \tag{28}$$

$$\delta^{mmeta}_{\tilde{x}} = p^{mmeta}_{1,\tilde{x}} - p^{mmeta}_{0,\tilde{x}} \tag{29}$$

We summarize all estimators in Table 1.

2.3 Mediated Effect Modification

In addition to covariates' role in confounding, we also consider these variables as potential mediators. As shown in Figure 1, the observed variable U_1 mediates the effect of the subgroup $\tilde{X}$ on the outcome, with or without confounding by U_2. Treating a variable on the causal path between $\tilde{X}$ and Y as a confounder corrupts the interpretation of other parameters. To see this, suppose the observed variable U completely mediates the treatment modification of subgroup $\tilde{X}$ and neither variable affects treatment (i.e., no confounding), then the true linear model is

$$E(Y|T,U,\tilde{X}) = \beta_0 + \beta_T t + \beta_U u + \beta_{TU} tu \ .$$

If we incorrectly assume U is a confounder and $\tilde{X}$ is a complete, direct effect modifier, we would fit the linear model

$$E(Y|T,U,\tilde{X}) = \beta^*_0 + \beta^*_T t + \beta^*_X \tilde{x} + \beta^*_U u + \beta^*_{TX} t\tilde{x} \ .$$

Assuming the mediator and moderator are both binary, a comparison of the true and modeled (starred) treatment differences for each value of U and $\tilde{X}$ yields a 2 by 2 table with entries:

			Subgroup	
			$\tilde{x}=0$	$\tilde{x}=1$
Covariate	$u=0$	True	β_T	β_T
		Modeled	β^*_T	$\beta^*_T + \beta^*_{TX}$
	$u=1$	True	$\beta_T + \beta_{TU}$	$\beta_T + \beta_{TU}$
		Modeled	β^*_T	$\beta^*_T + \beta^*_{TX}$

Both the modeled treatment and effect modification parameters β_T^* and β_{TX}^* are biased relative to the truth. One possible approach is to fit a saturated model with all interaction terms and rely on the data to determine which parameters should be zero (those with daggers),

$$E(Y|T,U,\tilde{X}) = \beta_0 + \beta_T t + \beta_U u + \beta_{TU} tu + \beta_X^\dagger \tilde{x} + \beta_{TX}^\dagger t\tilde{x} + \beta_{UX}^\dagger u\tilde{x} + \beta_{TXU}^\dagger t\tilde{x}u \; . \qquad (30)$$

3 Applying these methods in CRT-D trial data

To examine sex differences in the effectiveness of CRT-D versus ICD alone, we implement the approaches discussed using data from five randomized controlled trials, summarized briefly in Table 2. Other publications detail the trial protocols (Saxon et al., 1999, Young et al., 2003, Tang et al., 2009, Linde et al., 2006, Moss et al., 2006) and main results (Higgins et al., 2003, Abraham et al., 2004, Tang et al., 2010, Linde et al., 2008, Moss et al., 2009). We exclude two trials (COMPANION and CARE-HF) that do not include both a CRT-D and ICD arm. Together, these trials enrolled 1051 women, more than twice the number in the largest trial, enhancing our ability to estimate treatment effects in women and men.

3.1 Variable definitions

The functional outcome measures and follow-up periods varied widely across trials. Only two functional outcomes were measured consistently, and no time point was in common across more than 3 trials. The time-to-death is recorded across trials, but the length of follow-up (and thus right-censoring proportions) varied from more than 7 years in RAFT (72% censored) to only 6 months in CONTAK (88% censored). Most of the trials are powered to a composite primary outcome measuring "progression" e.g., mortality or hospitalization for heart failure. One study recorded only death, not also heart failure events.

We define a binary outcome equal to 1 if an individual's NYHA class improves (i.e., decreases) from baseline to follow-up and 0 otherwise. We also consider a continuous outcome: distance walked in 6 minutes, a functional outcome measured across all studies. For CONTAK and MIRACLE-ICD, we use the 3- or 6-month follow-up time[2] and for the remaining trials, we use 12 months.

We consider two baseline clinical characteristics of individuals. First, ischemic disease etiology is a binary indicator of whether the underlying disease is caused by coronary artery disease and previous myocardial infarction. Second, we use baseline 6-minute walk distance which is strongly correlated with the follow-up 6-minute walk distance outcome. Both showed some evidence of imbalance across arms (see below) and are clinically important as predictors of outcomes (Barsheshet et al., 2012), thus potential confounders. We also consider a trial-level variable, the proportion of patients with intraventricular (IV) condition delay. This is measured across all studies and showed some evidence of a relationship with the strength of the treatment effect across trials, so we use it in the meta-regression approaches.

[2]CONTAK had two phases, one with 3-month follow-up and the other with 6-month; we pool both phases in our analyses.

3.2 Assessing modeling assumptions

We begin by studying heterogeneity across trials, arms, and subgroups, which informs the choice of modeling strategy. Differences in baseline medical history variables across trials largely reflect eligibility criteria. Medications also vary, perhaps reflecting secular trends over the years trials enrolled patients. Across all trials, women are more likely to have LBBB, while men are more likely to have ischemic etiology and previous bypass surgery. Functional outcome differences at baseline are smaller. Women have slightly shorter 6-min walk, VO_2, and LVEDD measurements and slightly higher quality of life scores.

The models in Section 2.1 assume that treatment-by-sex groups require no covariate adjustment. This is important because most of these trials did not randomize by gender, so imbalances in important predictive factors may bias results. The left panel of Figure 3 displays absolute standardized mean differences (ASMDs) between arms, sorted in descending order of across-trial averages (plotted with stars). Differences for women (in black) are generally larger than for men (in grey) and vary more across trials, likely due to the smaller number of women in all the trials. The most important factor in confounding adjustment is the prognostic strength of the covariate. The right panel of Figure 3 displays ASMDs of covariates across outcome groups (i.e., NYHA class improved vs same/worse) by sex subgroup and trial.

3.3 Individual trial sex-specific estimates

Table 3 summarizes how we implement the methods described in Section 2 for the CRT-D scenario. There are four coefficients in the unadjusted models (intercept, treatment, subgroup, and treatment×subgroup) and five in the adjusted models (adding the covariate effect). For the binary outcome, we have a generalized linear regression model

$$Y_{t,\tilde{x},j} \sim \text{Binomial}(n_{t,\tilde{x},j}, p_{t,\tilde{x},j}) \tag{31}$$

$$\text{probit}(p_{t,\tilde{x},j}) = \beta_{0j} + \beta_{Xj}\tilde{x} + \beta_{Tj}t + \beta_{TXj}t\tilde{x} \ , \tag{32}$$

where $j = 1, \ldots, 5$ indexes trial. For the continuous outcome, we first divide the outcomes by the trial-level standard deviations so that we can fix the error variance at 1. Then we have a linear regression model,

$$Y_{t,\tilde{x},j} \sim \text{Normal}(\mu_{t,\tilde{x},j}, 1) \tag{33}$$

$$\mu_{t,\tilde{x},j} = \beta_{0j} + \beta_{Xj}\tilde{x} + \beta_{Tj}t + \beta_{TXj}t\tilde{x} \ . \tag{34}$$

In the hierarchical models, we specify a multivariate normal shrinkage distribution on the trial-level effects and constrain the correlations to zero for identifiability. Thus we have only variance parameters in the second level of the basic Bayesian hierarchical model,

$$\boldsymbol{\beta}_j \sim N(\boldsymbol{\beta}^{Bayes}, \Sigma) \ \text{ where} \tag{35}$$

$$\Sigma = \begin{pmatrix} \sigma_0^2 & 0 & 0 & 0 \\ & \sigma_X^2 & 0 & 0 \\ & & \sigma_T^2 & 0 \\ & & & \sigma_{TX}^2 \end{pmatrix} \ . \tag{36}$$

The meta-regression models add a second-level parameter vector $\boldsymbol{\alpha}$ containing the effect of the study-level covariate on each random effect mean,

$$\boldsymbol{\beta}_j \sim \text{Normal}\left(\boldsymbol{\beta}^{Bayes} + \mathbf{Z}_j\boldsymbol{\alpha}, \Sigma\right) \tag{37}$$

where $\mathbf{Z}_j = \text{diag}\left(v_j, v_j, v_j, v_j\right)'$ and v_j is the proportion of IVCD patients in each trial. For simplicity, we assume that the same covariate affects each coefficient, but allow the coefficients to differ.

Figure 4 plots the effects from all the strategies in Table 1 applied to the binary NYHA improvement outcome. Crossed, solid, and open symbols indicate unadjusted, conditional, and marginal effects. Despite the discrepancies across arms, ischemic etiology is not confounding these treatment effects; there is little to no difference among the three estimate types. Square, circle, and triangle plotting symbols indicate independent, hierarchical Bayesian ("shrunk"), and Bayesian meta-regression pooling strategies, respectively. In the trial-specific estimates, there are few differences between the independent and hierarchical estimates, except that shrinkage produces narrower trial-level credible intervals for small trials. However, in the global parameters, we clearly see the price we pay for increasing the model complexity: increasing uncertainty about global treatment effects. The conventional pooling approach treats the between-study heterogeneity and trial-level effect estimator variances as known, while the Bayesian hierarchical model treats these as unknown parameters, and meta-regression model adds still more second-level parameters. This is asking a lot of the data, which comprise only five observations per study (in a binary model with a binary confounder) in only five studies. Compare this to the nominal parameter count in the hierarchical models: five regression coefficients per study and five variances.

Models that substitute baseline 6-minute walk distance as a potential confounder produce similar results, see Figure 5. This continuous baseline covariate requires that we switch from trial-level specification to individual-level. Because confounder adjustment has no essentially impact, we do not pursue marginalization here.

We can quantify the extent of shrinkage by comparing the variance of each study-level estimate to the estimated across-study variance, $1 - [SD(\beta_{jk})/\hat{\sigma_k}]$ for $k \in \{0, T, X, TX\}$. Small values indicate that the uncertainty of the trial-level estimate is large compared to the across-trial variation. Figure 6 displays the shrinkage factors for parameters of the models that produce the estimates in Figure 4. Most of the effects are moderately shrunk in the hierarchical models, with the notable exception of the interaction effects. These have negative shrinkage factors, indicating trial-level parameters are estimated with *more* error than the variability across trials.

Figure 7 displays the effect estimates for linear models of the 6-minute walk test outcome. These models use individual-level data and we have added a saturated model as in Eq. (30) to study the potential for mediation or effect modification by ischemic etiology. Now we see substantial impacts on the effect estimates for women with ischemic disease. This is due to a large and significantly non-zero coefficients for three interaction effects in the MADIT-CRT trial: treatment in women, ischemic disease in women, and the interaction of all three (treatment, sex, and etiology). This figure shows substantially smaller (negative) treatment differences for women with ischemic disease. The estimates for women with non-ischemic disease are large and significantly positive (not shown).

We repeated this exercise with left bundle branch block as a potential effect modifier, either alone or in combination with sex. Linear models for the 6-minute walk test outcome (not shown) again produce significantly heterogeneous treatment effects only in the MADIT-CRT trial. There, patients with LBBB benefit significantly more from CRT-D vs ICD than those without, but the global effects are not significantly different from zero in either group, whether using hierarchical modeling or conventional pooling. Similarly, combining both sex and LBBB in a saturated model, only MADIT-CRT shows any differences. Men and women without LBBB have *negative* estimates for the treatment effect, while women with LBBB appear to benefit the most from treatment (the effect for men with LBBB was null).

We should exercise caution in over-interpreting these results. Dividing the data by numerous covariates into small subgroups may lead to unstable estimates. For example, even in MADIT-CRT, the largest trial, there are only 33 women with ischemic disease in the ICD arm and 61 in the CRT-D arm. The counts of women without LBBB are similarly small: 19 in the ICD arm and 27 in the CRT-D arm of that trial. We would expect extensive shrinkage of factors that are poorly estimated in each trial, but only if there are sufficient numbers of studies to estimate the shrinkage variance well. Figure 8 shows the extent of shrinkage for coefficients in a linear model for 6-minute walk distances. Here, we see a pronounced difference between the intercepts, which are highly shrunk, and the rest, which are relatively un-shrunk. The black lines for treatment effects in the MADIT-CRT trial in Figure 7 indicate a significant interaction between treatment and sex subgroup in the independent models. In the hierarchical models, this treatment heterogeneity is attenuated as the estimates are shrunk toward the rest of the trials, which do not strongly support sex heterogeneity or moderation by LBBB.

4 Discussion

We combine information across randomized controlled trial data using models that allow treatment effect heterogeneity in subgroups, confounding and mediation by clinical covariates, and trial-level effect modification. With only a handful of trials, we found little benefit to hierarchical models compared to conventional pooling mechanisms. Although we observed relatively little shrinkage of the trial-level estimates, the choice of pooling mechanism impacted the global effect estimates, particularly their credible intervals. The global point estimates from conventional pooling are similar to the hierarchical model point estimates, but with much narrower confidence intervals.

The difficulty of estimating variance parameters in hierarchical models is well known, particularly when the number of groups is small. However, we estimate the variances with reasonable precision. For example, in linear models for the 6-minute walk distance, the upper bounds of the posterior 95% credible intervals for the random effect standard deviations are all < 1.5. The greatest posterior uncertainty about global effect estimates is in our meta-regression models. This is mostly due to the lack of information about the second-level α parameters.

Our Bayesian hierarchical models do allow direct inference on the global estimates of treatment effects in men and women separately, even with transformation of the model parameters onto a more interpretable scale. For example, in our probit models, we transform

from regression coefficients to probability differences. These parametric approaches also allow us to incorporate variables at either the trial or individual level, relatively simple implementation, and efficient use of information.

Because a characteristic such as sex cannot be randomly assigned, post-hoc comparisons between subgroups may be confounded. In this paper, the treatment effects for men and women in each trial are unchanged by controlling for a few potential confounders: the unadjusted, conditional, and marginal estimates are essentially similar, both at the trial and global level. This indicates that the randomization in these trials successfully ameliorates confounding, at least on these measured covariates.

However, models that include interactions between sex and known clinical modifiers of treatment response (LBBB and non-ischemic disease) show some evidence of mediation beyond sex differences. Notably, this is strongest in the trial that identified the strongest signal for sex heterogeneity, MADIT-CRT. Shrinking the few trials with substantial male-female differences (especially MADIT-CRT) toward the generally smaller effects in the remaining trials mutes the differences. From a policy perspective, this may be preferable, as it can clarify the robustness of a subgroup effect encountered in a single trial. Treatment effect heterogeneity in only one trial among many is more likely due to features of the trial procedures or enrolled population than true biological differences in response to treatment.

The next step in clarifying these results for clinical practice should be further analyses of large data sets using well-measured clinical markers to distinguish between sex heterogeneity due to differences in etiology versus "true" heterogeneity by sex. Recall that interacting LBBB or ischemic etiology with sex and treatment arm resulted in very counts of patients in our data. Unfortunately large databases such as the National Cardiac Device Registry of ICD and CRT-D devices lack follow-up information for clinical outcomes. Administrative databases (claims) may include longer follow-up for outcomes such as death and hospitalization, but do not include functional measures and usually lack precise clinical markers. Thus it will likely be necessary to synthesize evidence across multiple data sources.

References

Abraham, W., Young, J., Leon, A., Adler, S., Bank, A., Hall, S., Lieberman, R., Liem, L., O'Connell, J., Schroeder, J., and Wheelan, K. (2004). Effects of cardiac resynchronization on disease progression in patients with left ventricular systolic dysfunction, an indication for an implantable cardioverter-defibrillator, and mildly symptomatic chronic heart failure. *Circulation*, 110(18):2864–8.

Arshad, A., Moss, A., Foster, E., Padeletti, L., Barsheshet, A., Goldenberg, I., Greenberg, H., Hall, W., McNitt, S., Zareba, W., Solomon, S., and Steinberg, J. (2011). Cardiac resynchronization therapy is more effective in women than in men: The MADIT-CRT (Multicenter Automatic Defibrillator Implantation Trial with Cardiac Resynchronization Therapy) trial. *Journal of the American College of Cardiology*, 57(7):813–20.

Barsheshet, A., Brenyo, A., Goldenberg, I., and Moss, A. (2012). Sex-related differences in patients' responses to heart failure therapy. *Nature reviews. Cardiology*, 9(4):234–42.

Berlin, J., Santanna, J., Schmid, C., Szczech, L., and Feldman, H. (2002). Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Statistics in Medicine*, 21(3):371–387.

Bilchick, K., Kamath, S., DiMarco, J., and Stukenborg, G. (2010). Bundle-branch block morphology and other predictors of outcome after cardiac resynchronization therapy in Medicare patients. *Circulation*, 122(20):2022–30.

Cheng, A., Gold, M., Waggoner, A., Meyer, T., Seth, M., Rapkin, J., Stein, K., and Ellenbogen, K. (2012). Potential mechanisms underlying the effect of gender on response to cardiac resynchronization therapy: Insights from the SMART-AV multicenter trial. *Heart Rhythm*, 9(5):736–41.

Debray, T., Koffijberg, H., Vergouwe, Y., Moons, K., and Steyerberg, E. (2012). Aggregating published prediction models with individual participant data: a comparison of different approaches. *Statistics in Medicine*, 31:2697–2712.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7:177–188.

Dunlay, S., Shah, N., Shi, Q., Morlan, B., VanHouten, H., Long, K. H., and Roger, V. (2010). Lifetime costs of medical care after heart failure diagnosis. *Circulation: Cardiovascular Quality and Outcomes*, 4(1):68–75.

Epstein, A., DiMarco, J., Ellenbogen, K., N.A. Estes, r., Freedman, R., Gettes, L., Gillinov, A., Gregoratos, G., Hammill, S., Hayes, D., Hlatky, M., Newby, L., Page, R., Schoenfeld, M., Silka, M., Stevenson, L., and Sweeney, M. (2008). ACC/AHA/HRS 2008 guidelines for device-based therapy of cardiac rhythm abnormalities. *Heart Rhythm*, 5(6):e1–62.

Ford, I., Norrie, J., and Ahmadi, S. (1995). Model inconsistency, illustrated by the cox proportional hazards model. *Statistics in Medicine*, 14(8):735–46.

Frison, L. and Pocock, S. (1992). Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statistics in Medicine*, 11(13):1685–704.

Goldstein, H., Yang, M., Omar, R., Turner, R., and Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *Applied Statistics*, 49:399–412.

Higgins, J., Whitehead, A., Turner, R., Omar, R., and Thompson, S. (2001). Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine*, 20(15):2219–41.

Higgins, S., Hummel, J., Niazi, I., Giudici, M., Worley, S., Saxon, L., Boehmer, J., Higginbotham, M., Marco, T. D., Foster, E., and Yong, P. (2003). Cardiac resynchronization therapy for the treatment of heart failure in patients with intraventricular conduction delay and malignant ventricular tachyarrhythmias. *Journal of the American College of Cardiology*, 42(8):1454–1459.

Jarcho, J. (2006). Clinical therapeutics: biventricular pacing. *New England Journal of Medicine*, 20:288–94.

Kirubakaran, S., Ladwiniec, A., Arujuna, A., Ginks, M., McPhail, M., Bostock, J., Carr-White, G., and Rinaldi, C. (2011). Male gender and chronic obstructive pulmonary disease predict a poor clinical response in patients undergoing cardiac resynchronisation therapy. *International Journal of Clinical Practice*, 65(3):281–8.

Kramer, D., Reynolds, M., and Mitchell, S. (2013). Resynchronization: Considering device-based cardiac therapy in older adults. *Journal of the American Geriatrics Society*, 61(4):615–21.

Lambert, P., Sutton, A., Abrams, K., and Jones, D. (2002). A comparison of summary patient-level covariates in meta-regression with indivudal patient data meta-analysis. *Journal of Clinical Epidemiology*, 55:86–94.

Libby, P. and Braunwald, E. (2008). *Braunwald's heart disease: A textbook of cardiovascular medicine*. Saunders/Elsevier, Philadelphia.

Linde, C., Abraham, W., Gold, M., Sutton, M. S. J., Ghio, S., and Daubert, C. (2008). Randomized trial of cardiac resynchronization in mildly symptomatic heart failure patients and in asymptomatic patients with left ventricular dysfunction and previous heart failure symptoms. *Journal of the American College of Cardiology*, 52(23):1834–43.

Linde, C., Gold, M., Abraham, W., and Daubert, J. (2006). Rationale and design of a randomized controlled trial to assess the safety and efficacy of cardiac resynchronization therapy in patients with asymptomatic left ventricular dysfunction with previous symptoms or mild heart failure–the REsynchronization reVErses Remodeling in Systolic left vEntricular dysfunction (REVERSE) study. *American Heart Journal*, 151(2):288–94.

Loring, Z., Canos, D., Selzman, K., Herz, N., Silverman, H., MaCurdy, T., Worall, C., J.Kelman, Riches, M., Pena, I., and Strauss, D. (2012). Abstract 15602: Left bundle branch block predicts better survival in women than men receiving cardiac resynchronization therapy: Long term follow-up of 145,000 patients. *Circ*, 126:A15602.

Loring, Z., Strauss, D., Gerstenblith, G., Tomaselli, G., Weiss, R., and Wu, K. (2013). Cardiac MRI scar patterns differ by gender in an implantable cardioverter defibrillator and cardiac resynchronization cohort. *Heart Rhythm*.

Mooyaart, E., Marsan, N., van Bommel, R., Thijssen, J., Borleffs, C., Delgado, V., van der Wall, E., Schalij, M., and Bax, J. (2011). Comparison of long-term survival of men versus women with heart failure treated with cardiac resynchronization therapy. *American Journal of Cardiology*, 108(1):63–8.

Moss, A., Brown, M., Cannom, D., Daubert, J., Estes, M., Foster, E., Greenberg, H., Hall, W., Higgins, S., Klein, H., Pfeffer, M., Wilber, D., and Zareba, W. (2006). Multicenter Automatic Defibrillator Implantation Trial– Cardiac Resynchronization Therapy (MADIT-CRT): Design and clinical protocol. *Annals of Noninvasive Electrocardiology*, 10(4):s34–s43.

Moss, A., Hall, W., Cannom, D., Klein, H., Brown, M., Daubert, J., N.A. Estes, r., Foster, E., Greenberg, H., Higgins, S., Pfeffer, M., Solomon, S., Wilber, D., and Zareba, W. (2009). Cardiac-resynchronization therapy for the prevention of heart-failure events. *New England Journal of Medicine*, 361(14):1329–38.

National Heart Lung and Blood Institute (2012). Morbidity & mortality: 2012 chart book on cardiovascular, lung, and blood diseases. Technical report, National Institutes of Health.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–710.

Perrin, M., Green, M., Redpath, C., Nery, P., Keren, A., Beanlands, R., and Birnie, D. (2012). Greater response to cardiac resynchronization therapy in patients with true complete left bundle branch block: A predict substudy. *Europace*, 14(5):690–5.

Pocock, S., Assmann, S., Enos, L., and Kasten, L. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine*, 21(19):2917–30.

Riley, R., Lambert, P., and Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*, 340:c221.

Riley, R., Simmonds, M., and Look, M. (2007). Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology*, 60:431–9.

Robinson, L. and Jewell, N. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 59:227–240.

Saxon, L., Boehmer, J., Hummel, J., Kacet, S., Marco, T. D., Naccarelli, G., and Daoud, E. (1999). Biventricular pacing in patients with congestive heart failure: Two prospective randomized trials. *American Journal of Cardiology*, 83:120D–123D.

Schmid, C., Stark, P., Berlin, J., and Landais, P. (2004). Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level factors. *Journal of Clinical Epidemiology*, 57:683–697.

Senn, S. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine*, 8(4):467–75.

Simmonds, M. and Higgins, J. (2007). Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. *Statistics in Medicine*, 26(15):2982–99.

Smith, C. T., Williamson, P., and Marson, A. (2005). Investigating heterogeneity in an individual patient data met-analysis of time to event outcomes. *Statistics in Medicine*, 24:1307–1319.

Strauss, D., Loring, Z., Selvester, R., Gerstenblith, G., Tomaselli, G., Weiss, R., Wagner, G., and Wu, K. (2013). Right, but not left, bundle branch block is associated with large anteroseptal scar. *Journal of the American College of Cardiology*, 62(11):959–67.

Sutton, A., Kendrick, D., and Coupland, C. (2008). Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine*, 27(5):651–69.

Tang, A., Wells, G., Arnold, M., Connolly, S., Hohnloser, S., Nichol, G., Rouleau, J., Sheldon, R., and Talajic, M. (2009). Resynchronization/defibrillation for ambulatory heart failure trial: Rationale and trial design. *Current Opinion in Cardiology*, 24(1):1–8.

Tang, A., Wells, G., Talajic, M., Arnold, M., Sheldon, R., Connolly, S., Hohnloser, S., Nichol, G., Birnie, D., Sapp, J., Yee, R., Healey, J., and Rouleau, J. (2010). Cardiac-resynchronization therapy for mild-to-moderate heart failure. *New England Journal of Medicine*, 363(25):2385–95.

The Criteria Committee for the New York Heart Association (1994). *Nomenclature and Criteria for Diagnosis of Diseases of the HEart and Great Vessels, 9th ed.* Little Brown and Company.

Turner, R., Omar, R., Yang, M., Goldstein, H., and Thompson, S. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, 19(24):3417–32.

VanderWeele, T. and Robins, J. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*, 18(5):561–8.

Whitehead, A., Omar, R., Higgins, J., Savaluny, E., Turner, R., and Thompson, S. (2001). Meta-analysis of ordinal outcomes using individual patient data. *Statistics in Medicine*, 20(15):2243–60.

Yang, J., Cheng, C., Yang, W., Pei, D., Cao, X., Fan, Y., Pounds, S., Neale, G., Trevino, L., French, D., Campana, D., Downing, J., Evans, W., Pui, C., Devidas, M., Bowman, W., Camitta, B., Willman, C., Davies, S., Borowitz, M., Carroll, W., Hunger, S., and Relling, M. (2009). Genome-wide interrogation of germline genetic variation associated with treatment response in childhood acute lymphoblastic leukemia. *JAMA*, 301(4):393–403.

Young, J., Abraham, W., Smith, A., Leon, A., Lieberman, R., Wilkoff, B., Canby, R., Schroeder, J., Liem, L., Hall, S., and Wheelan, K. (2003). Combined cardiac resynchronization and implantable cardioversion defibrillation in advanced chronic heart failure: The MIRACLE ICD Trial. *JAMA*, 289(20):2685–94.

Tables and Figures

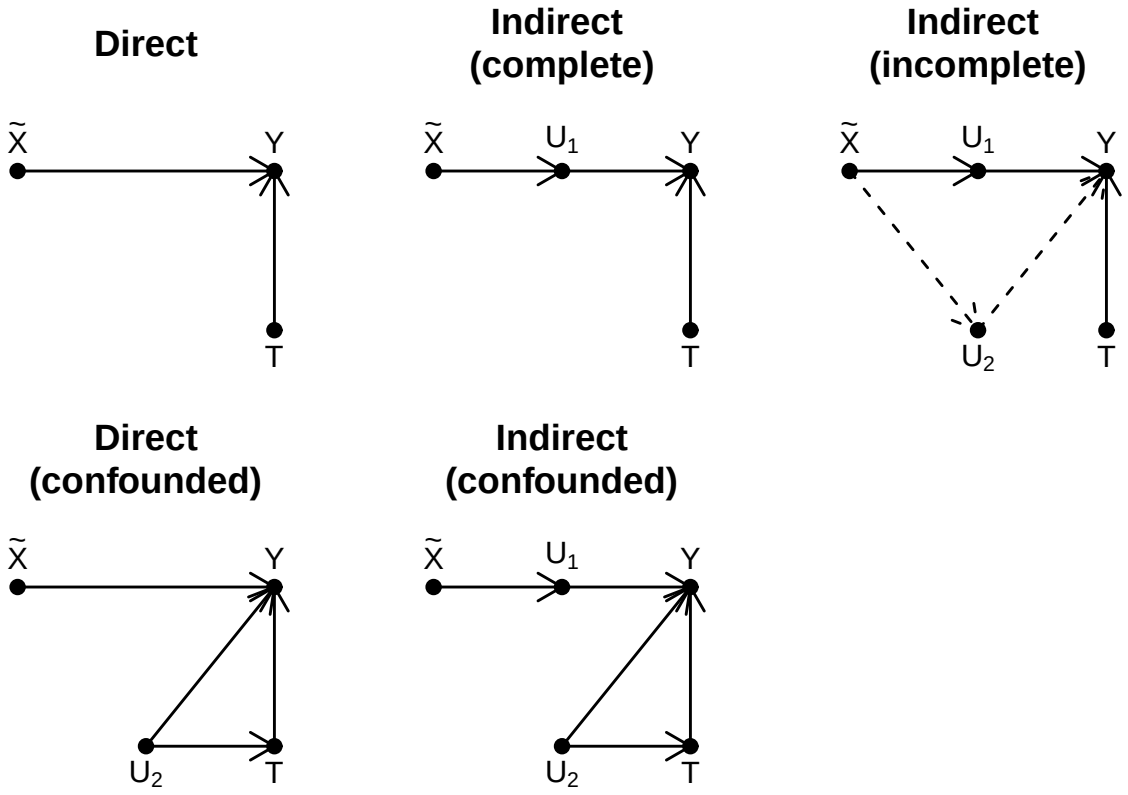

Figure 1: Causal diagrams for effect modification by sex with treatment T, outcome Y, effect modifier sex $\tilde{X}$, and observed prognostic factors U_1 and U_2.

Estimate	Combining Method	Covariate adjustment	Equation
$\delta_{\tilde{x}}^{pool}$	Conventional	(none)	(6)
$\delta_{\tilde{x}}^{Bayes}$	Hierarchical	(none)	(10)
$\delta_{\tilde{x}}^{meta}$	Meta-regression	(none)	(24)
$\delta_{\tilde{x},u}^{cpool}$	Conventional	conditional	(14)
$\delta_{\tilde{x},u}^{cBayes}$	Hierarchical	conditional	(19)
$\delta_{\tilde{x},u}^{cmeta}$	Meta-regression	conditional	(27)
$\delta_{\tilde{x}}^{mpool}$	Conventional	marginalized	(17)
$\delta_{\tilde{x}}^{mBayes}$	Hierarchical	marginalized	(21)
$\delta_{\tilde{x}}^{mmeta}$	Meta-regression	marginalized	(29)

Table 1: Description and equation numbers for all the proposed estimators

	Inclusion criteria			Baseline				Follow-up	
Study	Max LVEF	Min QR	NYHA	% CRT-D	% Women	% Ischemic etiology	Median 6-min walk (IQR)	Median 6-min walk (IQR)	% NYHA improved
CONTAK	35	120	II-IV	51	15	71	335	357	47
n=490, 1998-2000							(245-401)	(273-426)	
MIRACLE-ICD	35	130	II-IV	48	19	67	307	356	45
n=555, 1999-2001							(226-381)	(292-440)	
RAFT	30	120	II-III	52	16	67	369	395	41
n=1798, 2003-2009							(304-433)	(323-454)	
REVERSE	40	120	I-II	68	22	54	406	429	29
n=610, 2004-2006							(315-479)	(335-509)	
MADIT-CRT	30	130	I-II	62	25	53	376	389	29
n=1820, 2004-2008							(304-426)	(304-449)	

Table 2: Sample sizes, and enrollment periods, inclusion criteria, baseline covariates, and outcomes in our CRT-D trials. LVEF=left ventricular ejection fraction, NYHA=New York Heart Association class, CRT-D=cardiac resynchronization therapy with defibrillation.

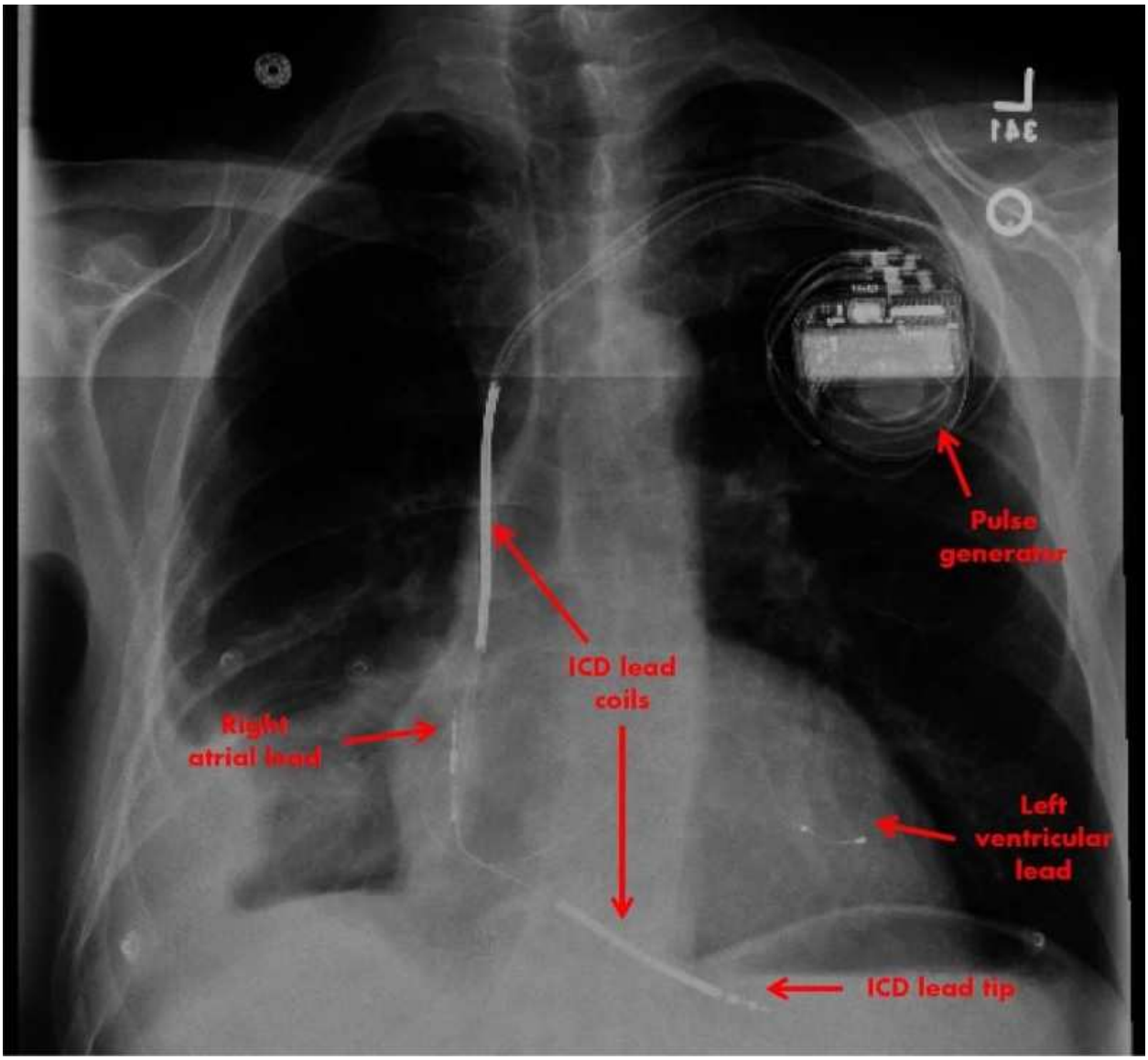

Figure 2: Chest x-ray of an implanted cardiac resynchronization defibrillator (CRT-D) system. The pulse generator sits in a pocket under the skin near the shoulder, and leads enter the patient's venous system near the clavicle. In addition to right atrial and left ventricular pacing leads, a high-voltage implantable cardioverter-defibrillator (ICD) lead goes to the right ventricular apex. Adapted with permission from Kramer et al. (2013).

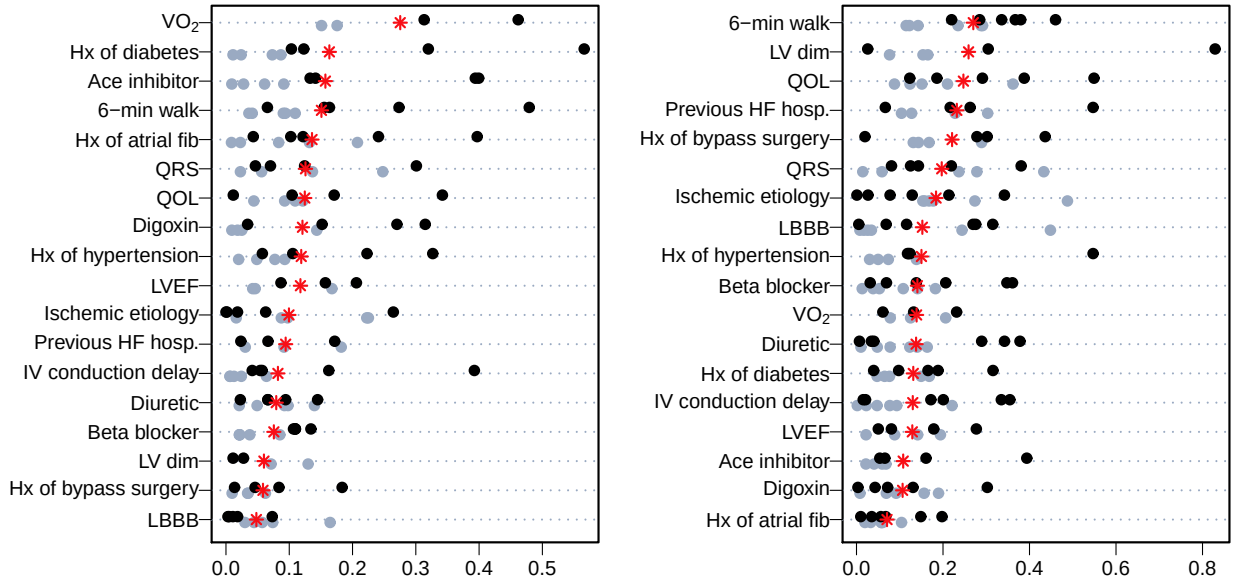

Figure 3: Differences in baseline measurements across arm (left) and outcome (right), by trial and sex. Each point is the absolute standardized mean difference (ASMD) between participants in the CRT-D and ICD arms (left) or improved and same/worse NYHA outcomes (right). Men are plotted in grey, women in black, and across-trial averages as asterisks.

Variable	Definition
Treatment arm T	CRT-D and ICD
Subgroups $\tilde{X}$	Female and male
Individual binary covariate U	Ischemic etiology and non-ischemic
Individual continuous covariate U	Baseline 6-minute walk distance
Binary outcome Y	NYHA improved and sameworse
Continuous outcome Y	6-minute walk distance
Trial covariate V	Proportion of patients with intraventricular conduction delay

Table 3: Variable definitions for implementing models of sex-specific treatment effects in CRT-D trials.

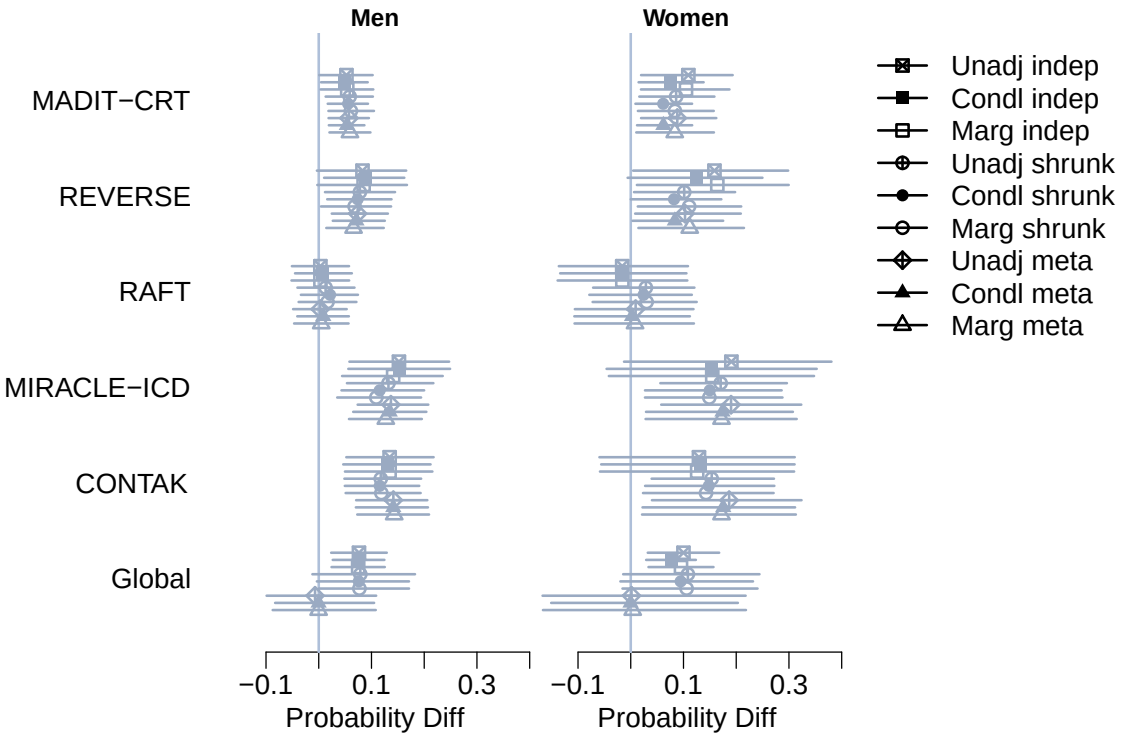

Figure 4: Differences (CRT-D minus ICD) in the probability of NYHA improvement for men (left) and women (right) in each Covariate is ischemic etiology and trial-level covariate is proportion of patients with intraventricular conduction delay. Plotting symbol indicates unadjusted (crossed), conditional (solid), and marginal (open) effects in independent (square), hierarchical Bayesian (circle), and meta-regression (triangle) models. There are no significant male-female differences across any models or trials.

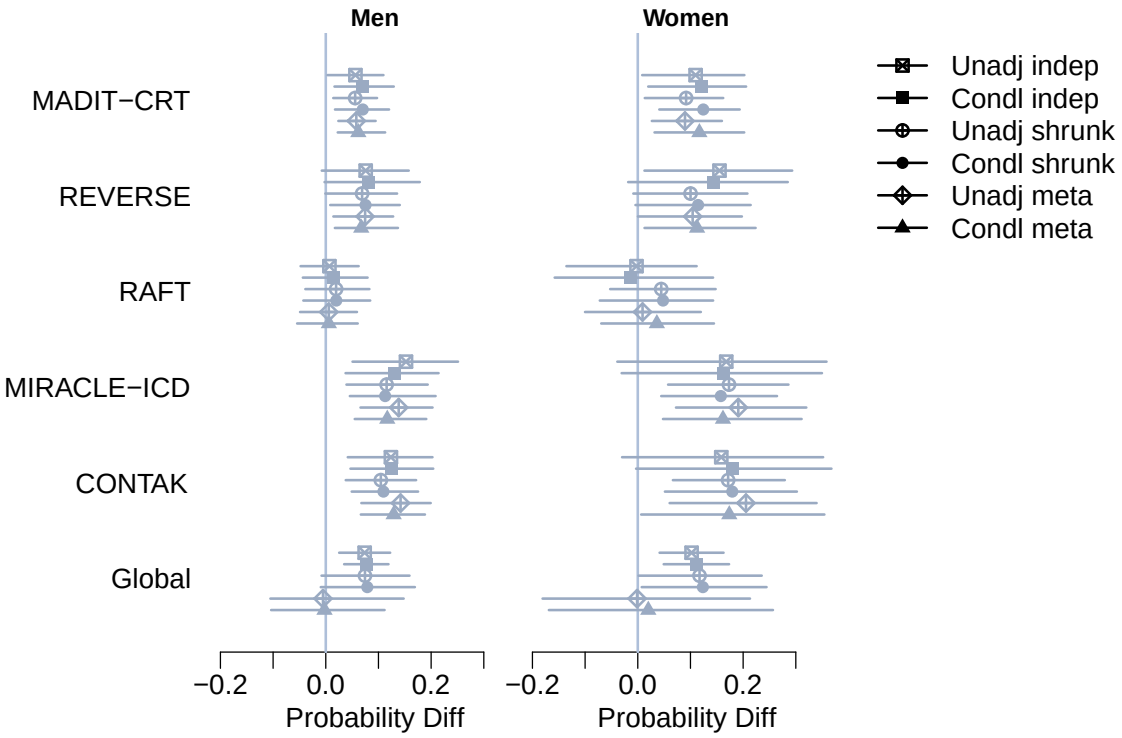

Figure 5: Differences (CRT-D minus ICD) in the probability of NYHA improvement for men (left) and women (right) in each trial. Covariate is baseline 6-minute walk distance and trial-level regressor is proportion of patients with intraventricular conduction delay. Plotting symbol indicates unadjusted (crossed), conditional (solid), and marginal (open) effects in independent (square), hierarchical Bayesian (circle), and meta-regression (triangle) models. There are no significant male-female differences across these models.

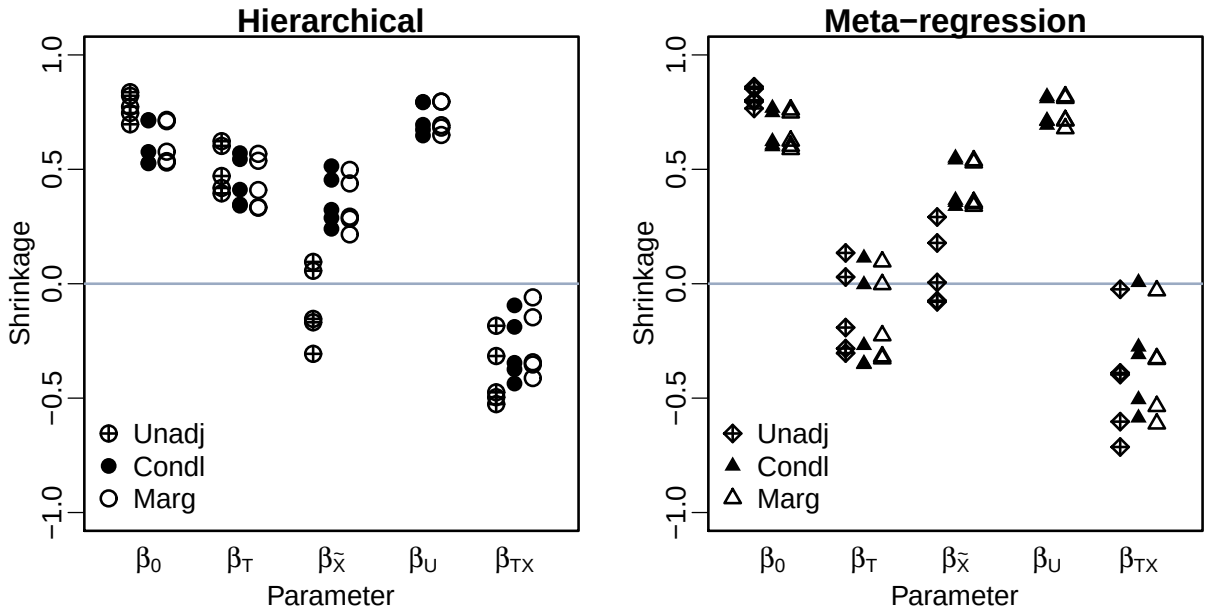

Figure 6: Shrinkage factors for each parameter in each trial of the hierarchical (left) and mega-regression (right) probit models applied to the NYHA improvement outcome.

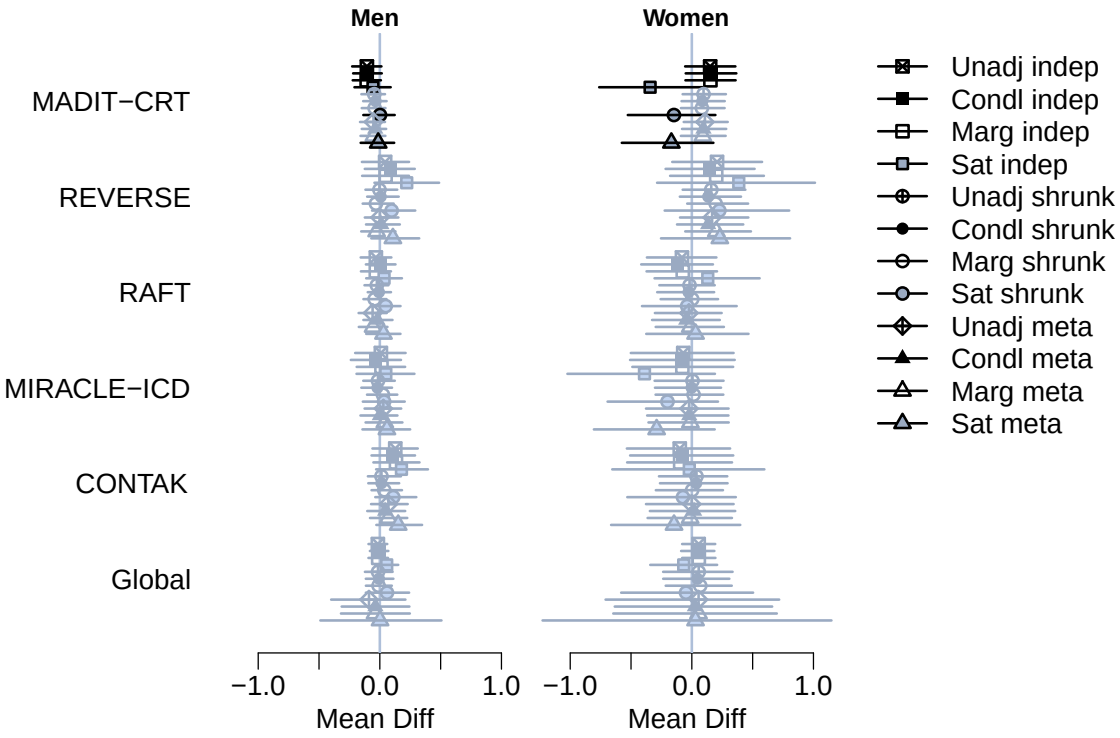

Figure 7: Difference (CRT-D minus ICD) in mean 6-minute walk distance for men (left) and women (right) in each trial. Covariate is ischemic etiology and trial-level regressor is proportion of patients with intraventricular conduction delay. Plotting symbol indicates unadjusted (crossed), conditional (solid), marginal (open), and saturated (bordered) effects in independent (square), hierarchical Bayesian (circle), and meta-regression (triangle) models. Effects from model with significant male-female differences in treatment effects are plotted in black.

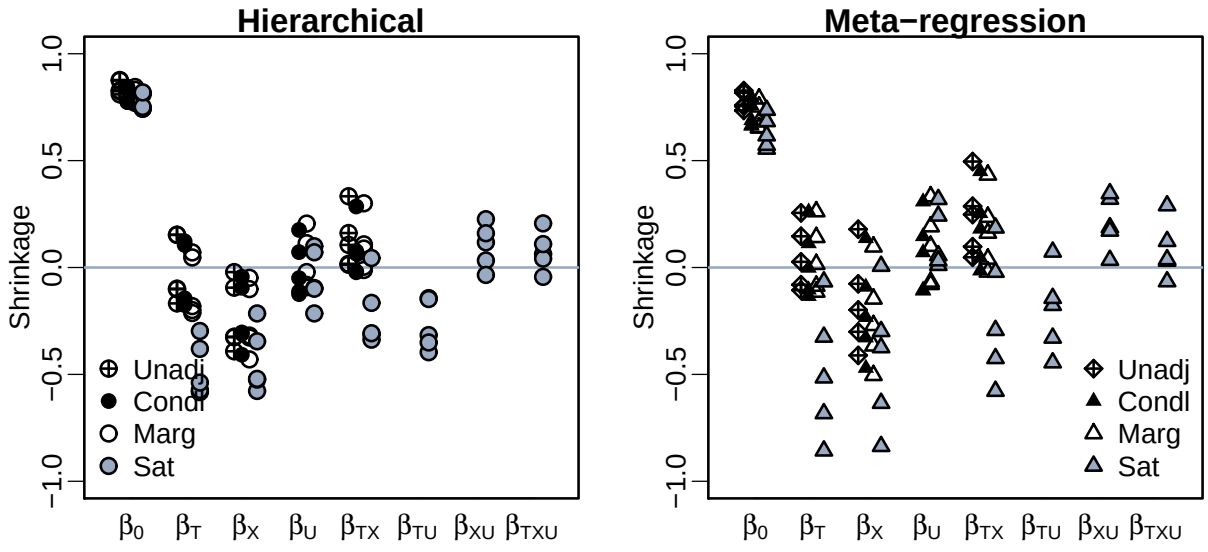

Figure 8: Shrinkage factors for each parameter in each trial of the hierarchical (left) and mega-regression (right) linear models applied to the 6-minute walk distance.