Analysis of Subgroup Effects in Randomized Trials When Subgroup Membership is Informatively Missing: Application to the MADIT II Study

Daniel Scharfstein
*The Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland U.S.A.*

Georgiana Onicescu
*Medical University of South Carolina, Charleston, South Carolina, U.S.A.*

Steven Goodman
*The Johns Hopkins School of Medicine, Baltimore, Maryland, U.S.A.*

Rachel Whitaker
*The Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland U.S.A.*

Summary.
In this paper, we develop and implement a methodology for drawing inference about subgroup effects in a two-arm randomized trial when subgroup status is only known for a non-random sample in each of the trial arms. The methodology is developed in the context of the MADIT II study, a randomized trial designed to evaluate the effectiveness of implantable defibrillators on survival.

*Keywords:* Bounds, Expert Opinion, Identifiability

1. Introduction

Missing data is a hallmark of most clinical trials. Whether it appreciably affects the subsequent inference depends on the extent and pattern of the missing data, the reasons for missingness, and the statistical procedures used to estimate the effect of interest and its uncertainty. It is a particularly difficult challenge because estimates of effect and their uncertainty depend on assumptions generally not testable from the observed data. Nonetheless, in biomedical settings, clinicians can sometimes provide plausible explanations for the missingness. Analyses that incorporate this clinical information can provide statistically and scientifically plausible bounds on effect estimates, facilitating the translation of clinical knowledge into quantitative inference.

In this paper, we present an example of a high-stakes medical technology being considered for reimbursement by government agencies and other payors. During the evaluation hearing, the discussion of the results of a pivotal two-arm randomized clinical trial focused on inference about a specific subgroup effect. Unfortunately, the trial was not designed to collect information on subgroup membership. Nonetheless, subgroup membership was recorded on a non-random sample of individuals in the trial, with the vast majority of these observations in the intervention arm. We present a statistical approach that incorporates plausible clinical/biological information to see what could have been learned about the subgroup effect in the presence of these missing covariate data.
Our approach is related, in spirit, to the work on bounds developed by Horowitz and Manski (2000), with two main exceptions. Their approach makes no assumptions and, thus, represents the “worst-case” inference about subgroup effects. In contrast, our approach seeks more precise inferences by introducing a set of plausible, application-specific assumptions. For a given observed data distribution, these assumptions serve to further constrain the set of compatible, complete data distributions. Horowitz and Manski (2000) characterize the sampling uncertainty associated with the observed data distribution through resampling-based methods; we adopt a Bayesian approach using vague priors.

The paper is organized as follows. In Section 2, we present background information about the medical technology and the issues surrounding the missing subgroup information. This background is essential for understanding our application-specific assumptions. In Section 3, we introduce our methodology. Section 4 presents an analysis of data from the pivotal study. The final section is devoted to a discussion.

2. Clinical Background

Sudden cardiac death (SCD) is typically initiated by ventricular tachycardia (VT), where the ventricles of the heart beat at a very rapid rate with inadequate pumping, and then progresses to ventricular fibrillation (VF), where the coordinated beating of the heart disintegrates into uncoordinated, ineffective contractions (Winslow, Mehta, and Fuster, 2005); death then results from cerebral hypoxia (Zipes, 2005). Pharmacological means of reducing VT or VF are only partially successful. The only effective treatment is defibrillation via electrical shock to the heart at the time of an arrhythmic event. The implantable cardiac defibrillator (ICD) was developed to monitor the heart rate and rhythm continuously, recognize VT and VF, and deliver corrective defibrillatory discharges when necessary (Mirowski, Mower, and Reed, 1980). The FDA first approved ICDs in 1985 and Medicare initiated coverage of ICDs in 1989, but only for an extremely limited subpopulation with particularly severe indications.

During the 1990s and early 2000s, a series of randomized clinical trials, comparing ICD to medical (drug) management, attempted to establish indications for implantation, i.e., to assess the existence and magnitude of benefit in different subpopulations. One of these trials was the Multicenter Automatic Defibrillator Intervention Trial (MADIT-I), comparing ICD to medical (drug) management (Moss et al., 1996). A key eligibility criteria included sustained VT or VF that was inducible by electrophysiological stimulation (EPS) and not suppressed by the administration of an antiarrhythmic agent. Patients with this condition are said to be “inducible.” 196 patients were enrolled in this trial, and ICDs were shown to have a dramatic effect on all-cause mortality; reducing it from 39% to 16%, with a hazard ratio of 0.46 (p=0.009) over a 27-month mean follow-up period. In 1999, based on the results of this and other trials, Medicare expanded coverage to inducible patients. A follow-up to the MADIT-I trial was initiated in 1997. This trial, dubbed MADIT-II, differed from its predecessor in that EPS testing was not required. The trial enrolled 1232 patients and was stopped due to efficacy in 2001. With an average follow-up time of 20 months, the ICD group exhibited a statistically and clinically significant 31% reduction in the overall mortality hazard (p=0.016), with 14.2% and 19.8% of patients who dying in the ICD and control arms, respectively.
Based on the results of MADIT-II, the Centers for Medicare and Medicaid Services (CMS) was asked to expand Medicare (United States federal health insurance for persons age 65 or older) coverage of ICDs to the MADIT-II population, including those without inducible arrhythmias. In their evaluation of the MADIT-II trial, CMS wanted to be assured that the benefit in the expanded subpopulation was not driven solely by the benefit to inducible patients already established in MADIT-I, especially since the observed effect (HR=0.69) was not as sizable as it was in the MADIT-I trial (HR=0.46). An analysis stratified by inducibility status was not possible, since EPS testing was not a prerequisite for enrollment into the study nor was it scheduled to be routinely performed post-randomization. Due to ICD reimbursement policies at the time and the invasive nature of EPS, a sizable fraction (79.3%) of patients enrolled in the ICD arm had EPS results recorded in the database provided to CMS; in contrast, only 2.4% of patients in the control arm had EPS recorded.

Interestingly, it is likely that the missing inducible status is informative. Many clinicians were reluctant to enroll their patients in MADIT II, already being convinced that ICDs benefited patients who who were inducible. So some physicians performed EPS testing, and if the patient was found to be inducible, implanted an ICD and did not enroll them in MADIT II. Thus, patients who were identified as non-inducible were more likely to be referred to MADIT II. Further, EPS testing post-randomization was more likely to occur in those who did not have prior EPS result. Since only post-randomization EPS test results were recorded in the database provided to CMS and pre- and post- EPS results are likely to be positively correlated, the patients, in each treatment group, who were not tested after randomization are more likely to be non-inducible than those with recorded results. It was not anticipated that treatment assignment would affect EPS inducibility status, so that a subgroup analysis based on post-randomization inducibility status was not considered by CMS as problematic.

The decision tree in Figure 1 provides a stylized illustration of why patients with post-randomization EPS results are more likely to be inducible than those without results. In this illustration, we assume all patients "Referred" are randomized and we ignore treatment assignment. Further we assume that, among those with EPS testing prior to randomization, patients who are non-inducible have 16 times the odds of being referred than those who are inducible. From the decision tree, we see that the conditional probability of being inducible after randomization given referral and EPS testing post-randomization is 0.31. In contrast, the conditional probability of being inducible given referred and not EPS tested is 0.23. These computations demonstrate that EPS inducibility status measured post-randomization is associated with EPS testing post-randomization.

3. Methodology

In Table 1, we display our notation. The goal of our analysis is to use the observed data to draw inference about the relative risk of death (ICD vs. control) for inducible ($RR_1$) and non-inducible ($RR_0$) patients. The observed data for each of 1232 individuals are $T$, $Y$, $M$ and, $I$ if $M = 0$. From these data, the tilde ($\tilde{\cdot}$) parameters are identifiable (i.e., estimable) from the distribution of the observed data. In contrast, the relative risks are not identifiable. We can, however, construct
Fig. 1. Example of informative missing inducibility status
Subgroup Effects in Randomized Trials When Subgroup Membership is Informatively Missing

Table 1. Notation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>Treatment group indicator (1 for ICD, 0 for control)</td>
</tr>
<tr>
<td>I</td>
<td>Inducible indicator (1 for inducible, 0 for non-inducible)</td>
</tr>
<tr>
<td>M</td>
<td>Missing EPS data indicator (1 for missing, 0 for observed)</td>
</tr>
<tr>
<td>Y</td>
<td>Death indicator (1 for dead, 0 for alive)</td>
</tr>
</tbody>
</table>

Parameter Definition

\[ p_{YIM|T=t}(y, i, m) = P[Y = y, I = i, M = m | T = t] \]
\[ \tilde{p}_{Y|T=t, M=0}(y, i) = P[Y = y, I = i | T = t, M = 0] \]
\[ \tilde{p}_{Y|T=t, M=1}(y) = P[Y = y | T = t, M = 1] \]
\[ \tilde{p}_M|T=t(m) = P[M = m | T = t] \]
\[ RR_i = \frac{\sum_{m=0}^{1} p_{YIM|T=t}(1, i, m)}{\sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=t}(1, i, m)} / \frac{\sum_{m=0}^{1} p_{YIM|T=0}(1, i, m)}{\sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=0}(1, i, m)} \]

To ensure that the independent variables form a proper probability distribution, we impose the following linear equality and inequality constraints

\[ \sum_{y=0}^{1} \sum_{i=0}^{1} \sum_{m=0}^{1} p_{YIM|T=t}(y, i, m) = 1 \quad \text{for } t = 0, 1 \]  
(1)

and

\[ 0 \leq p_{YIM|T=t}(y, i, m) \leq 1 \quad \text{for } y, i, m, t = 0, 1 \]  
(2)

The next set of constraints result from the fact that certain features of the treatment-specific joint distribution of Y, I and M are identifiable from the distribution of the observed data, which we treat as known for the moment. Specifically, we have the following linear equality constraints:

\[ \sum_{y=0}^{1} \sum_{i=0}^{1} p_{YIM|T=t}(y, i, m) = \tilde{p}_M|T=t(m) \quad \text{for } m, t = 0, 1 \]  
(3)

\[ p_{YIM|T=t}(y, i, 0) = \tilde{p}_M|T=t(0)\tilde{p}_{Y|M=0,T=t}(y, i) \quad \text{for } y, i, t = 0, 1 \]  
(4)

\[ \sum_{i=0}^{1} p_{YIM|T=t}(y, i, 0) = \tilde{p}_M|T=t(1)\tilde{p}_{Y|M=1,T=t}(y, i) \quad \text{for } y, t = 0, 1 \]  
(5)

The final set of constraints are based on five plausible scientific assumptions. The plausibility of these assumptions stem from our reading of the transcript from the 2003 Medicare Coverage Advisory Committee Meeting as well as personal communications with (1) Dr. Hugh Calkins, Professor of Medicine, Director of the
Arrhythmia Service and Clinical Electrophysiology Laboratory, Johns Hopkins University, (2) Dr. Joseph Chin, Director, Coverage and Analysis Group, CMS, and (3) Dr. Alfred Buxton, Director, Cardiology Division, Rhode Island and Miriam Hospitals, Professor of Medicine, Brown University.

Assumption 1: Treatment is randomized and treatment does not causally affect EPS inducibility status measured post-randomization.

Statistically, this means that \( T \) is independent of \( I \) (i.e., \( P[I = i|T = 1] = P[I = i|T = 0] \)). Thus, we have the linear equality constraint:

\[
\sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=1}(y, i, m) = \sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=0}(y, i, m) \quad \text{for } i = 0, 1 \quad (6)
\]

Assumption 2: Within each treatment group, there is a higher proportion of non-inducibles among patients with missing EPS results than those with recorded results.

This assumption \( (P[I = 0|M = 1, T = t] \geq P[I = 0|M = 0, T = t]) \) derives from our discussion in the previous section and can be expressed as the linear inequality constraints:

\[
\frac{\sum_{y=0}^{1} p_{YIM|T=t}(y, 0, 1)}{\tilde{p}_{M|T=t}} \geq \frac{\sum_{y=0}^{1} p_{YIM|T=t}(y, 0, 0)}{\tilde{p}_{M|T=t}} \quad \text{for } t = 0, 1 \quad (7)
\]

Assumption 3: With control treatment, the proportion dying among inducibles is greater than among non-inducibles.

This assumption \( (P[Y = 1|I = 1, T = 0] \geq P[Y = 1|I = 0, T = 0]) \) is justified based on prior trials such as MADIT-I and CABG-Patch. It can be expressed as the non-linear inequality constraint:

\[
\frac{\sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=0}(1, 1, m)}{\sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=0}(y, 1, m)} \geq \frac{\sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=0}(1, 0, m)}{\sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=0}(y, 0, m)} \quad (8)
\]

Assumption 4: With control treatment, the proportion expected to die (over an average 20 month follow-up period) among inducibles and non-inducibles is between 5 and 50 percent.

This very conservative assumption \( (0.05 \leq P[Y = 1|I = i, T = 0] \leq 0.50) \) is based on the results of prior trials such CABG-Patch, MADIT-I and can be expressed as the linear equality constraints:

\[
0.05 \leq \frac{\sum_{m=0}^{1} p_{YIM|T=0}(1, i, m)}{\sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=0}(y, i, m)} \leq 0.50 \quad \text{for } i = 0, 1 \quad (9)
\]

Assumption 5: For inducibles, the proportion dying is lower for those treated with an ICD than those treated with control.

This assumption \( (P[Y = 1|I = I, T = 1] \leq P[Y = 1|I = 1, T = 0]) \) is justified by the results of MADIT-I. It can be expressed as the non-linear inequality constraint:

\[
\frac{\sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=1}(1, 1, m)}{\sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=1}(y, 1, m)} \leq \frac{\sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=0}(1, 1, m)}{\sum_{y=0}^{1} \sum_{m=0}^{1} p_{YIM|T=0}(y, 1, m)} \quad (10)
\]
The lower (upper) bounds are derived by minimizing (maximizing) \( RR_i \) subject to constraints (1)-(10). The resulting bounds are a function of the distribution of the observed data (i.e., \( \tilde{\text{parameters}} \)). Computationally, we utilize the function \texttt{fmincon} in MATLAB to perform the constrained optimization.

We adopt a Bayesian approach to account for uncertainty. Specifically, we assume independent Dirichlet priors on \( \tilde{p}_{Y|T=t,M=0}(\cdot, \cdot) \), \( \tilde{p}_{Y|T=t,M=1}(\cdot) \), and \( \tilde{p}_{M|T=t}(\cdot) \), where all the prior parameters are all set to 1. Let \( \tilde{N}_{Y|T=t,M=0}(y,i) = \sum_j 1(Y_j = y, I_j = i, T_j = t, M_j = 0) \), \( \tilde{N}_{Y|T=t,M=1}(y) = \sum_j 1(Y_j = y, T_j = t, M_j = 1) \), and \( \tilde{N}_{M|T=t}(m) = \sum_j 1(M_j = m, T_j = m) \), where the subscript \( j \) denotes individual \( j \) and \( 1(A) \) takes on the value 1 if the event \( A \) is true and 0 otherwise. Given \( \tilde{p}_{Y|T=t,M=0}(\cdot, \cdot), \tilde{p}_{Y|T=t,M=1}(\cdot) \), and \( \tilde{p}_{M|T=t}(\cdot) \), \( \tilde{N}_{Y|T=t,M=0}(\cdot, \cdot), \tilde{N}_{Y|T=t,M=1}(\cdot) \), and \( \tilde{N}_{M|T=t}(\cdot) \) are independent multinomials with parameters \( \tilde{p}_{Y|T=t,M=0}(\cdot, \cdot), \tilde{p}_{Y|T=t,M=1}(\cdot) \), and \( \tilde{p}_{M|T=t}(\cdot) \), respectively. Then the posterior of \( \tilde{p}_{Y|T=t,M=0}(\cdot, \cdot), \tilde{p}_{Y|T=t,M=1}(\cdot) \), and \( \tilde{p}_{M|T=t}(\cdot) \) are independent Dirichlets with parameters \( \tilde{N}_{Y|T=t,M=0}(\cdot, \cdot) + 1, \tilde{N}_{Y|T=t,M=1}(\cdot) + 1 \), and \( \tilde{N}_{M|T=t}(\cdot) + 1 \), respectively. The joint posterior of the bounds are found by simulating from the posterior of \( \tilde{p}_{Y|T=t,M=0}(\cdot, \cdot), \tilde{p}_{Y|T=t,M=1}(\cdot) \), and \( \tilde{p}_{M|T=t}(\cdot) \) and, for each simulate, performing constrained optimizations to solve for lower and upper bounds on \( RR_0 \) and \( RR_1 \).

4. Results

Table 2 presents the observed data. Figures 2-6 display 25,000 simulates (dots) from the joint posteriors of the minimum and maximum of \( RR_i \) (\( i = 0 \), left panel; \( i = 1 \), right panel), under no assumptions, assumption 1, assumptions 1-2, assumptions 1-3, assumptions 1-4, and assumptions 1-5, respectively. In each panel, the solid line represents a 95\% highest posterior density confidence set and the circle represents the posterior mode. Table 3 summarizes the posterior modal bounds and modal length of bounds under the six sets of assumptions. In addition, this table displays the posterior probability that the maximum relative risk for each inducibility stratum is less than or equal to 1.

As expected, the figures and table show that the precision of the inferences non-decreases as more assumptions are imposed. In our analysis, Assumption 2 (regarding missing inducibility status) does not add any additional precision above and beyond Assumption 1 (regarding randomization). However, Assumption 2 does increase precision if Assumption 1 is removed.

Even under the strongest assumptions, there is not enough evidence to conclude that ICD’s are effective among non-inducibles. In contrast, under Assumptions 1-3 alone, there is 65\% posterior probability that the maximum relative risk among inducibles is less than or equal to one.

We also evaluated whether there is statistical evidence to suggest the effect of ICD’s is different between inducible and non-inducibles. To address this question, we computed the posterior distribution of the minimum and maximum of the ratio of relative risks of death (ICD vs. control) for inducibles vs. non-inducibles under the various sets of assumptions. Under Assumptions 1-5, the posterior probability that the maximum of the ratio was less than one (suggesting that the effect of ICD’s is greater for inducibles than non-inducibles) is 1.67\%. Thus, the data do not provide evidence to suggest a differential effect.
Table 2. Observed data

<table>
<thead>
<tr>
<th></th>
<th>( M = 0 )</th>
<th>( M = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T = 0 ) (( N_{M</td>
<td>T=0} ) = 12)</td>
<td>( T = 1 ) (( N_{M</td>
</tr>
<tr>
<td>( Y = 0 )</td>
<td>( \tilde{N}_{Y</td>
<td>M=0,T=0}(0) = 4 )</td>
</tr>
<tr>
<td>( Y = 1 )</td>
<td>( \tilde{N}_{Y</td>
<td>M=0,T=0}(1) = 6 )</td>
</tr>
<tr>
<td>( I = 0 )</td>
<td>( \tilde{N}_{Y</td>
<td>M=0,T=0}(0,0) = 4 )</td>
</tr>
<tr>
<td>( I = 1 )</td>
<td>( \tilde{N}_{Y</td>
<td>M=0,T=0}(0,1) = 6 )</td>
</tr>
</tbody>
</table>

\( \tilde{N}_{Y|M=0,T=1}(1,0) = 62 \)
Table 3. Posterior modes of the bounds and length of bounds under varying sets of assumptions, stratified by inducibility status. Posterior probabilities that the maximum relative risk in each stratum is less than or equal one. Posterior probability that the minimum relative risk among non-inducibles is greater than or equal to 0.46, a reported relative risk.

<table>
<thead>
<tr>
<th>Assumptions</th>
<th>Inducibles</th>
<th>Non-Inducibles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bounds</td>
<td>Length</td>
</tr>
<tr>
<td>-</td>
<td>[0.06, 26.84]</td>
<td>21.72</td>
</tr>
<tr>
<td>(1)</td>
<td>[0.13, 11.19]</td>
<td>8.83</td>
</tr>
<tr>
<td>(1)-(2)</td>
<td>[0.13, 11.19]</td>
<td>8.83</td>
</tr>
<tr>
<td>(1)-(3)</td>
<td>[0.13, 0.89]</td>
<td>0.78</td>
</tr>
<tr>
<td>(1)-(4)</td>
<td>[0.16, 0.89]</td>
<td>0.75</td>
</tr>
<tr>
<td>(1)-(5)</td>
<td>[0.19, 1.00]</td>
<td>0.75</td>
</tr>
</tbody>
</table>
5. Discussion

At the CMS hearing, Guidant Corporation, the sponsor of MADIT-II, argued that coverage should not be based on inducibility status, and made available previously undisclosed data on 257 patients who were identified to be non-inducible based on pre-enrollment EP testing. Of these patients, 113 were randomized to the control arm and 114 to the ICD arm (Anderson, 2003). For this possibly non-representative subset of non-inducible patients, they reported that 19.5% died under conventional therapy versus only 9% in the ICD group, yielding a relative risk of death for non-inducibles of 0.46. As can be seen in the last column of Table 3, under Assumptions (1)-(3) alone, the posterior probability that the minimum relative risk is greater than or equal to 0.46 is 97%. Thus, our analysis provides strong evidence that their reported point estimate is too optimistic. Of course, our analysis cannot rule out more modest beneficial effects or even harmful effects of ICD’s among non-inducibles.

Panel members tended to find the analysis of the pre-randomization inducibility data compelling, and unanimously voted that the “evidence is adequate to draw conclusions about health outcomes in patients identical to the patients enrolled in the MADIT II trial.” Risk stratification based on inducibility status was not part of their recommendation (Anderson, 2003).

While worst case bounds can often be too wide to practically useful, we have shown that by introducing clinically plausible, conservative assumptions, the precision of the inferences can be greatly increased. Our bounding methodology has general applicability to the setting of missing low-dimensional outcomes and covariates, like that considered by Horowitz and Manski (2000).

References


Fig. 2. Joint posterior distribution of minimum and maximum of $RR_i$, under no assumptions. Dots denote simulations. Solid line denotes 95% credible set. Circle denotes mode. Left (right) panel refers to the relative risk for non-inducibles (inducibles).
Fig. 3. Joint posterior distribution of minimum and maximum of $RR_i$, under assumption (1). Dots denotes simulates. Solid line denotes 95% credible set. Circle denotes mode. Left (right) panel refers to the relative risk for non-inducibles (inducibles).
Fig. 4. Joint posterior distribution of minimum and maximum of $RR_i$, under assumptions (1) and (2). Dots denotes simulates. Solid line denotes 95% credible set. Circle denotes mode. Left (right) panel refers to the relative risk for non-inducibles (inducibles).
**Fig. 5.** Joint posterior distribution of minimum and maximum of $RR_i$, under assumptions (1), (2), and (3). Dots denotes simulates. Solid line denotes 95% credible set. Circle denotes mode. Left (right) panel refers to the relative risk for non-inducibles (inducibles).
Fig. 6. Joint posterior distribution of minimum and maximum of $RR_i$, under assumptions (1), (2), (3) and (4). Dots denotes simulates. Solid line denotes 95% credible set. Circle denotes mode. Left (right) panel refers to the relative risk for non-inducibles (inducibles).
Fig. 7. Joint posterior distribution of minimum and maximum of $RR_i$, under assumptions (1), (2), (3), (4) and (5). Dots denote simulates. Solid line denotes 95% credible set. Circle denotes mode. Left (right) panel refers to the relative risk for non-inducibles (inducibles).


