A simple method for analyzing data from a randomized trial with a missing binary outcome
Stuart G Baker* 1 and Laurence S Freedman2

Address: 1Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, USA and 2Department of Mathematics, Statistics, and Computer Science, Bar Dan University, Ramat Gan 52900, Israel
Email: Stuart G Baker* - sb16i@nih.gov; Laurence S Freedman - lsf@actcom.co.il
* Corresponding author

Abstract

Background: Many randomized trials involve missing binary outcomes. Although many previous adjustments for missing binary outcomes have been proposed, none of these makes explicit use of randomization to bound the bias when the data are not missing at random.

Methods: We propose a novel approach that uses the randomization distribution to compute the anticipated maximum bias when missing at random does not hold due to an unobserved binary covariate (implying that missingness depends on outcome and treatment group). The anticipated maximum bias equals the product of two factors: (a) the anticipated maximum bias re were complete confounding of the unobserved covariate with treatment group among subjects with an observed outcome and (b) an upper bound factor that depends only on the fraction missing in each randomization group. If less than 15% of subjects are missing in each group, the upper bound factor is less than .18.

Results: We illustrated the methodology using data from the Polyp Prevention Trial. We anticipated a maximum bias under complete confounding of .25. With only 7% and 9% missing in each arm, the upper bound factor, after adjusting for age and sex, was .10. The anticipated maximum bias of .25 × .10 = .025 would not have affected the conclusion of no treatment effect.

Conclusion: This approach is easy to implement and is particularly informative when less than 15% of subjects are missing in each arm.

Background

Missing outcome data are common in clinical studies [1,2]. Many approaches assume missing at random (MAR) as a base case. MAR means that the probability of missing depends only on observed variables [3]. Four strategies for examining the bias or sensitivity of results when MAR does not hold are to (i) fit all saturated MAR and non-MAR missing-data models [4,5], (ii) add a parameter to various MAR models to make them non-MAR and test if the fit is significantly improved [6,7], (iii) impute the missing data in one arm using the observed proportion of events in the other arm [8,9], (iv) estimate results under a non-MAR missing-data mechanism with key parameters specified by the investigator [1,10]-[13].

We propose a variation of method (iv) for randomized trials with binary outcome that explicitly uses the randomization distribution to reduce user input. To our knowledge this is the only method that exploits the randomization distribution for missing-data adjustment.
We illustrate the methodology using data from the Polyp Prevention Trial (PPT) in which 2079 men and women with recently removed colorectal adenoma were randomized to receive either intensive counseling to adopt a low-fat diet (intervention) or a standard brochure on healthy eating (control) [14]. The binary outcome was at least one adenoma detected on colonoscopy following randomization. In the control arm 9% of the subjects were missing the outcome, and in the intervention arm 7% were missing the outcome. Dropping the data from subjects with a missing outcome gives an estimated difference of -0.002 (s.e. = .022) in the probability of adenoma recurrence between the intervention and control groups. Thus there was very little evidence that intensive counseling to adopt a low-fat diet reduced the probability of adenoma recurrence. An important question was whether or not an adjustment for the missing outcomes would have changed this conclusion.

Methods

Adjusting for Observed Covariates

As a starting point, we assume the data are missing at random (MAR). Let \( Y \) denote the binary outcome of adenoma recurrence. Let \( Z = 0 \) denote random assignment to the control group and \( Z = 1 \) denote random assignment to the intervention group. Also let \( R = 0 \) if the outcome is missing and 1 if the outcome is observed. Suppose we also have data on the observed variable \( S \), which represents either strata formed by the cross-classification of categorical baseline covariates or outpoints of a continuous variable. Under the MAR assumption, the probability of missing depends on \( Z \) and \( S \) but not \( Y \), namely,

\[
pr(R = 1|Z, s, Y = 1) = pr(R = 1|Z, s). \tag{1}
\]

Because \( R \) and \( Y \) are conditionally independent given \( Z \) and \( S \), it follows from (1) that

\[
pr(Y = 1|Z, s, R = 1) = pr(Y = 1|Z, s). \tag{2}
\]

In other words, under the MAR assumption in (1), the probability of adenoma recurrence conditional on treatment assignment and baseline covariates is the same in all subjects as in subjects not missing outcome. Baker and Laird [6] proved the related result that under MAR the maximum likelihood estimate of the probability of outcome conditional on covariates is the same in all subjects as in subjects not missing outcome.

With binary outcomes, the overall measure of treatment effect is typically a difference, a relative risk, or an odds ratio. We focus on the difference because it is easy to interpret [15] and because it simplifies our formulation. Let \( \Delta_s \) denote the treatment effect for stratum 5, namely

\[
\Delta_s = pr(Y = 1|Z = 1, s) - pr(Y = 1|Z = 0, s). \tag{3}
\]

By virtue of the randomization \( pr(S = s|Z = 1) = pr(S = s|Z = 0) = pr(S = s) \). Therefore we can write the overall treatment effect as

\[
\Delta = \sum_s \Delta_s pr(S = s). \tag{4}
\]

If the missing-data mechanism is given in (1), then from (2), the treatment effect in stratum \( s \) (3) equals the treatment effect in stratum \( s \) among subjects with observed outcomes,

\[
\Delta_s = pr(Y = 1|Z = 1, s, R = 1) - pr(Y = 1|Z = 0, s, R = 1). \tag{5}
\]

Let \( n_{sz} \) denote the number of subjects in treatment group \( z \) and stratum \( s \) who have observed outcome \( y \). Based on (5), we estimate \( \Delta_s \) by

\[
d_s = q_{s1} - q_{s0}, \quad q_{sz} = n_{sz1}/n_{sz}, \tag{6}
\]

where \( *+* \) denotes summation over the indicated subscript. Let \( N_{sz} \) denote the number of subjects (with either observed or missing outcomes) in treatment group \( z \) and stratum \( s \). We estimate \( pr(S = s) \) by \( w_s = N_{s+}/N_{s+} \), giving an overall estimate of treatment effect,

\[
\hat{\Delta} = \sum s d_s w_s . \tag{7}
\]

The estimate in (7) is closely related to the estimate proposed by Horvitz and Thompson [16]. It is also maximum likelihood because it is a function of maximum likelihood estimates of the parameters. Using the delta method, and noting that \( \hat{\Delta} = d_1 w_1 + d_2 w_2 + ... + d_h w_h + d_h (1 - \sum_{s=1}^{h-1} w_s) \), we obtain

\[
\hat{\omega} (\hat{\Delta}) = \sum_{s=1}^{h} \left( \frac{d\hat{\Delta}}{d\delta_s} \right)^2 \hat{\omega} (\delta_s) + \sum_{s=1}^{h-1} \left( \frac{d\hat{\Delta}}{d\delta_s} \right)^2 \hat{\omega} (\delta_s) = \sum_{s=1}^{h} w_s^2 \sum_{s=1}^{h-1} q_{sz} (1 - q_{sz}) / n_{sz} + \sum_{s=1}^{h-1} (d_s - d_h) \hat{w}_s (1 - \hat{w}_s) / N_{s+}. \tag{8}
\]

where \( w_h = 1 - \sum_{s=1}^{h-1} \hat{w}_s \).

Bias from an omitted binary covariate

Suppose that instead of (1), the probability of missingness depends on treatment assignment, baseline strata, and an unobserved binary covariate \( x \). For our example from the Polyp Prevention Trial, \( x \) could be an unreported indicator of a family history of colon cancer. Then

\[
pr(R = 1|z, s, x, Y = 1) = pr(R = 1|z, s, x). \tag{9}
\]
In other words the data would be MAR if $x$ were observed. The model in (9) implies that, when $x$ is not observed, missingness depends on outcome and on treatment group via

$$
pr(R = 1|z,s,y) = \frac{\sum pr(R = 1|z,s,x)pr(y|z,s,x)pr(x|s)}{\sum pr(y|z,s,x)pr(x|s)}.
$$

(10)

We assume that for each level of $x$ within stratum $s$, the treatment effect is the same, namely

$$
\Delta_s = pr(Y = 1|Z = 1, s, x) - pr(Y = 1|Z = 0, s, x)
$$

$$
= pr(Y = 1|Z = 1, s, x, R = 1) - pr(Y = 1|Z = 0, s, x, R = 1)
$$

from (9) (11)

Importantly $\Delta_s$ in (11) does not depend on $x$. Let $\Delta_s^{apparent}$ denote the apparent treatment effect in stratum $s$ after collapsing over $x$, namely,

$$
\Delta_s^{apparent} = \sum_{z=0} \frac{pr(Y = 1, x|Z = 1, s, R = 1) - \sum_{z=0} \frac{pr(Y = 1, x|Z = 0, s, R = 1)}{\sum_{z=0} pr(Y = 1, x|Z = 1, s, R = 1)}
$$

$$
= \sum_{z=0} \frac{pr(Y = 1|Z = 1, s, R = 1)pr(x = 1|Z = 1, s, R = 1)}{\sum_{z=0} pr(Y = 1|Z = 1, s, R = 1)}
$$

$$
- \sum_{z=0} \frac{pr(Y = 1|Z = 0, s, R = 1)pr(x = 1|Z = 1, s, R = 1)}{\sum_{z=0} pr(Y = 1|Z = 1, s, R = 1)}
$$

(12)

To formalize the relationship between $\Delta_s^{apparent}$ and $\Delta_s$ let

$$
\alpha_{s} = pr(Y = 1|Z = 0, s, x, R = 1)
$$

(13)

$$
\psi_s = \alpha_{1s} - \alpha_{0s}
$$

(14)

$$
\phi_{2s} = pr(X = 1|z, s, R = 1).
$$

(15)

$$
\epsilon_{s} = \phi_{1s} - \phi_{0s}.
$$

(16)

Combining (11) and (13), we can write

$$
pr(Y = 1|Z = 1, s, x, R = 1) = \alpha_{s} + \Delta_s.
$$

(17)

Substituting (13)-(17) into (12) gives

$$
\Delta_s^{apparent} = (\alpha_{0s} + \Delta_s)(1 - \phi_{1s}) + (\alpha_{1s} + \Delta_s)\phi_{1s} - (\alpha_{0s} - \phi_{0s} + \alpha_{1s}\phi_{0s})
$$

$$
= \Delta_s + \psi_s\epsilon_s.
$$

(18)

For a graphical depiction based on the BK-plot [17,18], see Figure 1.

From (18) the bias from omitting $x$ in stratum $s$ is $\psi_s\epsilon_s$.

The first factor

$$
\psi_s = pr(Y = 1|Z = 0, s, X = 1, R = 1) - pr(Y = 1|Z = 0, s, X = 0, R = 1)
$$

(19)

is the effect of $X$ on subjects in the control group with observed outcomes. By virtue of the MAR assumption in (9), we could also write $\psi_s = pr(Y = 1|Z = 0, s, X = 1) - pr(Y = 1|Z = 0, s, X = 0)$, which is the effect of $X$ on all subjects in the control group. The second factor,

$$
\epsilon_s = pr(X = 1|Z = 1, s, Z = 1) - pr(X = 1|Z = 0, s, R = 1),
$$

(20)

ranges from -1 to 1 and measures the degree of confounding between $X$ and $Z$ among subjects with observed outcomes (i.e. $R = 1$). If $\epsilon_s = 0$, there is no confounding and
To bound the overall bias \( \psi \), one might argue that if \( x \) were a strong unobserved inherited gene, \( \psi \) would be close to 1. However because "eligible subjects had no history of colorectal cancer, surgical resection of adenomas, bowel resection, the polyposis syndrome, or inflammatory bowel disease" [14], it is unlikely that many subjects had an unobserved high-penetrance gene related to the recurrence of adenomas. We therefore believe that unobserved factors that might affect both adenoma recurrence and missingness could have an effect of similar magnitude as observed baseline covariates. Thus to obtain a plausible value for \( \psi \), we suggest estimating \( \psi \) based only on the fraction with observed outcomes in stratum \( s \) and \( w_s \) is the fraction of subjects in stratum \( s \).

If only 15% of the subjects are missing in each arm \( \psi_{(\text{max})} \) is less than .18. If we let \( \psi_{(\text{max})} \) be the anticipated maximum value of \( \psi \), then substituting (24) into (22) gives the anticipated maximum bias,

\[
\text{bias}_{(\text{max})} = \pm \psi_{(\text{max})} \sum_s \varepsilon_s \psi \frac{w_s}{\pi_{ss}}
\]

where the anticipated maximum bias under complete confounding, \( \psi_{(\text{max})} \), is specified by the investigator; the upper bound factor, \( \varepsilon_s \psi \), is based on the fraction with observed outcomes in stratum \( s \); and \( w_s \) is the fraction of subjects in stratum \( s \).

Thus the investigator need only specify \( \psi_{(\text{max})} \). One might argue that if \( x \) were a strong unobserved inherited gene, \( \psi_{(\text{max})} \) would be close to 1. However because "eligible subjects had no history of colorectal cancer, surgical resection of adenomas, bowel resection, the polyposis syndrome, or inflammatory bowel disease" [14], it is unlikely that many subjects had an unobserved high-penetrance gene related to the recurrence of adenomas. We therefore believe that unobserved factors that might affect both adenoma recurrence and missingness could have an effect of similar magnitude as observed baseline covariates. Thus to obtain a plausible value for \( \psi_{(\text{max})} \), we suggest estimating \( \psi_{(\text{max})} \) based only on the fraction with observed outcomes in stratum \( s \) and \( w_s \) is the fraction of subjects in stratum \( s \).
Results

We applied our approach to data from the PPT trial stratified by age and sex (Table 2). We first assumed MAR and applied (7) and (8) to estimate the difference in the probabilities of adenoma recurrence between the two groups.

We obtained $\hat{\Delta} = -0.003$ with standard error $0.022$. We define $\varepsilon_{(\text{max})s} = \max((1 - \pi_0s)/(1 - \pi_1s), (1 - \pi_1s)/(1 - \pi_0s))$, where $\pi_zs$ equals one minus the fraction missing in group $z$ and stratum $s$. The anticipated maximum bias is $\psi_{\text{max}} = \sum_s \varepsilon_{(\text{max})s} w_s = 0.10 \psi_{\text{max}}$, where $\psi_{\text{max}}$ is the anticipated bias if there were complete confounding of the unobserved covariate and treatment.

To compute the anticipated maximum bias (25) we first computed $\varepsilon_{(\text{max})s}$ using (24) and estimated $w_s$ from the observed fractions (Table 2). This gave $\sum_s \varepsilon_{(\text{max})s} w_s = 0.10$. We then specified $\psi_{\text{max}}$ the anticipated maximum bias under complete confounding. To obtain a plausible value for $\psi_{\text{max}}$ we estimated $\psi_s$ in (19) pretending either sex or age was the unobserved covariate $x$. This gave $\hat{\psi}_s = 0.23, 0.18, 0.18, 0.19$ for the four age categories when $x = \text{sex}$ and $0.07$ and $0.09$ for the two sex categories when $x = \text{age}$. Treating the largest $\psi_s$ as a realistic lower bound for $\psi_{\text{max}}$ we specified a slightly larger value, $\psi_{\text{max}} = 0.25$, so that the anticipated maximum bias is $\text{bias}_{\text{max}} = 0.25 \times 0.10 = 0.025$. The MAR confidence interval is shifted to the right or left by the anticipated maximum bias (Figure 2).

For purpose of comparison, we also computed estimates and confidence intervals under a worst (best) case imputation [9,19], where missing outcome data in each stratum were imputed as no recurrence (recurrence) in controls and recurrence (no recurrence) in the intervention group. (These stratum-specific estimates were combined over strata using weights inversely proportional to the stratum-specific variances.) In the worst and best case imputations the confidence intervals did not overlap zero (Figure 2).

Our sensitivity analysis showed that the worst and best case imputations were too extreme. Because the absolute value of the anticipated maximum bias, $0.025$, is smaller than $1.96 \times \text{se}(\hat{\Delta}) = 0.043$, the bias-adjusted confidence intervals overlap zero. Thus the anticipated maximum bias of $\pm 0.025$ did not change our conclusion of little evidence of an effect of treatment on adenoma recurrence. However it did increase our uncertainty, as the more extreme lower and upper bounds indicated that the true effect of treat-

Table 2: Results of Polyp Prevention Trial

<table>
<thead>
<tr>
<th>stratum s</th>
<th>adenoma no</th>
<th>adenoma yes</th>
<th>adenoma missing</th>
<th>difference in observed rates of recurrence $d_s$</th>
<th>weight $w_s$</th>
<th>bias factor $\varepsilon_{(\text{max})s}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>control</td>
<td>study</td>
<td>control</td>
<td>study</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30–49</td>
<td>573</td>
<td>374</td>
<td>94 (9%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>578</td>
<td>380</td>
<td>76 (7%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40–59</td>
<td>58</td>
<td>12</td>
<td>3 (4%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>76</td>
<td>7 (4%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60–69</td>
<td>94</td>
<td>76</td>
<td>9 (5%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70–79</td>
<td>70</td>
<td>71</td>
<td>29 (17%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>women</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>control</td>
<td>study</td>
<td>control</td>
<td>study</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30–49</td>
<td>54</td>
<td>11</td>
<td>3 (4%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>47</td>
<td>12</td>
<td>4 (6%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40–59</td>
<td>69</td>
<td>24</td>
<td>4 (4%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>69</td>
<td>27</td>
<td>4 (4%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60–69</td>
<td>77</td>
<td>31</td>
<td>13 (11%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70–79</td>
<td>54</td>
<td>29</td>
<td>11 (12%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The overall estimate of the difference in probabilities of recurrence between study and control groups is $\hat{\Delta} = \sum_s d_s w_s = -0.003$ with a standard error $0.022$. We define $\varepsilon_{(\text{max})s} = \max((1 - \pi_0s)/(1 - \pi_1s), (1 - \pi_1s)/(1 - \pi_0s))$, where $\pi_zs$ equals one minus the fraction missing in group $z$ and stratum $s$. The anticipated maximum bias is $\psi_{\text{max}} = \sum_s \varepsilon_{(\text{max})s} w_s = 0.10 \psi_{\text{max}}$, where $\psi_{\text{max}}$ is the anticipated bias if there were complete confounding of the unobserved covariate and treatment.
We also agree with Shih [1] that one should collect information on the cause of missingness. In particular we recommend reporting whether any of the missing outcomes were definitely MAR, for example, due to random technical problems, to accidents, or to leaving the study for reasons completely unrelated to the investigation. Suppose that outcome was definitely MAR in a proportion \( \pi_{zs} \) of subjects. Then it is more informative to write \( \nu_{zs} \) as \( pr(R = 1, \text{MAR}|z, s) + \nu_{zs} \). Because \( \nu_{zs} \) contains no information about the effect of \( X \) on missingness, one can replace \( \pi_{zs} \) by \( \pi_{zs} \cdot \nu_{zs} \), which reduces \( \epsilon_{(\max)} \) and hence reduces the anticipated maximum bias.

Although we applied our methodology to a cross-classification of categorical covariates, it could also be applied to continuous covariates or a univariate combination of covariates in a manner analogous to a propensity score [22]. Let \( u \) denote a vector of covariates and \( \epsilon_z = pr(R = 1|z, u) \). Following the derivation of propensity scores [22], we can write, \( pr(R = 1|z, \epsilon_z) = E(E(R|z, \epsilon_z)|z)|z, \epsilon_z = E(\epsilon_z|z, \epsilon_z) = \epsilon_z \). Therefore \( pr(R = 1|z, u) = pr(R = 1|z, \epsilon_z) \), and thus \( \epsilon_z \) contains the same information for the probability of being observed as \( u \). This calculation justifies using \( \epsilon_z \) to summarize the covariates predicting missingness. To form five strata for randomized group \( z \), we would compute \( \epsilon_z \) for each subject in group \( z \) and then divide the distribution of \( \epsilon_z \) into quintiles.

**Conclusion**

The bias due to an unobserved binary covariate could arise when the probability of missingness depends on both treatment and outcome. Computation of the bias is easy because it equals the maximum anticipated bias under complete confounding multiplied by an upper bound factor. The maximum anticipated bias might require some expert input but some lower bound values can be obtained using observed baseline covariate. The upper bound factor is easily computed from the fraction missing in each group. The methodology is particularly useful in the common situation when no more than 15% of the subjects (in excess of those definitely MAR) have missing outcomes, so that the upper bound factor in the bias is less than .18.

**Contributions**

SGB devised the basic model with the unobserved covariate, worked out the unconstrained maximization, and wrote the initial draft of the manuscript. LSF worked out the constrained maximization and provided substantive improvements to the manuscript.

---

**Figure 2**

Comparison of missing data adjustments for Polyp Prevention Trial. The graph plots the estimated differences in the probability of adenoma recurrence between the intervention and control groups and the 95% confidence intervals. MAR is missing at random within strata. MAR ± bias shifts the MAR confidence interval based on the anticipated maximum bias. Worst and best case imputes missing data to the randomization group that would give the largest positive and negative effect, respectively.
Additional file 1

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-3-8-s1.pdf]

References

1. Shih WJ Problems in dealing with missing data and informative censoring in clinical trials. *Current Controlled Trials in Cardiovascular Medicine* 2002, 3(4)
17. Wainer H *The BK-Plot: Making Simpson’s paradox clear to the masses*. Chance 15:60-62

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/3/8/prepub

Publish with *BioMed Central* and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:
- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp