HHS Public Access

Author manuscript

J Biopharm Stat. Author manuscript; available in PMC 2019 January 01.

Published in final edited form as:

J Biopharm Stat. 2018 ; 28(4): 633–644. doi:10.1080/10543406.2017.1372773.

Differential losses to follow-up that are outcome-dependent can vitiate a clinical trial: Simulation results

Richard F. Potthoff

Cancer Statistical Center, Duke University Medical Center, 2424 Erwin Road, Suite 802, Durham, NC 27705, U.S.A. potthoff@duke.edu ; phone: (919)681-1038

Abstract

Loss to follow-up (LTFU) in clinical trials represents a potential threat to their soundness that may not be adequately recognized. We consider a log-rank test in a trial with two arms, experimental and control, and with a single unfavorable binary endpoint such as death. Commonly, one applies censoring to patients with LTFU. That approach is valid if LTFU is independent of outcome, but can lead to bias otherwise. Unfortunately, there is no statistical test for independence, so the legitimacy of the approach rests on unverifiable assumptions. For two cases, we evaluate the impact of the approach based on simulations that use reasonable models for outcome dependent LTFU. In each case, LTFU in one arm disproportionately suppresses recognition of relatively early deaths or other outcomes, thus producing bias favoring that arm. The first case has extra LTFU in the experimental arm and the treatment has no benefit. The second case has extra LTFU in the control arm and the treatment is effective. The simulation results show severe inflation of Type I error in the first case and major loss of power in the second case. Remedies for LTFU are scarce but include avoiding it in the first place where possible.

Keywords

Bias; Clinical trials; Log-rank test; Loss to follow-up; Simulations

1 Introduction

This paper uses simulations to examine the effects of outcome-dependent loss to follow up (LTFU) whose extent or nature differs between the two arms of a randomized clinical trial. The setting posits that one uses a log-rank test, with LTFU handled through censoring, to assess whether an experimental arm is better than a control arm with respect to a single (i.e., non longitudinal) unfavorable binary endpoint or outcome, such as death. LTFU occurs for a patient if one cannot determine when or whether the patient's outcome took place after loss of contact.

If LTFU is independent of outcome but its extent or distribution differs between arms, then the log-rank test still has the correct Type I error. Likewise, if LTFU is outcome-dependent but the joint distribution of LTFU and the outcome is the same in both arms under the null hypothesis, then again the log-rank test is valid. But if LTFU is outcome-dependent and also may have a different (null) joint distribution with the endpoint in the two arms, then the log-

rank test can produce faulty results. It is at this last situation that the present paper is directed.

To enable slightly simpler language, we often refer henceforth to "death" rather than "outcome." Of course, everything that follows applies to any binary outcome, not just death.

LTFU poses challenges. Per the results of Tsiatis (1975) on competing risks, if one is given any pair of ("crude") distribution functions, consisting of the distribution of death where it precedes LTFU and the distribution of LTFU where it precedes death, then there exists a joint distribution consistent with those two in which ("potential") death and LTFU times are independent. The implication is that (for either arm) one can never determine from clinical trial data whether death and LTFU are independent, because, for any (crude) data, one can always find an independent joint distribution that conforms to the data.

Since it is thus not safe to assume independence of death and LTFU (where LTFU does exist), the use of the log-rank test (or any test that presumes independence) is problematic. It can be more so the greater the LTFU difference between arms, and the greater the LTFU in total.

The purpose of this paper is rather modest: to assess the possible impact of LTFU on Type I error and on power, rather than to propose remedies. But the troubles that our simulations expose are disturbing. Although there are both empirical and theoretical indications that one should sometimes expect distortions from LTFU with respect to both Type I error and power loss (as one may gather from points covered later in this Introduction), the possible severity of the consequences of the biases and distortions for some clinical trials may not be well appreciated. This paper provides quantification of those consequences, an undertaking that may be novel.

Outcome-dependent LTFU that differs between the two arms can take varied forms. One arm (either treatment or control) but not the other may involve undesirable side effects that lead to greater LTFU among patients who are sicker, more vulnerable to the side effects, and likely to die sooner; or possibly LTFU is greater among the healthier patients if the side effects bother them more. A treatment or control alternative that succeeds in causing some patients but not others to feel better may entail greater LTFU for the latter patients if they see no benefit in continuing. Or, conversely, an alternative that is demanding but somewhat successful may result in greater LTFU among patients who feel better and are destined to have better survival, if they think it is no longer necessary to continue. In all these situations, lack of information about what happened to the patients with LTFU can lead to bias.

Different previous authors, though not many, have written in some depth about LTFU. Akl et al. (2012) examined LTFU in 235 reports (published in 2005-2007 in five leading medical journals) of randomized controlled clinical trials with binary outcome and statistical significance at the 0.05 level. Analyzing those reports, they found a median LTFU of 6% of participants (interquartile range, 2% to 14%) in studies reporting LTFU; widespread lack of adequate attention to details and implications regarding LTFU; difficulties in evaluating (through different means that they considered) the impact of LTFU; and loss of statistical

significance in up to a third of trials under plausible assumptions about the outcomes for patients with LTFU.

Other articles that call attention to the bias that can result from LTFU in clinical trials include Ranganathan and Pramesh (2012) and Walsh et al. (2015), in a general context; Clark et al. (2003), in the context of cancer trials; and Stinner and Tennent (2012), in the context of surgical trials. The last article noted a sharp difference in LTFU occurrence between the surgical and nonsurgical arms of a trial (6% versus 25%, respectively). In an observational study pertaining to hip replacement, Murray et al. (1997) found that patients who had LTFU were in worse condition (before LTFU) than patients who were otherwise comparable.

For specific observational studies, Geng et al. (2008, 2010), Wu et al. (2008a), te Riele et al. (2010), and Schomaker et al. (2012) each reported empirical evaluations of possible LTFU-related bias in different contexts, through efforts to obtain outcome information about patients with LTFU. All these studies except Wu et al. (2008a) found that patients with LTFU had worse outcomes than other patients. Use of outcome information that is tracked down about patients with LTFU is dealt with also by Frangakis and Rubin (2001), in a work that is more methodological than empirical.

In principle at least, one can try to compensate for bias resulting from LTFU by predicting the unknown outcomes for the patients who have LTFU and then using those predictions in the survival analysis, although such a technique can be heavily dependent on modeling assumptions. A recent article by Liu (2016) explores this type of approach.

Somewhat related to LTFU are the works of Snapinn et al. (2004) and Jiang et al. (2004),which focus not on LTFU but rather on noncompliance (patient's discontinuation of study drug) and its effect on power and bias and on required sample size. Scharfstein et al. (2014) focus on informative LTFU in longitudinal studies rather than those with single-event endpoints, developing and examining methodology to analyze sensitivity to model assumptions. Shih (2002) discusses various issues regarding LTFU, mainly in the context of cardiology clinical trials, and notes that LTFU often comes about after adverse events.

Ways to measure and report LTFU are the subject of several works. They include Clark et al. (2002), Siskind (2002), and Wu et al. (2008b).

Obviously, the best way to deal with LTFU is to prevent it in the first place, to the extent that that can be done. Possible strategies to reduce clinical-trial LTFU in different contexts include those considered by Sprague et al. (2003), Cleland et al. (2004), and McCarthy et al. (2016).

Whether in regard to clinical trials with a binary outcome, those with longitudinal data, or observational studies, LTFU essentially involves a missing survival outcome, so missing-data theory is applicable. Results just cited suggest that generally the missing data resulting from LTFU are not missing completely at random (MCAR) or even missing at random (MAR), since often the probability of missingness (of LTFU) appears to depend on (the sometimes unobserved) outcome. With data missing not at random (MNAR), standard tools

of survival analysis such as the log-rank test, Cox regression, and the Kaplan-Meier estimator are generally unsuitable to use [see also, e.g., Little (1993, 1995); Frangakis and Rubin (2001); Little and Rubin (2002); Schomaker et al. (2014)]. Thus the consequences of using the log-rank test in clinical trials that have LTFU are important to assess.

Section 2 describes our model and Section 3 the simulations. Simulation results are in Section 4. Section 5 concludes.

2 Model

Various models might be formulated to investigate how outcome-dependent differential loss to follow-up can distort Type I error or power in a two-arm randomized clinical trial that has a single binary endpoint (i.e., a single dichotomous time-to-event endpoint). Our model seems reasonable and is also mathematically convenient.

For simplicity, we consider a situation where LTFU occurs in only one arm. Basic conclusions may differ little whether (e.g.) LTFU occurrences in the two arms are 7% and 0% or 12% and 5%.

We use K to denote the time from the randomization of the first patient until the data cutoff. The time from the first patient randomization to the finish of all patient randomizations is denoted by $F (F < K)$.

The model has three basic elements. First, t, a patient's time of death (or other outcome) counted from the time of that patient's randomization, is assumed to be measured exactly (rather than somewhere within a specified interval) and to follow a Weibull distribution,

$$f(t) = a^b b t^{b-1} e^{-(at)^b}, \quad 0 < t < \infty . \quad (1)$$

The median of the distribution (1) is $M = (\log 2)^{1/b}/a$. Our simulations vary a and b. Let q denote the time from the patient's randomization until the data cutoff at the end of the trial (which means that $q \leq K$). If $t > q$, then t will not be known to the investigator.

Second, for a patient who is in the arm where LTFU can occur and whose death is at time t, we specify that LTFU will occur (at some time before t) with probability p_{*t}, and not occur with probability $(1 - p_{*t})$, where

$$p_{*t} = p_0 e^{-ut} \text{ with } u = \frac{1}{K}\log\frac{p_0}{p_K} \qquad \text{for } 0 < p_K \leq p_0 \leq 1$$

or

$$p_{*t} = 1 - (1 - p_0)e^{-ut} \text{ with } u = \frac{1}{K}\log\frac{1 - p_0}{1 - p_K} \qquad \text{for } 1 > p_K \geq p_0 \geq 0 . \quad (2)$$

In both of equations (2), $p_{*0} = p_0$ and $p_{*K} = p_K$, where p_0 and p_K (the respective LTFU probabilities for patients who die at 0 and K time units after *their* randomizations) are parameters to be set to different values for the simulations. The formulas (2) simply use K as a convenient anchoring point but instead could have used some value other than K for that purpose. Note that t can be $< q$, between q and K, or $> K$. In either of the last two cases, death occurs after data cutoff (because $t > q$) but LTFU (if it takes place) can occur either before or after data cutoff.

Finally, let y denote the time from the patient's randomization until LTFU occurs, if the patient is in the arm that can have LTFU and is one of those who do realize LTFU. For such a patient, LTFU has to occur before death, so we use $g(y|t)$, $0 \le y < t$, to denote the conditional density of y given t for the patient. Of course, $g(y|t)$ has to satisfy $\int_0^t g(y \mid t)\, dy = 1$. For the simulations, we apply two choices for $g(y|t)$, one uniform (rectangular),

$$g(y \mid t) = \frac{1}{t},\ 0 \le y < t, \quad (3)$$

and the other triangular,

$$g(y \mid t) = \frac{2y}{t^2},\ 0 \le y < t\,. \quad (4)$$

For given t, LTFU tends to occur closer to death in (4) than in (3). Observe that the value drawn for y will be merely hypothetical if $q < y < t$, but not if $y < t < q$ or $y < q < t$. In either of the last two cases, the patient is censored at y, the time of LTFU. In the first case, the patient is censored at q.

The model of (2), (3), and (4) obviously has outcome-dependent LTFU. It was designed with the intent that patients with relatively early deaths or other outcomes would be overrepresented among those who realized LTFU, thus leading to unduly favorable evaluation of the related arm. LTFU prevents the posting of the death of a patient with $y < t < q$.

The overrepresentation of patients who die earlier among those with LTFU, for the above model versus one with LTFU independent of survival, is perhaps best understood in a rough way if one considers patients with very small t. Under the independence model, such patients have only a tiny chance of realizing LTFU ($\Pr\{y < t|t\}$), far lower than any value of p_{*t} we will have under our model above.

Simulations of clinical trials involving outcome-dependent LTFU appear to be uncommon, although Schomaker et al. (2014, section 4) do report one that uses a model that they devised. In our model, the use of the Weibull distribution (1) for time until death, t, seems standard. But the rest of our model, which prescribes the conditional distribution of y (time

until LTFU) given t, through the use of (2) and (3)-(4), allows for considerable added versatility and appears to be novel.

Specifying the probability of LTFU (p_{*t}) to be a function of t, as in (2), provides broad flexibility. Further flexibility is provided through $g(y|t)$, although $g(y|t)$ is restricted to the interval from 0 to t, as is $E(y|t)$, the mean time until LTFU. Both (3) and (4) prescribe beta distributions for $g(y|t)$, with $E(y \mid t) = \int_0^t y\, g(y \mid t)\, dy = \frac{1}{2}t$ for (3) and $= \frac{2}{3}t$ for (4). Obviously, other beta distributions for $g(y|t)$ could be used, although ones with $E(y|t)$ below $\frac{1}{2}t$ would entail relatively early LTFU and might thus be deemed unrealistic.

3 Simulations

The simulations were all carried out with SAS, including SAS/IML. Broadly, they deal with two cases. In Case 1, LTFU occurs in the experimental arm (but not the control arm) and $f(t)$ of (1) is the same in both arms (i.e., the treatment has no benefit). The simulations gauge the extent to which the Type I error (the chance of finding the treatment beneficial when it is not) is too high. Apparently the LTFU gives the treatment a bias in its favor because deaths of patients who die earlier are less likely to be recognized.

Case 2 has LTFU in the control arm only and has $f(t)$ different in the two arms so that the treatment is beneficial. The simulations evaluate the reduction in power (the chance of finding the treatment beneficial) below what was intended. The LTFU produces a bias that helps the control arm, thereby impairing the power by lessening the difference between the two arms.

Each of the two cases has simulations for 30 conditions. These 30 conditions consist of all combinations of the following: three pairs of Weibull distributions $f(t)$ of (1) (one distribution for the experimental arm, one for the control arm); five sets of (p_0, p_K) for use in (2) for the arm that has LTFU; and either the uniform distribution (3) or the triangular distribution (4) for $g(y|t)$ for the arm that has LTFU.

For Case 1, all three pairs of Weibull distributions have medians $M_E = M_C = 3$ (units of time), where the subscripts E and C refer to the experimental and control arms, respectively. The values of the Weibull shape and scale parameters for the three pairs (same for E and C) are

$$b_E = b_C = 0.8,\ a_E = a_C = \tfrac{1}{3}(\log 2)^{1.25};$$
$$b_E = b_C = 1,\ a_E = a_C = \tfrac{1}{3}(\log 2);$$
$$b_E = b_C = 1.25,\ a_E = a_C = \tfrac{1}{3}(\log 2)^{0.8}.$$

The three pairs of Weibull distributions of (1) for Case 2 each have the median pairs $M_E = 4.2$, $M_C = 3$, with shape and scale parameters of

$$b_E = b_C = 0.8,\ a_E = \frac{1}{4.2}(\log 2)^{1.25}, a_C = \frac{1}{3}(\log 2)^{1.25};$$
$$b_E = b_C = 1,\ a_E = \frac{1}{4.2}(\log 2), a_C = \frac{1}{3}(\log 2);$$
$$b_E = b_C = 1.25,\ a_E = \frac{1}{4.2}(\log 2)^{0.8}, a_C = \frac{1}{3}(\log 2)^{0.8}.$$

For either arm and for either Case 1 or Case 2, the value of t for a patient comes from the applicable Weibull distribution using the formula $t = (-\log x)^{1/b}/a$, where x is drawn randomly from the uniform distribution on [0, 1].

For the experimental arm for Case 1 and the control arm for Case 2, the five sets of (p_0, p_K) used in calculating p_{*t} of (2) are

$$(p_0, p_K) = (0, 0), (0.05, 0.05), (0.05, 0.15), (0.1, 0.1), (0.15, 0.05).$$

LTFU occurs for a patient with time of death t if $x < p_{*t}$, where x is from the uniform distribution on [0, 1]. Of course, for the control arm for Case 1 and the experimental arm for Case 2, $p_{*t} = 0$ for all t and all five sets.

For a patient with time of death t who realizes LTFU (in either the experimental arm for Case 1 or the control arm for Case 2), the time of the LTFU is determined as $y = tx^{1/w}$, where $w = 1$ for the rectangular $g(y|t)$ distribution (3) and $w = 2$ for the triangular distribution (4), with x uniform [0, 1]. As noted before, y plays no role if (and only if) the value drawn for it exceeds q.

Certain elements, representing reasonable specifications, are common to all the simulations. Each arm has 300 patients. They are randomized uniformly over a period of $F = 5$ time units. Data cutoff occurs after $K = 7$ time units. Thus q is uniform [2, 7]. The median survival times (shown above) provide acceptable numbers of events.

The median $M_E = 4.2$ for Case 2 (with $M_C = 3$) was chosen so that power in the absence of LTFU would take a reasonable value. For 300 patients in each arm, accrual time of 5 and follow-up time of 2 (i.e., $F = 5$ and $K = 7$), no LTFU, and exponential survival with median survival times of 4.2 and 3, PROC POWER of SAS shows that the power of the log-rank test is 87.4% and 70.2% for respective one-sided significance levels of 0.025 and 0.005.

For each case and each of the 30 conditions, the simulations ran 10,000 trials, thus yielding a total of 2 × 30 × 10,000 = 600,000 simulated trials involving 2 × 300 × 600,000 = 360,000,000 patients. The log-rank test was run for each trial, with a patient treated as censored at the time of LTFU if LTFU occurred (with $y < q$) or as censored at the time of data cutoff if the patient was still alive then (but had not realized LTFU). For each of the 600,000 trials, we recorded the value of z that was obtained as the square root of the log-rank χ^2-value from PROC LIFETEST of SAS, with a minus sign attached if the control arm was the one with the worse result.

Each simulation result in our tables below is shown as a percentage, and comes from Bernoulli observations that are mostly based on whether or not $z < -1.96$ or $z < -2.576$ and are generated from (at least) 10,000 simulated clinical trials. The standard error for each one is thus no greater than $\sqrt{0.5 \times 0.5 / 10,000} = 0.005 = 0.5\,\%$, and is far lower for many of the percentages in the tables.

4 Results of the simulations

Table 1 shows the simulation results for Case 1, for which bias that favors the treatment and thus inflates the Type I error is the consequence of the LTFU in the experimental arm. The treatment has no benefit in Case 1. The top row in each of the three groups of five rows in Table 1 shows that rejection frequency conforms with what it is supposed to be (0.025 or 0.005) for the null case with no LTFU. Generally, for all three values of b (the Weibull shape parameter), rejection frequencies climb as LTFU becomes more influential, with a bit greater impact for $b = 0.8$ than $b = 1$ and for $b = 1$ than $b = 1.25$. The impact for triangular $g(y|t)$ exceeds that for uniform $g(y|t)$, apparently because LTFU comes closer to death with the triangular distribution and can thus lead to (artificially) better survival results recorded for affected patients. The greatest distortion occurs with $b = 0.8$, $(p_0, p_K) = (0.15, 0.05)$, and triangular $g(y|t)$, for which the null hypothesis of no benefit for the treatment is rejected almost 17% of the time where it should be 2.5% and almost 6% of the time where it should be 0.5%.

Case 2, for which the treatment is beneficial and whose simulation results appear in Table 2, involves power diminution stemming from control-arm LTFU that causes bias in favor of the control. In the middle group of five rows in the table (exponential distribution), the top row (no LTFU) shows simulated power close to the values of 87.4% and 70.2% found from PROC POWER of SAS. For all three values of b, power drops sharply as LTFU increases. As in Table 1, the impact is heavier with triangular than with uniform $g(y|t)$. The three values of b differ considerably with respect to power even when no LTFU exists, with the top row in each group of five rows in Table 2 showing $b = 1.25$ with greater power and $b = 0.8$ with lower power than $b = 1$; not unexpectedly, that relationship endures for the remaining four rows.

Although the power loss resulting from LTFU in Case 2 is substantial, it is at least somewhat inflated because the effective sample-size reduction stemming from LTFU would, by itself, cause some reduction in power. One thus has to evaluate how the results Table 2 in are affected by the sample-size factor. We do so by examining how power would be affected if LTFU is *independent of* outcome. (For Case 1, there is no concern like that for Case 2, since no similar consideration of sample-size effect for Case 1 would diminish the distortion shown by Table 1.)

As a first step, we show in Table 3, for each of the 30 conditions and for the LTFU arms of Cases 1 and 2, the percentage of simulation patients who had LTFU and died before data cutoff ($y < t < q$, "Lost and died"); who had LTFU (that necessarily occurred before data cutoff) and survived beyond data cutoff ($y < q < t$, "Lost and survived"); and who had LTFU in total ($y < t$). [The "Lost and survived" columns in Table 3 come from the combined results

of Cases 1 and 2, and are thus each based on 20,000 trials. The LTFU arm has the same median survival time of $M = 3$ in the two cases, and the circumstances also are otherwise the same, so the results can be fused because the simulation estimates the same percentage for both cases. For the "Lost and died" columns, the percentages are obtained by combining not only for the two cases but also for the two w-values, thus producing results that are based on 40,000 trials and are the same in the two "Lost and died" columns in Table 3. This combining of results from $w = 1$ and $w = 2$ can be done for "Lost and died" ($y < t < q$) but not for "Lost and survived" ($y < q < t$) because, for the former, different $g(y|t)$ distributions have no effect since $t < q$, whereas, for the latter, $g(y|t)$ affects the frequency of $y < q < t$ (versus $q < y < t$).]

To try to gauge the impact of the LTFU sample-size effect on inflating the power loss in Table 2, we focus on the middle five rows of Table 3, which are for $b = 1$ (exponential distribution). The aim is to find joint distributions of survival and LTFU that have LTFU independent of survival and that match the results in these rows as well as possible. One can then use PROC POWER of SAS to evaluate power for those distributions. Such computation with PROC POWER can be done only when $b = 1$. But the top five and bottom five rows of Table 3 are still of some relevance, because the close similarity of their values to those of the middle five rows and of each other suggests that the evaluation of the sample-size effect may not differ much for $b = 0.8$ or 1.25 versus $b = 1$.

To try to match the values in the middle five rows of Table 3 with those from joint distributions with survival and LTFU independent, we first need to calculate the probabilities of $y < t < q$ and $y < q < t$ when t and y are independent, both are exponential, and q is uniform [V, K], where $V = K - F$. We specify the distributions

$$f(t) = ae^{-at}, \quad 0 < t < \infty; \quad h(y) = se^{-sy}, \quad 0 < y < \infty; \quad \varphi(t, y) = f(t)h(y). \quad (5)$$

(The outcome $y > t$ is considered to be effectively equivalent to $y = t$, i.e., no LTFU. The outcome $q < y < t$ does not entail LTFU because no LTFU occurs before data cutoff.) From (5) one obtains

$$\Pr\{y < t < q \mid q\} = \int_0^q \left[\int_0^t \varphi(t, y)dy\right]dt = \frac{s}{a+s} - e^{-aq} + \frac{a}{a+s}e^{-(a+s)q},$$

which leads to

$$\begin{aligned} \Pr\{y < t < q\} &= \frac{1}{F}\int_V^K \Pr\{y < t < q \mid q\}dq \\ &= \frac{s}{a+s} - \frac{1}{Fa}(e^{-Va} - e^{-Ka}) + \frac{a}{F(a+s)^2}\left[e^{-V(a+s)} - e^{-K(a+s)}\right]. \end{aligned} \quad (6)$$

Then also

$$\Pr\{y < q < t \mid q\} = \left[\int_q^{\infty} f(t)dt\right]\left[\int_0^q h(y)dy\right] = e^{-aq}(1 - e^{-sq}),$$

from which

$$\begin{aligned}\Pr\{y < q < t\} &= \frac{1}{F}\int_V^K \Pr\{y < q < t \mid q\}dq \\ &= \frac{1}{Fa}\ (e^{-Va} - e^{-Ka}) - \frac{1}{F(a+s)}\left[e^{-V(a+s)} - e^{-K(a+s)}\right]\end{aligned} \tag{7}$$

follows.

Ideally, one would like to choose values of s so as to get a close match both between (6) and "Lost and died" in the middle rows of Table 3 and between (7) and "Lost and survived" in those rows. As it turns out, though, that is not feasible: Regardless of the extent of LTFU, the ratio of (7) to (6) is far higher than the ratio of "Lost and survived" to "Lost and died" in the relevant part of Table 3. We therefore take a conservative approach and, with $F = 5$, $V = 2$, $K = 7$, and $a = (\log 2)/M_C = (\log 2)/3$, find s so that (6) matches exactly the values in the middle rows of the two "Lost and died" columns in Table 3. The resulting values of s are in the first column of Table 4. The "Lost and die" values in that table, calculated from (6) using those values of s, agree exactly with the corresponding values in Table 3. The "Lost and survive" values in, though, are obtained from (7) and are sharply higher than the corresponding values in Table 3. The distributions reflected in the left part of Table 4 thus have greater overall LTFU than the corresponding distributions in the simulations. They thereby provide a conservative basis for assessing (through overstatement) the part of the power loss that is attributable to the sample-size factor rather than to LTFU separate from the sample-size factor.

The last six columns of Table 4 show the power comparisons. The four columns appearing under "$w = 1$" and "$w = 2$" are simply the simulation results for power for $b = 1$ copied from Table 2. The other two columns are the outputs from PROC POWER of SAS for the log-rank test using one-tailed significance levels of 0.025 and 0.005 with exponential LTFU hazards for the control arm equal to each value of s [and with 300 patients in each arm, exponential survival medians of 4.2 (experimental) and 3 (control), accrual time of 5, and follow-up time of 2]. The model that PROC POWER uses to calculate power assumes that LTFU is independent of survival.

In the last six columns of Table 4, the power decreases in the two PROC POWER columns are far less than in the other four columns. Thus the conclusion that LTFU brings about a substantial loss of power in Case 2 is left largely unaltered after considering the effects of the sample-size factor. That is, only a small part of the power loss shown in Table 2 is due to the sample-size effect rather than to the bias stemming from outcome-dependent LTFU.

5 Summary and concluding remarks

For different conditions, our simulations have demonstrated troubling effects of outcome-dependent LTFU whose degree differs between two arms of a clinical trial. Our simulation model is set up so that the deaths or other outcomes suppressed by LTFU tend to be earlier ones.

For Case 1, our simulations tested the effects on Type I error of extra LTFU in the experimental arm when the treatment provides no benefit relative to the control. For Case 2, our work tested the effects on power of adding the LTFU in the control arm when the treatment does provide benefit. Simulation results for Case 1 show seriously inflated Type I error, thus leading to inordinate probability that the treatment will be determined to be effective when it is not. Results for Case 2 show major impairment of power, thereupon entailing a sharp reduction in anticipated ability to detect the effectiveness of a treatment that really does yield an advantage.

Analyses of clinical trials may often assume that LTFU is independent of outcome and can therefore be legitimately handled with censoring in a log-rank test. But not only may that assumption be invalid; no test for the existence of outcome-dependent LTFU is even possible. Thus outcome-dependent LTFU whose extent or nature differs between arms, which this paper focuses upon, is especially pernicious. With lower LTFU in total, or lower LTFU difference between the two arms, problems may be less but do not disappear. One generally cannot find out whether LTFU is causing bias.

Remedies for LTFU are not easily found. Some forms of statistical adjustment might be attempted but would be based on models, probably complex ones, whose foundations could be hard to verify. Greater attention to LTFU details in journal articles reporting clinical trials would raise awareness, aid interpretation, and be highly desirable, but would not itself provide resolution.

The remaining remedy consists of increased efforts to avoid LTFU in the first place. Innovative strategies to improve retention could potentially be of significance. For trials with a mortality endpoint, intensive efforts to find the vital status of LTFU patients who are otherwise unaccounted for would be an obvious way to attack the problems of LTFU. For trials whose endpoint is not mortality, though, the challenge is far greater.

Some of the references cited in the Introduction (Section 1 above) provide further information related to some of the possibilities just mentioned. Unfortunately, finding a good overall approach for tackling potentially severe damage from LTFU cannot be easy.

Thus it is difficult to provide guidance, to those engaged in the design and analysis of clinical trials, as to how to avoid unfavorable impact from LTFU. Recognizing the problem in the design stage, rather than waiting until data analysis, is obviously helpful. Running a pilot mini-trial beforehand might provide a better gauge as to the seriousness of the consequences of LTFU and also uncover ideas for minimizing LTFU. Anything that avoids LTFU to begin with (or, for a mortality endpoint, finds the vital status for more patients with LTFU) is beneficial.

Acknowledgments

Funding: This work was partially supported by Grant CA142538 from the National Cancer Institute. The author thanks two referees and an associate editor for their helpful comments and suggestions.

References

Akl EA, Briel M, You JJ, Sun X, Johnston BC, Busse JW, Mulla S, Lamontagne F, Bassler D, Vera C, Alshurafa M, Katsios CM, Zhou Q, Cukierman-Yaffe T, Gangji A, Mills EJ, Walter SD, Cook DJ, Schünemann HJ, Altman DG, Guyatt GH (2012). Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review. BMJ 344:e2809. [PubMed: 22611167]

Clark TG, Altman DG, De Stavolumea BL (2002). Quantification of the completeness of follow-up. The Lancet 359:1309–1310.

Clark TG, Bradburn MJ, Love SB, Altman DG (2003). Survival analysis Part I: basic concepts and first analyses. British Journal of Cancer 89:232–238. [PubMed: 12865907]

Cleland JGF, Torp-Pedersen C, Coletta AP, Lammiman MJ (2004). A method to reduce loss to follow-up in clinical trials: informed, withdrawal of consent. The European Journal of Heart Failure 6:1–2. [PubMed: 15012911]

Frangakis CE, Rubin DB (2001). Addressing an idiosyncrasy in estimating survival curves using double sampling in the presence of self-selected right censoring. Biometrics 57:333–342. [PubMed: 11414553]

Geng EH, Emenyonu N, Bwana MB, Glidden DV, Martin JN (2008). Sampling-based approach to determining outcomes of patients lost to follow-up in antiretroviral therapy scale-up programs in Africa. JAMA 300:506–507. [PubMed: 18677022]

Geng EH, Glidden DV, Emenyonu N, Musinguzi N, Bwana MB, Neilands TB, Muyindike W, Yiannoutsos CT, Deeks SG, Bangsberg DR, Martin JN (2010). Tracking a sample of patients lost to follow-up has a major impact on understanding determinants of survival in HIV-infected patients on antiretroviral therapy in Africa. Tropical Medicine and International Health 15(June, Supplement s1):63–69.

Jiang Q, Snapinn S, Iglewicz B (2004). Calculation of sample size in survival trials: the impact of informative noncompliance. Biometrics 60:800–806. [PubMed: 15339304]

Little RJA (1993). Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association 88:125–134.

Little RJA (1995). Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association 90:1112–1121.

Little RJA, Rubin DB (2002). Statistical Analysis with Missing Data. Hoboken, NJ: Wiley.

Liu Y (2017). Sensitivity analysis for informative censoring in survival data: a trial example. Journal of Biopharmaceutical Statistics 27:595–610. [PubMed: 27010683]

McCarthy O, French RS, Roberts I, Free C (2016). Simple steps to develop trial follow-up procedures. Trials 17:29. [PubMed: 26767505]

Murray DW, Britton AR, Bulstrode CJK (1997). Loss to follow-up matters. The Journal of Bone and Joint Surgery 79-B:254–257.

Ranganathan P, Pramesh CS (2012). Censoring in survival analysis: potential for bias. Perspectives in Clinical Research 3:40. [PubMed: 22347702]

Scharfstein D, McDermott A, Olson W, Wiegand F (2014). Global sensitivity analysis for repeated measures studies with informative dropout: a fully parametric approach. Statistics in Biopharmaceutical Research 6:338–348.

Schomaker M, Gsponer T, Estill J, Fox M, Boulle A (2014). Non-ignorable loss to follow-up: correcting mortality estimates based on additional outcome ascertainment. Statistics in Medicine 33:129–142. [PubMed: 23873614]

Shih WJ (2002). Problems in dealing with missing data and informative censoring in clinical trials. Current Controlled Trials in Cardiovascular Medicine 3(1):4. [PubMed: 11985778]

Siskind V (2002). Quantification of completeness of follow-up. The Lancet 360:724.

Snapinn S, Jiang Q, Iglewicz B (2004). Informative noncompliance in endpoint trials. Current Controlled Trials in Cardiovascular Medicine 5:5. [PubMed: 15233844]

Sprague S, Leece P, Bhandari M, Tonetta P, III, Schemitsch E, Swiontkowski MF, on behalf of the S.P.R.I.N.T. investigators. (2003). Limiting loss to follow-up in a multicenter randomized trial in orthopedic surgery. Controlled Clinical Trials 24:719–725. [PubMed: 14662277]

Stinner DJ, Tennent DJ (2012). Losses to follow-up present a risk to study validity: differential attrition can be a shortcoming in clinical research. AAOS Now 6(2):38.

te Riele WW, Boerma D, Wiezer MJ, Borel Rinkes, I. H. M., van Ramshorst B (2010). Long-term results of laparoscopic adjustable gastric banding in patients lost to follow-up. British Journal of Surgery 97:1535–1540. [PubMed: 20564686]

Tsiatis A (1975). A nonidentifiability aspect of the problem of competing risks. Proceedings of the National Academy of Sciences of the United States of America 72:20–22. [PubMed: 1054494]

Walsh J, Devereaux PJ, Sackett DL (2015). Clinical trialist rounds: 28. When RCT participants are lost to follow-up. Part I: why even a few can matter. Clinical Trials 12:537–539. [PubMed: 26253052]

Wu Y, Furnary AP, Grunkemeier GL (2008a). Using the national death index to validate the noninformative censoring assumption of survival estimation. The Annals of Thoracic Surgery 85:1256–1260. [PubMed: 18355506]

Wu Y, Takkenberg JJM, Grunkemeier GL (2008b). Measuring follow-up completeness. The Annals of Thoracic Surgery 85:1155–1157. [PubMed: 18355488]

Table 1.

Percentage of trials with $z < -1.96$ and $z < -2.576$, for 10,000 simulated trials with 300 patients in each arm, for each of 30 conditions, with experimental and control median survival times both equal to 3 but with loss to follow-up (LTFU) for some experimental patients.

		***y* is uniform (*w* = 1) % rejection (one-tailed) at level of**		***y* is uniform (*w* = 2) % rejection (one-tailed) at level of**	
***b*, Weibull shape parameter**	**Experimental patients, LTFU parameter**	**0.025**	**0.005**	**0.025**	**0.005**
0.8	0%	2.62%	0.62%	2.49%	0.58%
	5	5.25	1.35	6.03	1.52
	5 to 15	4.97	1.20	8.01	2.02
	10	9.80	2.86	12.54	3.88
	15 to 5	15.67	4.43	16.76	5.74
1 (exponential distribution)	0	2.54	0.52	2.34	0.46
	5	4.55	1.04	5.77	1.64
	5 to 15	5.10	1.16	7.45	2.13
	10	8.73	2.34	11.74	3.78
	15 to 5	13.71	4.22	15.29	4.79
1.25	0	2.46	0.43	2.45	0.41
	5	4.89	1.11	5.53	1.28
	5 to 15	4.88	1.09	7.56	1.90
	10	8.24	2.40	10.89	3.18
	15 to 5	11.68	3.58	14.70	4.80

Table 2.

Percentage of trials with $z < -1.96$ and $z < -2.576$, for 10,000 simulated trials with 300 patients in each arm, for each of 30 conditions, with respective experimental and control median survival times of 4.2 and 3 and with loss to follow-up (LTFU) for some control patients.

		***y* is uniform ($w = 1$) % rejection (one-tailed) at level of**		***y* is triangular ($w = 2$) % rejection (one-tailed) at level of**	
***b*, Weibull shape parameter**	**Control patients, LTFU parameter**	**0.025**	**0.005**	**0.025**	**0.005**
0.8	0%	69.28%	45.24%	69.17%	44.63%
	5	55.58	32.29	52.68	28.99
	5 to 15	53.94	30.68	46.40	24.08
	10	41.36	20.72	36.21	17.07
	15 to 5	32.88	14.22	29.54	12.13
1 (exponential distribution)	0	87.50	69.60	87.41	69.59
	5	79.25	58.00	77.08	54.59
	5 to 15	78.57	56.35	71.44	48.22
	10	67.31	44.22	62.35	38.14
	15 to 5	60.71	36.41	57.01	32.24
1.25	0	97.38	91.49	97.86	91.16
	5	95.25	85.54	94.36	83.30
	5 to 15	94.62	83.87	91.70	78.22
	10	91.20	76.97	87.93	71.44
	15 to 5	87.19	69.60	84.47	65.29

Table 3.

Mean % of 300 patients who realized loss to follow-up (LTFU) and died, realized LTFU and survived, and realized LTFU in total, for the arm with LTFU (experimental arm for Table 1, control arm for Table 2), calculated from 40,000 or 20,000 simulated trials for each of 30 conditions.

		***y* is uniform (*w* = 1) % of patients in LTFU arm who were**			***y* is triangular (*w* = 2) % of patients in LTFU arm who were**		
***b*, Weibull shape parameter**	**Patients in LTFU arm, LTFU parameter**	**Lost and died**	**Lost and survived**	**Lost, total**	**Lost and died**	**Lost and survived**	**Lost, total**
0.8	0%	0%	0%	0%	0%	0%	0%
	5	3.0	1.0	4.0	3.0	0.6	3.6
	5 to 15	4.5	3.3	7.8	4.5	1.8	6.3
	10	6.0	2.0	8.0	6.0	1.3	7.3
	15 to 5	7.1	1.0	8.1	7.1	0.7	7.8
1 (exponential distribution)	0	0	0	0	0	0	0
	5	3.1	1.1	4.2	3.1	0.7	3.8
	5 to 15	4.9	3.2	8.1	4.9	2.0	6.9
	10	6.3	2.1	8.4	6.3	1.4	7.7
	15 to 5	7.1	1.2	8.3	7.1	0.9	8.0
1.25	0	0	0	0	0	0	0
	5	3.3	1.1	4.4	3.3	0.8	4.0
	5 to 15	5.4	3.0	8.4	5.4	2.0	7.4
	10	6.6	2.2	8.7	6.6	1.5	8.1
	15 to 5	7.2	1.3	8.5	7.2	1.0	8.2

Table 4.

Comparison of power when loss to follow-up (LTFU) is independent of survival versus dependent on it, for exponential survival, exponential LTFU in LTFU arm(control arm) for the independence condition, and respective experimental and control survival medians of 4.2 and 3.

					Power with one-tailed Type I error at level of					
		For LTFU independent of survival			0.025			0.005		
s, exponential parameter for LTFU*	Associated median LTFU time, (log 2)/*s*	Values found from (6) and (7): % of patients in LTFU arm (control arm) who are			LTFU independent of survival (power from PROC POWER)	LTFU dependent on survival (power from simulation, per Table 2)		LTFU independent of survival (power from PROC POWER)	LTFU dependent on survival (power from simulation, per Table 2)	
		Lost and die	Lost and survive	Lost, total		*w* = 1	*w* = 2		*w* = 1	*w* = 2
0	---	0%	0%	0%	87.4%	87.5%	87.4%	70.2%	69.6%	69.6%
0.02705	25.6	3.1	3.8	7.0	86.5	79.3	77.1	68.8	58.0	54.6
0.04337	16.0	4.9	5.9	10.8	86.0	78.6	71.4	67.9	56.4	48.2
0.05638	12.3	6.3	7.5	13.8	85.6	67.3	62.4	67.2	44.2	38.1
0.06489	10.7	7.1	8.5	15.6	85.3	60.7	57.0	66.7	36.4	32.2

* Values chosen so that “Lost and die” column above matches middle five rows of “Lost and died” columns in Table 3