Methods of Assessing Treatment Failure or Response with Informative Censoring

Jun Zhao[1], Qi Tang[1], Bo Fu[1], Qin Pan[2], Claire Tsao[3]

[1]Abbvie, Inc., 1 North Waukegan Road, North Chicago, IL 60045
[2]Chiltern International Ltd., 1016 W 9th Ave, King of Prussia, PA 19406
[3]Abbvie Biopharmaceuticals, Inc., 1500 Seaport Blvd., Redwood City, CA 94063

Abstract

In longitudinal clinical trials, we may define the onset of treatment response or treatment failure based on assessing relative change on clinical instruments that are commonly used in clinical practices. In some cases, the defined event needs to be confirmed at pre-defined later visit/visits after the event onset. However, when patients drop out of the study early or get an alternative medical intervention after the initial onset of the event, the subsequent clinical assessments are missing or considered as missing for the latter visits. Therefore, the event (either response or failure) cannot be confirmed and the information becomes censored. In this research, we compare several methods that are commonly used to deal with the sustained or confirmed response or failure when the missing data exist and the censoring may be informative. Simulation results are given to illustrate the methods to be applied and the bias introduced is also assessed for the corresponding method.

Key Words: longitudinal, confirmed treatment failure, informative censoring

1. Background

In longitudinal clinical trials, we may define the onset of treatment response or treatment failure based on assessing relative change on clinical instruments or assessments that are commonly used in clinical practices [1]. In schizophrenia trails, a well-accepted treatment response is defined as 30% or 50% improvement [2] from baseline in Positive and Negative Symptom Scale (PANSS) total score [3]. Sometimes, clinicians may use the change score of the Clinical Global Impression (CGI) to make decision on treatment failure or treatment response [4], or change score of the Expanded Disability Status Scale (EDSS) [5] to define the disability progression or disability improvement in multiple sclerosis.

In some cases, the defined event may need to be confirmed at pre-defined later visit/visits after the event onset. For example, in defining a sustained response in schizophrenia in a weekly assessed trial, the patients need to have at least two consecutive visits that meet the criteria of 30% decrease from baseline in PANSS total score [6].

When patients early drop out of the study or get an alternative medical intervention after the initial onset of the event, the subsequent clinical assessments are missing or

considered as missing for the latter visits. Even though in some cases the visits during the alternative medical intervention may be used to confirm the initial event occurred prior to the start of the alternative medical intervention, there is high chance that the event (response or failure) cannot be confirmed, therefore the information becomes censored.

For example in a 6-visits (weekly visits) schizophrenia trial you can expect the following scenarios of data pattern in defining the sustained/confirmed event (Table 1). The question arises on the scenarios 5 and 6, on whether they can be classified as the confirmed events.

Table 1: Scenarios of data pattern in defining the sustained response

Scenario	Visit 1	Visit 2	Visit 3	Visit 4	Visit 5	Visit 6	Event
[1]	o	o	o	o	o	o	No
[2]	o	o	o	response	o	o	No
[3]	o	o	response	response			Yes
[4]	o	o	o	.	.	.	censor
[5]	o	o	response	.	.	.	?
[6]	o	o	o	o	o	response	?

o: observation that doesn't meet criteria

Another example (Table 2) is to assess the confirmed disability progression (trimonthly visits), defined as patients who have at least a 1.0 point increase on the EDSS score (range 0-10) from baseline that is confirmed at the next visit.

Table 2: Scenarios of data pattern in defining the confirmed disability progression

Scenario	BL	Visit x	Visit x	Next Visit	Visit x	Last Visit	Event
[1]	o	o	o	o	o	o	No
[2]	o	o	progression	o	o	o	No
[3]	o	o	progression	confirmed	o		Yes
[4]	o	o	o	.	.	.	censor
[5]	o	o	progression	.	.	.	?
[6]	o	o	o	o	o	progression	?

o: observation that doesn't meet criteria

In clinical trials, it is expected the following issues may occur when we use the clinical assessment to define the event of interest. Observations are longitudinally collected at protocol defined visits, either weekly or monthly visit. Therefore, the events can only be detected at these pre-defined visits. The information between visits may be censored when the visit intervals are too far apart. Assessment instrument is subjective, e.g., PANSS, CGI, or EDSS scales. The endpoint is not normally distributed, and it is expected that the distribution might be skewed. Missing data occur due to early patient drop-out with a high rate, e.g., 30%-40% during study in schizophrenia trials [7]. Most importantly, when we assess the sustained/confirmed events, after we find the onset of the event, the confirmation visit may be missing, and the mechanism of missingness may not be random. This leads to the issue of informative censoring, which may cause biased resulting inference when performing survival analysis on the time-to-event data, e.g., Kaplan-Meier estimation [8].

In this paper, we are interested in exploring various statistical methods to analyze time-to- "confirmed event" data when non-random dropouts may exist after the initial event occurs. We refer this initial event as the "unconfirmed event". To explore the impact on the estimation by using different methods, we first simulated a true population without missing data and then created missing pattern and applied to the true population dataset to generate our "base dataset" that contains dropouts. From this "base dataset", we resampled 1000 datasets (referred as "studies") and then summarized the results with respect to hazard ratio, coverage probability, relative bias, % studies showing significant treatment difference for each method.

2. Methods of handling missing confirmation data

Assuming we only consider two treatment groups with an endpoint of time-to-confirmed event, which is derived from a longitudinally assessed clinical endpoint.

- Longitudinal assessments: Y_{ij} , i=1, … n (i^{th} subject) at j=1 to K (j^{th} visit)
- Event data: $T_i = (t_i, \delta_i)$ where t_i is the event onset time, and δ_i is the censorship indicator, i=1, … n (i^{th} subject)

We focus on how to handle missing data (i.e., missing confirmation visit resulted in unconfirmed event). Standard statistical analysis methods are used to analyze such data. They include but not limited to Fisher's exact, chi-square, or logistic regression for categorical endpoints, and survival analysis (log-rank, PH model) for the time-to-event endpoints.

The methods we are going to explore to deal with missing confirmation data after the initial onset of the events include the following:

- Observed case approach: only count confirmed events as the event of interest
- All unconfirmed events as confirmed
- Multiple imputation method/algorithm
- Joint modeling approach for longitudinal assessments and time-to-event data

2.1 Observed case approach

The simplest way to deal with the missing data in the analysis is to ignore them and only analyze observed data using a standard survival analysis method. In this approach, only confirmed events are considered as the events in the analysis. This approach assumes that the unconfirmed events with missing confirmation have small chance to be an event of interest, i.e., confirmed event, or the rate of missing data is small and can be ignored. Statistically speaking, if the missing mechanism is missing completely at random (random censoring in survival analysis) for both treatment groups, then this will be an appropriate approach to estimate the difference of event rates of interest between treatment groups. However, even though the approach is simple and interpretable, using only the observed data, the event rates may be underestimated. On the other hand, if the assumption of missing complete at random is violated, bias may occur.

2.2 All unconfirmed events as confirmed

Another simple way to deal with missing confirmation visit is to assume all unconfirmed events as the event of interest. In contrast to the previous observed cases approach, this

method can lead to overestimate the event rates of interest. In this method, both the confirmed events and unconfirmed events with missing confirmation are considered as event of interest. This approach has the same weakness as the "last observation carried forward" approach to impute the longitudinal missing data. By the same token, if the missing mechanism is missing completely at random (random censoring in survival analysis) for both treatment groups, it is a reasonable approach to estimate the difference of event rates of interest between treatment groups. However, bias may occur when the missing completely at random assumption is violated.

For the events of sustained failure or confirmed disability progression, another approach, which may be more reasonable to apply, is to include patients who have confirmed events or unconfirmed events but dropped out early due to lack of efficacy or due to adverse events associated with underlying disease. The unconfirmed events with early dropped-out due to other reasons are considered not confirmed events.

2.3 Multiple imputation method/algorithm

When the objective is to estimate the event rate of confirmed response or failure, or the corresponding time to confirm treatment response or failure, the above approaches in Section 2.1 and 2.2 may lead to biased estimations when drop-outs depend on treatment and/or event confirmation. We concentrate on the case where a subject has an unconfirmed treatment response or failure at their last available assessment and there is no further assessment available to confirm this response or failure. This scenario leads us to consider applying a multiple imputation method that may have a better property.

Algorithm: the presence or absence of a confirmed disability progression will be imputed using a Multiple Imputation (MI) approach [9]:

1. Among subjects with at least one event, regardless of whether confirmed or not, the probability of confirmation for those with a missing clinical assessment to confirm the previous event will be estimated via a logistic model, per treatment, adjusting for variables such as baseline score, change score to the tentative event.
2. Based on these probability estimates, confirmed events will be imputed via multiple imputations. The multiple imputation from this logistic regression model will be conducted multiple (e.g. 30-50) times to generate these (e.g. 30-50) complete analysis datasets.
3. Each of these complete data sets will be analyzed using a specific model (e.g. Cox PH, log-rank).
4. Finally, the statistics (e.g. hazard ratio, standard error, p-value) will be combined using Rubin's rule. SAS Procedures, PROC MI and PROC MIANALYZE will be used for analysis.

2.4 Joint modeling method

Another approach is to jointly consider the repeated measurements and the confirmed events simultaneously. Theraratically, the joint modeling [10] approach can increase efficiency by using longitudinal and time-to-event information and reducing uncertainty, and avoid intermittent measurements [11]. This is a likelihood based approach. The likelihood function is based on a joint distribution of $(Y_{ij}, T_i|v_i)$, where Y_{ij} is the assessment for subject i at visit j, T_i is the event time of that subject, i=1…N and

$j=1,\ldots,n_i$, and v_i is common random effects of the process of Y_{ij} and T_i. The likelihood function is as follows:

$$L = \prod_{i=1}^{N} \int f_Y\left(Y_{ij} \middle| v_i\right) f_T(T_i|v_i) f_v(v_i) dv$$

$$f_Y\left(Y_{ij}|v_i\right) \sim \beta_l^T X_{li} + v_{0i} + v_{1i} s_{ij}$$

$$\left(\text{event}_i,\ t_i\right) \sim \beta_s^T X_{si} + r_1 v_{0i} + r_2 v_{1i}$$

Where s_{ij} is the measurement time at visit j for subject i; t_i is the time-to-treatment failure for subject i ; X's are the baseline covariates for subject i.

$v_i = \{v_{0i}, v_{1i}\}$ is random effect with bivariate normal distribution $\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\alpha_1 & \alpha_{12}\\ \alpha_{21} & \alpha_2\end{pmatrix}\right)$, shared by both processes of the repeated longitudinal measurements and the confirmed events. A Weibull baseline hazards model with shape parameter γ is used as the time-to-event analysis in simulation studies.

3. Simulation

We use simulation to evaluate different approaches to deal with missing confirmation in the time to event data.. The simulated data set is for illustration purpose, and may or may not apply to certain disease area.

One "true" population was simulated without any missing data. This data set included longitudinal assessments at predefined visits for two treatment groups (A vs B). The simulation of these longitudinal assessments by the following two steps:

- Step 1: Simulated baseline scores (0 to 5 to avoid ceiling effect, with a mean of 2.5). For illustration purpose, we assume the scores are continuous. The data generated are by 0.5 increment.
- Step 2: Simulated post-baseline visits scores (9 visits) using baseline score and a transitioning matrix (from previous to current visit).

From the "true" population, generated a "base dataset" with an pre-assigned missing pattern. We assume the missing is monotone. This missing pattern was based on a pre-defined probability function from some observed change scores for each visit. The early drop-out rates: A vs B is summarized in Table 3 below.

Table 3: Monotone drop-out (%) in the simulated data set

Visit	1	2	3	4	5	6	7	8	9
A	0.8	3.3	6.0	7.8	10.1	12.8	16.0	16.5	31.4
B	2.6	6.5	9.2	11.9	14.2	16.7	18.8	20.5	36.0

Summary of simulated events in the "base dataset":

- At least one event: A vs B = 31.3% vs 39.9%
- Confirmed: A vs B = 15.3% vs 20.1%
- Confirmation unavailable: A vs B = 3.3 % vs 3.1 %
- Other: did not meet definition: A vs B = 12.7% vs 16.7%

In order to evaluate the performance of each method applying the dataset we simulated from the "base dataset", we resampled 1000 studies using bootstrapping. Each data set includes two treatment parallel groups (A and B) with a sample size of n=800 each group. Some of the key simulation results are summarized in the Table 4 below. The results are based on the 1000 bootstrapping data sets, by applying the Cox PH regression on time to event adjusted by the baseliine value.

Table 4: Summary of simulation results for each method used

Method	**Hazard Ratio (A vs B)**			**P-value**	
	Mean(Median) (True: 0.661)	Coverage Prob.	Relative Bias	Mean(Median) (True: <0.000)	% (<=0.05)
Observed cases	0.715 (0.713)	97.6%	8.2%	0.020 (0.004)	88.6%
Unconfirmed evants as events	0.671 (0.670)	99.3%	1.5%	0.001 (<0.001)	99.8%
Imputation methods (MI)	0.664 (0.663)	99.5%	0.5%	0.003 (<0.001)	99.6%
Joint modeling*	0.724 (0.717)	94.3%	9.4%	0.079 (0.021)	65.8%

Coverage prob. = % of 95% CIs covering the true HR. Relative bias = 100x[mean HR – true HR)/true HR], where i=1, …1000.

*: Only the studies with convergent estimates are included. The coefficients of random terms used as connection in the joint modeling are both positive and significant for random slop and intercept.

4. Discussion

This paper was motivated by an informative dropout issue which we observed in some clinical trials. We attempted to simulate some datasets with the features we would like to study. However, the base dataset we simulated turned out to have a large effect size (i.e., HR=0.66) and high power. As a result, the comparisons among the 4 approaches did not differentiate too much. Nonetheless, our illustration shows the following points.

- The "unconfirmed events as events" and the MI approaches have a higher power, reasonable coverage probability, and less bias.
- The "observed case" approach suggests lower power, large bias, and lower coverage probability.

We did a similar exercise using our study where the effect size was worse than that for this exercise. The conclusion about these 3 approaches was the same. For the joint modeling approach, more work needs to be done before we can comment on its appropriateness for the scenarios we studied in this paper.

There is possibility that the confirmation visit might be missing when we observe unconfirmed events for the sustained/confirmed treatment failure or response endpoints. The missingness might be informative. Multiple methods can be applied to deal with the confirmed events with missing data. However, the results may not be consistent and bias may be introduced due to violation of certain analysis assumptions for a specific method. It is suggested, as stated in other literature [12, 13], to minimize the informative missing, and prospectively identify a primary analysis method followed by several supportive analyses (i.e. sensitivity analyses) to confirm the primary findings.

Acknowledgements

The authors acknowledge the programmer Yang Han for his help in fixing simulation programs. This manuscript was sponsored by AbbVie, Inc. AbbVie contributed to the design, research, and interpretation of data, writing, reviewing, and approving the content. Jun Zhao, Qi Tang, Bo Fu, and Claire Tsao are employees of AbbVie, Inc. Qin Pan was an employee of AbbVie when the research was conducted, and is now an employee of Chiltern International Ltd.

References

1. Micha Mandel, Susan A. Gauthier, Charles R. G. Guttmann, Howard L. Weiner, and Rebecca A. Betensky. Estimating Time to Event From Longitudinal Categorical Data: An Analysis of Multiple Sclerosis Progression. J Am Stat Assoc. 2007 Dec; 102(480): 1254–1266.
2. Leucht S, Davis JM, Engel RR, Kissling W, Kane JM: Definitions of response and remission in schizophrenia: recommendations for their use and their presentation. Acta Psychiatr Scand Suppl 2009, 7-14.
3. Kay SR, Fiszbein A, Opler LA: The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull 1987, 13:261-276.
4. Haro JM, Kamath SA, Ochoa S, et al. The Clinical Global Impression-Schizophrenia Scale: A simple instrument to measure the diversity of symptoms present in schizophrenia. Acta Psychiatr Scand. 2003; 107(416):16–23.
5. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology. 1983 Nov; 33(11):1444-52.
6. Leucht S, Zhao J. Early improvement as a predictor of treatment response and remission in patients with schizophrenia: a pooled, post-hoc analysis from the asenapine development program. J Psychopharmacol. 2014 Apr; 28(4): 387–94.
7. Lieberman JA, Stroup TS, McEvoy JP, et al. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. N Engl J Med 2005; 353(12):1209-1223.
8. Federico Campigotto and Edie Weller, Impact of Informative Censoring on the Kaplan-Meier Estimate of Progression-Free Survival in Phase II Clinical Trials. Journal of Clinical Oncology. Volume ew Number 27, September 2014.
9. Yue Zhao, Amy H. Herring, Haibo Zhou, Mirza W. Ali, and Gary G. Koch. A Multiple Imputation Method for Sensitivity Analysis of Time-to-Event Data with Possible Informative Censoring. J Biopharm Stat. 2014 ; 24(2): 229–253.

10. Michael S. Wulfsohn, Anastasios A. Tsiatis. A Joint Model for Survival and Longitudinal Data Measured with Error. Biometrics. 1997 Mar; Vol. 53, No. 1, pp. 330-339.
11. Dimitris Rizopoulos. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Chapman and Hall/CRC, June 2012.
12. National Research Council. The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press 2010.
13. Fotios Siannis, John Copas, Guobing Lu. Sensitivity analysis for informative censoring in parametric survival models. Biostatistics (2005), **6**, 1, pp. 77–91.