Published in final edited form as:

Med Care. 2009 July ; 47(7 Suppl 1): S115–S119. doi:10.1097/MLR.0b013e31819bc08a.

Cost Analysis with Censored Data

Yijian Huang, PhD

Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia

Abstract

Economic evaluation of medical interventions has become an accepted, and often required, adjunct to the standard effectiveness and safety assessment in clinical research. However, the statistical analysis can be challenging due to censored cost data, as commonly obtained in medical studies. Over the past decade, important statistical issues that arise from censored cost data have been identified and a number of advances made to tackle them. In this article, we will describe these issues, including induced dependent censoring and limited identifiability with the cost distribution, and review recent statistical developments. Available methods address either time-restricted medical cost or lifetime medical cost jointly with survival time. Their applicability and limitation in various practical situations will be discussed.

Keywords

Identifiability; Induced dependent censoring; Lifetime medical cost; Statistical method; Time-restricted medical cost

While demands on our health care system continue to outgrow the resources available, the need to effectively control medical care expenditure becomes ever more urgent. Cost assessment has become an integral component of comprehensive medical treatment evaluation, in addition to standard effectiveness and safety assessment. A recent example is a phase III clinical trial to treat advanced non-small-cell lung cancer patients, comparing Vinorelbine plus Cisplatin versus Paclitaxel plus Carboplatin.¹ The primary endpoint was the effectiveness measure of survival time. Meanwhile, an important secondary study objective was to assess medical cost associated with these two competing treatments. As the leading cause of cancer-related death in the United States, lung cancer has been estimated to cost society \$4.7 billion annually in direct medical costs.² Treating lung cancer with minimal economic burden possible has become a priority. Consequently, such cost evaluation in clinical studies is increasingly being used to inform and influence decision making regarding the use of new interventions.

However, cost analysis can be challenging due to the presence of censoring, i.e., with incomplete follow-up data. Censoring may be due to study termination or participant drop-out. In the aforementioned lung cancer trial, resource utilization was tracked for supportive care medications, blood products, medical procedures, protocol and non-protocol related treatments, and medical care for inpatient days or outpatient visits. These cost data were collected every 3 months for the first 6 months, and every 6 months thereafter, up to 24 months. Figure 1 shows the scatter plot of cumulative medical cost (in 1998 US dollars) and follow-up duration for 408 eligible study participants, where a dot indicates an observed death and a circle indicates a censored individual. Clearly, censoring was not uncommon in this study. One

prominent cause was the two-year duration of cost data collection, beyond which cumulative cost was censored for all surviving participants.

Apparently, mean cost estimates calculated by averaging observed cumulative cost from all the study participants or else from complete, i.e., uncensored, individuals are biased. On the other hand, standard survival analysis techniques³ have been shown to be problematic if directly applied to cost analysis.⁴ In fact, there are several unique issues associated with censored cost data, posing serious challenges to the analysis. Over the past decade, a number of advances have been made in statistical methodology to address these challenges. We provide a review of these developments in this article, focusing on cost analysis in clinical trials and observational studies. Important statistical issues associated with censored cost data will be described first. Available statistical methods will then be reviewed. The advantages and limitations of these methods will be discussed.

Statistical issues associated with censored cost data

Medical cost is cumulative over time. Of interest is typically the cost incurred from the initiation of the study until death, i.e., lifetime medical cost. Unfortunately, death may not be observed for every participant in a clinical study and therefore lifetime medical cost is subject to censoring. A couple of unique issues then emerge.

Induced dependent censoring

Write T as survival time and C as the censoring time. Due to censoring, these underlying random variables are only observed through follow-up time X and censoring indicator Δ (i.e., whether death is observed):

$$X = T \wedge C, \quad \Delta = I(T \leq C),$$

where $\wedge$ denotes minimization and $I(\cdot)$ is the indicator function. As in the lung cancer trial, survival time T is typically the primary outcome and standard survival analysis techniques are used to accommodate censoring. These standard survival analysis techniques include the Kaplan-Meier estimator for the estimation with a homogeneous sample (i.e., the one-sample problem), logrank test for sample comparison (i.e., the two-sample problem), and the Cox proportional hazards model for regression analysis with a heterogeneous sample. It is important to note that a basic but critical assumption behind these standard methods is the independent censoring mechanism, i.e., T and C are independent (though possibly upon conditioning on covariates in the regression problem). Fortunately, this assumption is often plausible in practice.

When medical cost is considered, one might find that lifetime medical cost, say U , bears much similarity with survival time T . Both are measures of the disease process (toward death), albeit on different scales, time and cost. Just as T is time to death, U can be referred to as cost to death. Denote the cost accumulated up to time t by $A(t)$. Then, lifetime medical cost is

$$U = A(T).$$

In the presence of censoring, U (like T) is not always observed. Indeed, the cost accumulation process $A(\cdot)$ is only observed over the follow-up period $[0, X]$. The accumulated medical cost at the end of the follow-up is

$$W=A(X).$$

Since $A(\cdot)$ is non-decreasing, we have

$$W=A(T) \wedge A(C)=U \wedge A(C),$$

where $A(C)$ can be viewed as the “censoring cost”, i.e., the counterpart of censoring time, C , on the cost scale. With this perspective, it is tempting to apply standard survival analysis techniques to the cost scale for the inference of U . Such an approach has been suggested in the literature.⁵ However, its validity fails since the censoring pattern on the cost scale is typically dependent, as will now be discussed.

Censoring on the cost scale is unique in that it is induced from censoring on the time scale. Suppose that the censoring time is independent of the disease process including survival time T and cost accumulation process $A(\cdot)$, though possibly upon conditioning on covariates in the regression problem. If the cost accumulation process is deterministic, the independence between T and C implies that between U and $A(C)$. But in general $A(\cdot)$ is stochastic instead, and consequently U and $A(C)$ are typically induced to be dependent.⁴ To understand this phenomenon, consider a hypothetical scenario where the cost accumulation rate is constant over time for each individual, but the rate may vary from individual to individual. In this case, one has

$$A(t)=Rt,$$

where random variable R is the cost accumulation rate. Then,

$$W=RX=RT \wedge RC.$$

Even though T and C are independent, RT and RC are generally dependent. An individual with a higher cost accumulation rate tends to incur more medical cost at both survival time and censoring time. In reality, the cost accumulation process can be more complicated, e.g., with higher rate typically after disease diagnosis and also prior to death. So long as the process is stochastic, the induced censoring pattern on the cost scale is generally dependent. This dependence implies that standard survival analysis techniques become invalid when adopted to the cost scale.

Marginal identifiability of the cost distribution

As one cause of censoring, clinical studies typically have a fixed study duration shorter than the longest survival time. In the lung cancer study, the cost collection was limited to two years. Consequently, an issue of concern is identifiability, i.e., whether or not a quantity of interest can be uniquely determined from the distribution of the observed random variables. In the case of survival time T , its distribution is only identifiable up to the study duration but not beyond. The estimation and inference of survival time are necessarily based on the distribution of T over the study duration only. Nevertheless, this proves adequate for most practical purposes.

Unfortunately, the identifiability issue becomes much more concerning and thorny when lifetime medical cost U is considered. In many studies, it is not uncommon to observe that a

certain proportion of the study participants incur little cost during the study. Thus, little information on lifetime medical cost for these individuals can be observed. Consequently, the marginal distribution of lifetime medical cost, in the one-sample setting, is not identifiable anywhere over the support of the cost distribution from 0 to ∞ . This identifiability issue is intimately related to induced dependent censoring. Both originate from the stochastic nature of the cost accumulation process.

A review of existing statistical methods

Due to the practical importance and perhaps the challenges as well, tremendous research efforts have been devoted to the development of statistically sound methods for medical cost estimation over the past decade. Since the marginal distribution of lifetime medical cost may not be identifiable anywhere beyond 0, an immediate question concerns the specific cost measure to address. This leads to two broad classes of statistical methods, one focusing instead on time-restricted medical cost and the other on the distribution of lifetime medical cost jointly with survival time.

Imposing time limit

To avoid the non-identifiability issue associated with the marginal distribution of lifetime medical cost, an obvious fix is to consider time-restricted medical cost instead. For example, two-year restricted cost is the cost accumulated up to two years or death, whichever occurs earlier. Then, identifiability is no longer an issue for the time-restricted medical cost so long as the time limit does not exceed the study duration. Often the time limit is chosen to be the same as the study duration.

We start with the one-sample problem, where the main focus is on estimating the mean of time-restricted medical cost. Let L be the time limit and L -restricted survival time $T^L = T \wedge L$. The L -restricted medical cost is $U^L = A(T^L)$. A number of estimators have been proposed, including those by Lin et al.,⁴ Bang and Tsiatis,⁶ Strawderman,⁷ and Zhao and Tian.⁸ For many of them, Zhao et al.⁹ showed that they are closely related, or even identical under certain conditions. To account for the effect of censoring, a fundamental approach is the Horvitz-Thompson-type method¹⁰ to reweight complete cases. Here, a complete case refers to an individual who either dies or is followed up to time limit L . Write

$$X^L = T^L \wedge C, \quad W^L = A(T^L \wedge C), \quad \Delta^L = I(T^L \leq C),$$

which all are observed random variables. A complete case corresponds to $\Delta^L = 1$. Let the survivor function of the censoring time be $S_C(t) = \Pr(C > t)$. Then, one can find

$$E \left\{ \frac{\Delta^L W^L}{S_C(X^L)} \right\} = E \left[E \left\{ \frac{\Delta^L W^L}{S_C(X^L)} \middle| T^L, U^L \right\} \right] = E \left[E \left\{ \frac{U^L}{S_C(T^L)} I(T^L \leq C) \middle| T^L, U^L \right\} \right] = E(U^L).$$

The above equation is actually quite intuitive: An individual with restricted survival time T^L has the probability $S_C(T^L)$ of being observed as uncensored. Therefore, censoring can be taken into account by weighting each uncensored individual with $S_C(X^L)^{-1}$; note that $T^L = X^L$ for an uncensored individual. Add subscript i to each random variable to denote realization of the random variable for individual i , $i = 1, \dots, n$. The above equation motivates the simple weighted estimator as referred to by Bang and Tsiatis:⁶

$$n^{-1} \sum_{i=1}^n \frac{\Delta_i^L W_i^L}{\widehat{S}_c(X_i^L)},$$

where $\widehat{S}_c$ is the Kaplan-Meier estimator for censoring time using data $\{X_i^L, \Delta_i^L\}_{i=1}^n$.

The preceding estimator utilizes data $\{X_i^L, W_i^L, \Delta_i^L\}_{i=1}^n$ only. When the cost accumulation process $A(\cdot)$ is also observed, Bang and Tsiatis⁶ proposed to partition $[0, L]$ into small intervals, compute a similar Horvitz-Thompson-type estimate for cost incurred in each interval, and sum over these intervals. This so-called partitioned estimator makes full use of the cost accumulation data. Although not always true, the partitioned estimator tends to be more efficient than the simple weighted estimator in most practical situations. Furthermore, based on the missing data theory,^{11,12} Bang and Tsiatis⁶ derived estimators in an attempt to improve on estimation efficiency. Nevertheless, their simulations suggested that the efficiency gain is typically limited, if any, over the partitioned estimator.

In the two-sample problem, these one-sample results can be utilized in a straightforward fashion to construct test statistics for the comparison of time-restricted medical cost.⁸ Furthermore, the Horvitz-Thompson method can be similarly used in the regression problem to account for censoring. Among others, Lin^{13,14} developed mean regression procedures, and Bang and Tsiatis¹⁵ suggested a median regression procedure. See also Bang¹⁶ for a discussion.

Imposing a time limit is effective in avoiding the identifiability issue. This strategy has been widely used in practice, and existing statistical methods are fairly abundant. For the lung cancer trial, the estimated survival rate at 24 months is approximately 15%, and the identifiability issue with lifetime medical cost is indeed of concern. Ramsey et al. considered two-year restricted cost to compare the two treatments.¹⁷ This analysis might be relatively easy to interpret since the difference in survival between the two treatments was not statistically significant. However, for a study where an intervention has a significant effect on survival, interpreting time-restricted medical cost can be difficult. Indeed, the time limit is typically artificial and time-restricted medical cost may not be a good approximation for lifetime medical cost. More importantly, identical distributions of lifetime medical cost do not imply identical distributions of time-restricted medical cost, because the survival distributions are different; see Figure 2. Therefore, caution should be exercised in the interpretation.

Joint distribution with survival time

Given the nowhere identifiability result, the estimation and inference on lifetime medical cost might appear hopeless. Nevertheless, Huang and Louis showed that lifetime medical cost can be estimated jointly with survival time although not marginally.¹⁸ In statistical terminology, a random variable that is observed only upon the occurrence of an event is a mark of the event.

Lifetime medical cost is a mark of death. With data $\{X_i^L, W_i^L, \Delta_i^L\}_{i=1}^n$ in the one-sample problem, Huang and Louis revealed that the joint distribution of T and U can be identified over $[0, \tau] \times [0, \infty)$, where τ is the study duration.¹⁸ This is illustrated in Figure 3. Due to limited study duration, the joint distribution is apparently not identifiable over $[\tau, \infty) \times [0, \infty)$, which leads to nowhere identifiability of the marginal distribution of U . In fact, lifetime medical cost U is always censored for an individual with survival time $T > \tau$. Nevertheless, the joint distribution is identifiable on $[0, \tau] \times [0, \infty)$. Furthermore, Huang and Louis derived the nonparametric maximum likelihood estimator for the joint distribution.¹⁸ This estimation procedure is a generalization of the Kaplan-Meier method. The Kaplan-Meier estimator for the distribution of T may be formulated through Efron's¹⁹ redistribution-to-the-right algorithm, where

probability mass, say $\hat{p}_i$, is assigned to each uncensored individual i , i.e., with $\Delta_i = 1$. Then, the estimator of Huang and Louis for the joint distribution function $F(t, u) = \Pr(T \leq t, U \leq u)$ is given by

$$\widehat{F}(t, u) = \sum_{i: \Delta_i = 1} \widehat{p}_i I(X_i \leq t, W_i \leq u).$$

The survival-time marginal of this joint distribution estimator is the Kaplan-Meier estimator.

This identifiability result in the one-sample problem has strong implications on the two-sample problem. It is clear now that testing for lifetime medical cost cannot be intuitively based on its marginal distribution. Huang and Lovato suggested a test based on the joint distribution with survival time.²⁰ Denote survival time and lifetime medical cost for sample j , $j = 1, 2$, by $T^{(j)}$ and $U^{(j)}$, respectively. In a hypothetical scenario where $T^{(1)}$ and $T^{(2)}$ have identical distributions, a reasonable test statistic can be constructed under the null hypothesis that the joint distributions of $(T^{(1)}, U^{(1)})$ and $(T^{(2)}, U^{(2)})$ are equivalent. However, more generally, $T^{(1)}$ and $T^{(2)}$ may have different distributions but the difference is a nuisance, i.e., not of direct interest, for the purpose of cost comparison. For this reason, the difference in survival time is calibrated first using a scale change model, i.e., $T^{(1)} = \beta T^{(2)}$. After the scale change parameter β is estimated, the equivalence of the joint distributions of $(T^{(1)}, U^{(1)})$ and $(\beta T^{(2)}, U^{(2)})$ is then tested as a means of comparing lifetime medical cost between the two samples.

The same idea for the two-sample problem can be extended to the regression problem. Write $\mathbf{Z}$ as the covariate vector. Huang proposed the following calibration regression model:²¹

$$\log \begin{pmatrix} T \\ U \end{pmatrix} = \begin{pmatrix} \beta'_0 \\ \beta'_1 \end{pmatrix} \mathbf{Z} + \varepsilon,$$

where β_0 and β_1 are regression coefficient vectors, and the error term ε has a completely unspecified bivariate distribution. This is a natural generalization of the accelerated failure time model for survival time. An estimation and inference procedure has been developed. The above model can be slightly extended with the logarithm transformation replaced with any known transformation.

By considering the joint distribution with survival time, this strategy overcomes the identifiability issue with the marginal distribution of lifetime medical cost. Huang applied calibration regression to the lung cancer trial data.²¹ The treatment effects on survival time and lifetime medical cost were assessed simultaneously, while adjusting for covariates including age and lactate dehydrogenase normality indicator. However, there are limitations with this approach as well. First, for the two-sample and regression problems, modeling assumptions are imposed on the joint distribution. If the study duration is short, these assumptions could be difficult to check. Second, as far as cost data are concerned, only lifetime medical cost for uncensored individuals are utilized for all the procedures previously described. This can be an advantage since it accommodates various cost collection schemes, from the least informative scenario with W observed only to the most informative scenario with continuous observation of $A(\cdot)$ over $[0, X]$. However, cost accumulation data, if available, are not used to good advantage. There might be room for efficiency gain, which requires further investigation.

Discussion

With censored cost data, induced dependent censoring is better-known than the identifiability issue with the marginal cost distribution. But the latter is more difficult. For cost analysis in a clinical trial, the strategy of targeting time-restricted medical cost would be reasonable if the intervention has little effect on survival time. Otherwise, interpreting the analysis results can be difficult. On the other hand, the joint modeling strategy suits the situation where the survival rate is reasonably small at the end of the study, say less than 20%. In this case, the model assumptions can be checked over a reasonable support of survival time and lifetime medical cost.²¹

However, in many clinical studies, one may face the situation that an intervention has an effect on survival time and the study duration is relatively short in the sense that the survival rate is around, say, 50% or higher at the end of the study. In this case, none of the existing methods (as discussed here) might be suitable to address the cost comparison. With such limitation in available data, lifetime medical cost may be too much to hope for. On the other hand, time-restricted medical cost can be misleading since it is tied to not only cost but also survival time. O'Brien discussed extrapolation beyond study duration by modeling future cost to estimate lifetime medical cost.²² This is a fairly common approach to coping with limited study duration in practice. However, the estimation can be sensitive to the modeling assumptions for extrapolation. Further developments are needed.

Throughout this article, it is assumed that the censoring time is independent of the disease process, conditionally on covariates if applicable. In the case of censoring due to study termination, this assumption holds true generally. However, it might be questionable with patient drop-out. When the assumption does not hold, modeling the censoring time becomes necessary and statistical methods described in this article might be extended.

Acknowledgments

This article was based on a presentation at the "Health Care Costs: Standardized Methods and Estimates for Research and Policy Applications" workshop, Rockville, MD, December 6-7, 2007, sponsored by the National Cancer Institute (NCI) and the Agency for Healthcare Research and Quality (AHRQ). The author thanks Dr Joseph Lipscomb, Dr Martin Brown, Dr Robin Yabroff, other participants of the workshop, and the reviewers for constructive comments and suggestions which have led to numerous improvements. Support by the National Cancer Institute is gratefully acknowledged.

References

1. Kelly K, Crowley J, Bunn PA Jr, et al. Randomized phase III trial of Paclitaxel plus Carboplatin versus Vinorelbine plus Cisplatin in the treatment of patients with advanced non-small-cell lung cancer: A Southwest Oncology Group Trial. *J Clin Oncol* 2001;19:3210–3218. [PubMed: 11432888]
2. Brown ML, Lipscomb J, Snyder C. The burden of illness of cancer: economic cost and quality of life. *Annu Rev Public Health* 2001;22:91–113. [PubMed: 11274513]
3. Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. Vol. 2. New York: Wiley; 2002.
4. Lin DY, Feuer EJ, Etzioni R, et al. Estimating medical costs from incomplete follow-up data. *Biometrics* 1997;53:419–434. [PubMed: 9192444]
5. Fenn P, McGuire A, Philips V, et al. The analysis of censored treatment cost data in economic evaluation. *Med Care* 1995;33:851–863. [PubMed: 7637406]
6. Bang H, Tsiatis AA. Estimating medical costs with censored data. *Biometrika* 2000;87:329–343.
7. Strawderman RL. Estimating the mean of an increasing stochastic process at a censored stopping time. *J Am Statist Assoc* 2002;95:1192–1208.
8. Zhao H, Tian L. On estimating medical cost and incremental cost-effectiveness ratios with censored data. *Biometrics* 2001;57:1002–1008. [PubMed: 11764238]

9. Zhao H, Bang H, Wang H, et al. On the equivalence of some medical cost estimators with censored data. *Statistics in Medicine* 2007;26:4520–4530. [PubMed: 17380543]
10. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Statist Assoc* 1952;47:663–685.
11. Robins JM.; Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell, N.; Dietz, K.; Farewell, V., editors. *AIDS Epidemiology-Methodological Issues*. Boston: Birkhauser; 1992. p. 297-331.
12. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Statist Assoc* 1994;89:846–866.
13. Lin DY. Linear regression analysis of censored medical costs. *Biostatistics* 2000;1:35–47. [PubMed: 12933524]
14. Lin DY. Regression analysis of incomplete medical cost data. *Statistics in Medicine* 2003;22:1181–1200. [PubMed: 12652561]
15. Bang H, Tsiatis AA. Median regression with censored cost data. *Biometrics* 2002;58:643–649. [PubMed: 12229999]
16. Bang H. Medical cost analysis: Application to colorectal cancer data from the SEER Medicare database. *Contemporary Clinical Trial* 2005;26:586–597.
17. Ramsey SD, Moinpour CM, Lovato LC, et al. Economic analysis of Vinorelbine plus Cisplatin versus Paclitaxel plus Carboplatin for advanced non-small-cell lung cancer. *JNCI* 2002;94:291–297. [PubMed: 11854391]
18. Huang Y, Louis TA. Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika* 1998;85:785–798.
19. Efron, B. *Proceedings of the Fifth Berkeley Symposiums in Mathematical Statistics IV*. New York: Prentice-Hall; 1967. The two sample problem with censored data; p. 831-853.
20. Huang Y, Lovato L. Tests for lifetime utility or cost via calibrating survival time. *Statistica Sinica* 2002;12:707–723.
21. Huang Y. Calibration regression of censored lifetime medical cost. *J Am Statist Assoc* 2002;97:318–327.correction, 97:661
22. O'Brien, BJ. Cost-effectiveness analysis in cancer: toward an iterative framework for integration of evidence from trials and models. In: Lipscomb, J.; Gotay, CC.; Snyder, C., editors. *Outcomes Assessment in Cancer*. Cambridge: Cambridge University Press; 2005. p. 503-521.

Figure 1. Scatter plot of cumulative medical cost (in 1998 US dollars) and follow-up duration of 408 eligible study participants in the lung cancer trial

Figure 2.
Two hypothetical individuals have the same lifetime medical cost but different time-restricted medical cost, as a result of different survival time

Figure 3. The joint distribution of survival time and lifetime medical cost is only identifiable in the shaded area, which results in the nowhere identifiability of the cost marginal distribution