Published in final edited form as:

Chance (N Y). 2008 September ; 21(3): 55–58.

Misreporting, Missing Data, and Multiple Imputation: Improving Accuracy of Cancer Registry Databases

Yulei He [Assistant Professor],

Department of Health Care Policy, Harvard Medical School, 180 Long-wood Avenue, Boston, MA, 02115 (E-mail: he@hcp.med.harvard.edu)

Recai Yucel [Assistant Professor], and

Department of Epidemiology and Biostatistics, School of Public Health, University at Albany, SUNY, One University Place, Rensselaer, NY 1214 (E-mail: ryucel@albany.edu)

Alan M. Zaslavsky [Professor]

Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA, 02115 (E-mail: zaslavsky@hcp.med.harvard.edu)

Cancer registries collect information on type of cancer, histological characteristics, stage at diagnosis, patient demographics, initial course of treatment including surgery, radiotherapy, and chemotherapy, and patient survival (Hewitt and Simone 1999). Such information can be valuable for studying the patterns of cancer epidemiology, diagnosis, treatment, and outcome. However, misreporting on registry information is unavoidable, and thus studies based solely on registry data would lead to invalid results.

Past literature has documented the inaccuracy of registry records on adjuvant, or supplemental, chemotherapy and radiotherapy. The Quality of Cancer Care (QOCC) project (Ayanian et al. 2003) used data from the the California Cancer Registry, the largest geographically contiguous population-based cancer registry in the world, to study the patterns of receiving and reporting adjuvant therapies for stage II/III colorectal cancer patients. The study surveyed the treating physicians for a subsample of the patients in the registry to obtain more accurate reports of whether they have received adjuvant therapies. This study confirmed the inaccuracy of the registry data in favor of underreporting. Table 1 (line 2 vs line 1), which is based on this study, implies substantial underreporting of 20% and 13% in chemotherapy and radiotherapy rates, respectively.

Given that the registry is a valuable data source in health services research, how can we improve quality of inferences using the comprehensive but inaccurate registry database? Consider, for example, our goal is to obtain accurate estimates of treatment rates from the misreported records in the registry. A simple approach is to use only the validation sample, i.e. the physician survey data collected in the QOCC project. However, due to logistic reasons, the survey sample (< 2000 patients) was much smaller than the registry sample (> 12000 patients) used in the study, and hence analyzing the validation sample alone would greatly reduce precision, especially for complex estimands such as regression estimates.

Another approach, the errors-in-variables method (Carroll et al. 2006), would analyze the registry data while adjusting for the reporting error. This approach typically involves modeling the relationship between the correct values and misreported ones, represented here by the validation sample and the corresponding registry data. Using information from both sources, it should yield valid results with increased precision. However, the statistical sophistication of the error-adjustment procedures might be challenging for analysts who typically do not possess statistical expertise to implement such methods.

A more appealing strategy might be multiple imputation (Rubin 1987). In a typical nonresponse problem, this method first “fills-in” (imputes) missing variables several times to create multiple completed datasets. Analysis can then be conducted for each set using complete-data procedures. The results obtained from separate sets of completed data are combined into a single inference using simple rules. In the problem of misreporting, the essence of applying this strategy is to impute the uncollected correct treatment variables in the remainder of the registry, and then to perform analysis on the completed/corrected data. Figure 1 illustrates this strategy. The corrected registry data can then be used by practitioners without any additional modeling effort. As with the errors-in-variables approach, the imputation model characterizes the measurement error process and makes the adjustment. The imputer may also incorporate additional information which may not generally be available to analysts, such as information from other administrative databases, into the imputation model to further improve the analyses (Yucel and Zaslavsky 2005; Zheng et al. 2006).

An Imputation Approach

Consider a single therapy variable in the QOCC data, e.g. the adjuvant chemotherapy. Let $Y_{(O)}$ and $Y_{(R)}$ denote the true and reported status of the treatment in the registry sample, respectively. Both of them are binary variables with 1= “Yes” and 0= “No”. In addition, we assume that only underreporting takes place in the registry, that is, $Y_{(R)} = 0$ if $Y_{(O)} = 0$, and $Y_{(R)}$ could be either 1 or 0 if $Y_{(O)} = 1$. Such an assumption is at least almost entirely accurate for the QOCC data. Variable $Y_{(R)}$ is complete for the registry. Variable $Y_{(O)}$ is observed in the validation sample but missing for the remainder of the registry.

We first consider a simplest case with no predictors. By our assumption of underreporting, the missing $Y_{(O)}$ must be 1 if the corresponding reported $Y_{(R)}$ is 1. On the other hand, $Y_{(O)}$ could be either 1 or 0 if $Y_{(R)}$ is 0. By Bayes’ theorem, we can compute the probability that $Y_{(O)}$ is 1 given that $Y_{(R)}$ is 0:

$$P(Y_{(O)}=1|Y_{(R)}=0)=\frac{P(Y_{(R)}=0|Y_{(O)}=1)P(Y_{(O)}=1)}{P(Y_{(R)}=0)} \quad (1)$$

The conditional probability that $Y_{(O)}$ is 0 is one minus this probability.

Because of the incomplete $Y_{(O)}$, $P(Y_{(O)} = 1)$ (the true treatment rate) and $P(Y_{(R)} = 0|Y_{(O)} = 1)$ (the rate of underreporting) are unknown and need to be estimated. If the values of $Y_{(O)}$ had been completely observed in the registry and assuming a uniform prior distribution for the two probabilities, these rates have Beta distributions whose parameters are determined by the corresponding number of 1’s and 0’s in the sample. The Beta distribution for estimating $P(Y_{(O)} = 1)$ is $\text{Beta}(\#(Y_{(O)} = 1) + 1, \#(Y_{(O)} = 0) + 1)$. The Beta distribution for estimating $P(Y_{(R)} = 0|Y_{(O)} = 1)$ is $\text{Beta}(\#(Y_{(R)} = 1, Y_{(O)} = 1) + 1, \#(Y_{(R)} = 0, Y_{(O)} = 1) + 1)$. One can easily generate values of these probabilities from their distributions given complete data. Given values of the probabilities, one can compute $P(Y_{(O)} = 1|Y_{(R)} = 0)$ and draw a value of $Y_{(O)}$ from a Bernoulli random variable with this probability. The imputation procedure first initializes the rates with reasonable prior guesses and then iterates between the following two steps until the rate estimates achieve convergence. Figure 2 illustrates this procedure. The steps in the iteration are stated in sequence below.

Step 1. Impute the missing values of $Y_{(O)}$ as Bernoulli draws with probability of success given by (1).

Step 2. Based on the completed data, update $P(Y_{(O)} = 1)$ and $P(Y_{(R)} = 0|Y_{(O)} = 1)$, and hence re-calculate $P(Y_{(O)} = 1|Y_{(R)} = 0)$.

Multiple imputations of $Y_{(O)}$ are created by repeating the procedure several times with different starting values for the initial (unknown) probabilities. After convergence of the algorithm, one more draw of all missing $Y_{(O)}$ values is taken, yielding one full set of imputed values.

We now describe the imputation model of $Y_{(O)}$ for the real data. Clinical and/or demographical information recorded in the cancer registry, such as age, gender, and comorbidity, are good candidates as predictors of both the receipt and reporting of adjuvant therapies. For example, younger patients are more likely to receive chemotherapy than elders (Ayanian et al. 2003). Furthermore, both treatment and reporting rates might vary across providers (hospitals). Incorporating patient-level covariates and clustering effects across hospitals in the imputation model helps to improve the model predictions.

Let X denote the covariate information and assume it is fully observed and recorded without error in the registry. The joint distribution of $Y_{(O)}$ and $Y_{(R)}$ given X can be decomposed into two parts, the outcome model $P(Y_{(O)}|X, \theta_{(O)})$ and the reporting model $P(Y_{(R)}|Y_{(O)}, X, \theta_{(R)})$, i.e.

$$P(Y_{(O)}, Y_{(R)}|X, \theta) = P(Y_{(O)}|X, \theta_{(O)})P(Y_{(R)}|Y_{(O)}, X, \theta_{(R)}).$$

The former corresponds to clinical processes relating receipts of the treatment with patient/hospital characteristics, and the latter characterizes the ways in which misreporting occurs in the registry. They can be viewed as generalizations of the true treatment and underreporting rates from (1). For each part, we can apply a logistic or probit regression model with hospital random effects, using $\theta_{(O)}$ and $\theta_{(R)}$ to denote the corresponding model parameters.

We assume the missingness of $Y_{(O)}$ in the QOCC data is at random (Little and Rubin 2002), i.e., dependent only on the observed covariate variables denoted by X . Because the validation sample included all patients from certain regions within a defined period, this can be considered as planned missingness, or missingness due to study design. Consequently, by capturing these factors in the imputation model, missing at random is a plausible assumption.

Similar to the simple case without any covariates, application of Bayes' theorem underlies the calculation of the probability used in imputation in the presence of covariates X ,

$$P(Y_{(O)} = 1|Y_{(R)} = 0, X, \theta) = \frac{P(Y_{(R)} = 0|Y_{(O)} = 1, X, \theta_{(R)})P(Y_{(O)} = 1|X, \theta_{(O)})}{P(Y_{(R)} = 0, X, \theta_{(R)})}. \quad (2)$$

As shown in Figure 3, the imputation algorithm iterates between estimating the outcome and reporting model probabilities and imputing $Y_{(O)}$ until the estimates for $\theta_{(O)}$ and $\theta_{(R)}$ achieve convergence. That is, the sampling continues until it is judged that the algorithm is sampling from the target posterior distributions. Then one final draw of all missing values of $Y_{(O)}$ produces a completed data set. The process is repeated multiple times with different starting probability values in order to produce multiple imputed data sets.

Application

Study sample

From the 10 regional cancer registries in California, the QOCC project selected all ($n = 12594$) patients age 18 or older who were newly diagnosed with stage III colon cancer or stage II or III rectal cancer and underwent surgery during the years 1994 to 1997. This registry sample included patients from 433 hospitals. From records of those patients diagnosed and treated in 1996 and 1997 in registry regions 1, 3, and 8, representing the San Francisco/Oakland, San Jose, and Sacramento areas in Northern California, respectively, the patients' treating physicians were identified and mailed a written survey, asking whether their patients received adjuvant chemotherapy or radiotherapy based on their medical records. The survey cohort included 1956 patients. Physician responses, or direct abstracts from medical records by registry staff, were obtained for 1450 (74%) of these patients treated at 98 hospitals.

In the combined dataset, patient-level covariates include age, gender, cancer stage at diagnosis, race, marital status, hospital transfer (whether the patient was transferred between diagnosis and treatment), comorbidity scores, and the median income of the patient's census block group. Hospital characteristics include hospital volume, presence of tumor registry accredited by the American College of Surgeons (ACOS) Commission on Cancer, teaching status, and location (urban versus non urban).

Imputation model fitting

We imputed the adjuvant chemotherapy and radiotherapy data separately for each patient for uncollected true therapy status of the registry sample. Each patient received 30 imputations. The adequacy of imputation models was checked by comparing predictions under the model to data that actually were observed. Indeed, the models generated predictions that were pretty similar to the observed data.

Several variables were predictive of *receiving treatments*. Chemotherapy was received more often by younger, married, or stage III rectal cancer patients, and less often by those with lower income, stage II rectal cancer, or more comorbidities. Patients who were transferred before surgery, typically to hospitals with more specialized facilities for cancer care, were also more likely to receive the treatment. Patients at ACOS hospitals were more likely to receive chemotherapy, whereas those at teaching hospitals appeared to receive it less often. Patients in the region in which the survey was conducted were also more likely to receive the treatment, as were patients who were treated in 1996–1997 (as opposed to 1994–1995). The effects of patients' age, marital status, comorbidity, hospital transfer, ACOS hospital status, and year of treatment on the receipt of radiotherapy were similar to those on the receipt of chemotherapy. Male patients were more likely to receive radiotherapy. Patients with stage II or III rectal cancer were much more likely to receive radiotherapy than those with stage III colon cancer.

Factors predicting *completeness of reporting* are also of interest to inform efforts to validate or improve the quality of registry data. Chemotherapy for older or married patients was more often underreported. High-volume hospitals reported chemotherapy more completely than others, as did urban hospitals; this might reflect greater investments in data management in these institutions. Radiotherapy was also more completely reported in the high-volume hospitals and for patients with stage II or III rectal cancer. There were fewer significant predictors for the reporting of radiotherapy than chemotherapy, suggesting a better and more consistent pattern of reporting in the former.

The random effect estimates showed that there existed moderate variation among hospitals in provision and reporting of chemotherapy but less for radiotherapy, and there is little correlation between the random effects of treatment receiving and reporting processes.

Registry data analyses

Table 1 lists the estimated rates of adjuvant therapies using the survey, the uncorrected registry, and the multiply imputed registry data. Rates calculated from the imputed data are significantly larger than those from the registry alone (line 4 vs. line 3), confirming the previous findings of underreporting. They are also substantially lower than those from the survey alone (line 4 vs. line 1) because the rates of therapies were estimated to be lower in the part of the state outside the survey area.

Adjuvant therapy variables can be important predictors of clinical outcomes. Using multiply imputed datasets, we fitted a logistic regression model for patients' two-year survival with predictors including the receipt of adjuvant chemotherapy and other covariates. Receiving chemotherapy was a strong prognostic factor of survival (odds ratio=1.26, standard error=0.09), consistent with results from clinical trial literature. Although certain variability has been introduced from different imputations, the multiple imputation analysis still increases the precision compared to analyzing the survey data alone because the survey sample is much smaller than the registry sample. For example, the standard error of the coefficient of receiving chemotherapy in the latter approach is about 70% larger than that in the former.

Public Health: A critical role for statistical methodology

To correct the underreporting of adjuvant therapies in a cancer registry, we multiply imputed the accurate treatment information using a model based on a validation sample. This illustrated how inferential tools using multiple imputation can be adapted to deal with a single inaccurately-reported therapy variable. There are several extensions with substantive and statistical importance that could be pursued. A first extension pertains to incorporating the dependency and correlation between the two therapies in either the receipt or reporting process. Another extension would be to build the imputation model based on information from multiple sources, such as the survey data, medical records, and claims data. This added information could improve predictions and statistical inference.

In addition to the adjuvant therapy variables, patient's birthplace and cancer stage also suffer from misreporting in the registry. Furthermore, misreporting often occurs for important quality indicators or indexes in other administrative systems such as claims databases. The multiple imputation strategy constitutes a promising tool to tackle this general problem in using public databases for health services research.

References

- Ayanian JZ, Zaslavsky AM, Fuchs CS, Guadagnoli E, Creech CM, Cress RD, O'Connor LC, West DW, Allen ME, Wolf RE, Wright WE. Use of adjuvant chemotherapy and radiation therapy for colorectal cancer in a population-based cohort. *Journal of Clinical Oncology* 2003;21:1293–1300.
- Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. *Measurement Error in Nonlinear Models: A Modern Perspective*. Vol. 3. New York, NY: CRC Press; 2006.
- Hewitt, M.; Simone, JV. *Ensuring Quality Cancer Care*. Washington, DC: National Academy Press; 1999.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. Vol. 2. New York, NY: Wiley press; 2002.
- Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley press; 1987.
- Yucel RM, Zaslavsky AM. Imputation of binary treatment variables with measurement error in administrative data. *Journal of American Statistical Association* 2005;100:1123–1132.
- Zheng H, Yucel RM, Ayanian JZ, Zaslavsky AM. Profiling providers on use of adjuvant chemotherapy by combining cancer registry and medical record data. *Medical Care* 2006;44:1–7. [PubMed: 16365606]

Figure 2.

Imputation Scheme for a Simplest Case: After initial values of $P(Y_{(O)} = 1)$ and $P(Y_{(R)} = 0| Y_{(O)} = 1)$ are chosen, $P(Y_{(O)} = 1|Y_{(R)} = 0)$ is computed. The algorithm iterates between imputing missing values of $Y_{(O)}$ on the left and drawing new probabilities on the right. After the iterations converge to the target posterior distribution, a final draw of $Y_{(O)}$ produces a set of imputations. The process is repeated multiple times.

Figure 3.

Imputation Scheme for the Real Data: After initial values of parameters in the models $P(Y_{(O)} = 1|X, \theta_{(O)})$ and $P(Y_{(R)} = 0|X, Y_{(O)} = 1, \theta_{(R)})$ are chosen, probabilities $P(Y_{(O)} = 1|X, Y_{(R)} = 0, \theta)$ are computed. The algorithm iterates between imputing missing values of $Y_{(O)}$ on the left and drawing parameters in the probability models on the right. After the iterations converge to the target posterior distribution, a final draw of $Y_{(O)}$ produces a set of imputations. The process is repeated multiple times.

Table 1

Adjuvant Therapy Rates % (SE)

Sample	Chemotherapy	Radiotherapy
Survey	73.3 (1.16)	25.4 (1.14)
Registry (in the survey region)	57.9 (0.79)	22.2 (0.67)
Registry (statewide)	51.4 (0.45)	19.6 (0.35)
Imputed registry (statewide)	61.2 (0.77)	23.1 (0.61)