Published in final edited form as:
Biometrics. 2009 September ; 65(3): 946–952. doi:10.1111/j.1541-0420.2008.01164.x.

Combining Information from Cancer Registry and Medical Records Data to Improve Analyses of Adjuvant Cancer Therapies

Yulei He and
Department of Health Care Policy, Harvard Medical School, Boston, 02115, USA

Alan M. Zaslavsky
Department of Health Care Policy, Harvard Medical School, Boston, 02115, USA

Yulei He: he@hcp.med.harvard.edu; Alan M. Zaslavsky: zaslavsky@hcp.med.harvard.edu

SUMMARY

Cancer registry records contain valuable data on provision of adjuvant therapies for cancer patients. Previous studies, however, have shown that these therapies are underreported in registry systems. Hence direct use of the registry data may lead to invalid analysis results. We propose first to impute correct treatment status, borrowing information from an additional source such as medical records data collected in a validation sample, and then to analyze the multiply imputed data, as in Yucel and Zaslavsky (2005). We extend their models to multiple therapies using multivariate probit models with random effects. Our model takes into account the associations among different therapies in both administration and probability of reporting, as well as the multilevel structure (patients clustered within hospitals) of registry data. We use Gibbs sampling to estimate model parameters and impute treatment status. The proposed methodology is applied to the data from the Quality of Cancer Care project, in which stage II or III colorectal cancer patients were eligible to receive adjuvant chemotherapy and radiation therapy.

Keywords

Cancer registry data; Hierarchical Bayesian model; Medical records data; Misclassification; Missing data; Multiple imputation

1. Introduction

Cancer registries collect information on type of cancer, histological characteristics, stage at diagnosis, patient demographics, initial course of treatment including surgery, radiation therapy, and chemotherapy, and patient survival (Hewitt and Simone, 1999). Such information can be valuable for studying variations in quality of cancer care, for example, across racial and ethnic groups. Concerns have been raised, however, about the completeness of treatment information in cancer registries. Bickel and Chassin (2000) and Malin et al. (2002) demonstrated underreporting of adjuvant chemotherapy and radiation therapy for breast cancer in hospital and state registries, respectively. Cress et al. (2003) reported similar treatment underreporting for colorectal cancer in a state registry and showed that it was associated with both patient and hospital characteristics. Thus, studies based solely on registry data would lead to invalid results.

Correspondence to: Yulei He, he@hcp.med.harvard.edu; Alan M. Zaslavsky, zaslavsky@hcp.med.harvard.edu.

The classical errors-in-variables approach (Carroll et al., 2006) often used in epidemiology would be to analyze the relationship of registry data on treatment to clinical outcomes, and adjust for reporting error. This approach might involve modeling the relationship between the correct values of therapy variables in the validation sample and misreported/misclassfied ones in the registry. The error-adjustment procedures are often complicated and are analysis-specific. On the other hand, the therapy variables may be used by many researchers in analyses for various scientific purposes. Implementation of the error-adjustment procedures might be challenging for analysts who do not possess the relevant specialized statistical expertise.

A more appealing strategy might be multiple imputation (Rubin, 1987). In a typical missing data problem, this method first "fills in" (imputes) missing variables several times to create multiple completed datasets. Analysis of each set can then be conducted using standard complete-data procedures. Finally, the results obtained from separate completed datasets are combined into a single inference using simple rules. In the presence of underreporting, this strategy is applied by imputing the uncollected correct treatment variables in the registry outside the validation sample. The imputer also may incorporate additional information which may not generally be available to other analysts, such as from other administrative databases (Rubin, 1987, 1996). The imputation model characterizes the misclassification process and makes the adjustment. The corrected registry data can then be analyzed without any additional modeling of underreporting.

Yucel and Zaslavsky (2005) (henceforth "YZ") proposed statistical models for imputing receipt of adjuvant chemotherapy using data from the California Cancer Registry and from medical records obtained from a physician follow-back survey, a validation sample for the registry data. Cancer treatment patterns may vary across hospitals. Similarly, the cancer registry data are aggregated from hospital registries, whose completeness of reporting may vary due to differences in registrar resources, provider network structures, and other organizational factors. Hence YZ's model included individual and hospital level predictors, as well as hospital random effects for provision and reporting of chemotherapy. They used multiply-imputed data sets to estimate models for mortality within two years of treatment. Using the same models, Zheng et al. (2006) profiled hospitals based on imputed rates of chemotherapy for colorectal cancer.

The method proposed by YZ focused on a single treatment variable. But patients may receive multiple therapies in the course of treatment. For example, Malin et al. (2002) developed individual quality scores to measure the receipt of each treatment (surgery, lymph node dissection, radiation therapy, and tamoxifen/chemotherapy) by eligible breast cancer patients, and added these scores to summarize overall quality. Furthermore, reporting completeness for different treatments may also be associated. Ignoring such associations when correcting the registry data may bias results of analyses concerning multiple therapies. In this paper, we extend YZ's method to impute the underreported status of multiple treatment variables. This approach borrows strength from the validation sample to correct the misclassification in the registry system, accommodating the associations among the multiple therapies.

In Section 2, we present the statistical models. In Section 3 we analyze data from our motivating example. Finally in Section 4 we suggest directions for future research.

2. Statistical models

2.1 General framework

As in YZ, we let S_1 and S_2 denote the validation sample (those for whom medical record abstraction was performed) and the remainder of the population (those with only registry data), respectively, so the entire population is $S = S_1 \cup S_2$. True treatment status is assumed to be positive (treatment was provided) if the treatment is recorded in either the registry or the medical records abstract data; if the registry records show that the patient did not receive a treatment, it might due to underreporting. We refer to receipt of each therapy as a statistical "outcome" of the imputation model, although these can act as either dependent or independent variables in the complete-data analyses (after imputation).

Let $\mathbf{Y}_O = (Y_{O1}, Y_{O2}, \dots, Y_{OL})$ represent the true treatment status of L therapies, with $Y_{Ol} = 1$ if the person has received treatment l ($l = 1, \dots , L$) and 0 otherwise. The corresponding treatment status as reported to the registry is $\mathbf{Y}_R = (Y_{R1}, Y_{R2}, . . . , Y_{RL})$. Under our assumptions, $\mathbf{Y}_O$ is directly observed among patients in S_1, for whom both the registry data and medical records are known, but not in S_2, while $\mathbf{Y}_R$ is always observed in both S_1 and S_2. The relationship between Y_{Ol} and Y_{Rl} is partially deterministic ($Y_{Rl} = 0$ if $Y_{Ol} = 0$), and partially stochastic (Y_{Rl} is a Bernoulli variable if $Y_{Ol} = 1$), reflecting our assumption of stochastic underreporting. Potential covariates $\mathbf{X}$ (characteristics of patients and health care providers) are also recorded for all patients in the registry and assumed to be accurate in both S_1 and S_2. The statistical goal is to impute $\mathbf{Y}_O$ in S_2 from the model $f(\mathbf{Y}_O, \mathbf{Y}_R|\mathbf{X}, \theta)$. Since $Y_{Ol} = 1$ if $Y_{Rl} = 1$, the imputation is stochastic for unobserved Y_{Ol} only when $Y_{Rl} = 0$.

YZ considered the case where $L = 1$. They factorized the joint distribution over S as

$$f(Y_{O1,}Y_{R1}|\mathbf{X},\theta) \propto f_O(Y_{O1}|\mathbf{X},\theta_O)f_R(Y_{R1}|Y_{O1},\mathbf{X},\theta_R),$$

where the outcome model $f_O(Y_{O1}|\mathbf{X}, \theta_O)$ represents the relationship of receipt of treatment with patient and hospital characteristics, and the reporting model $f_R(Y_{R1}|Y_{O1}, \mathbf{X}, \theta_R)$ characterizes the ways in which misclassification occur in the data. Note that this taxonomy differs from that used in the classical errors-in-variables approach (Clayton, 1992). Our reporting model corresponds to the "error/measurement" component of the classical model. However, whereas the latter specifies a model for a disease outcome predicted by an exposure measured with error, the "outcome" in our imputation model could be either exposure or outcome in the analysts' complete-data models.

The implicit assumption that the outcome and reporting models hold with the same parameter values in S_1 and S_2 is referred to as *transportability across different studies* in classical measurement error theory (Carroll et al., 2006, Chap. 1). In our motivating example, the medical records data (validation study) are collected from a subsample of the registry sample and hence constitute an *internal* part of the main study. If we assume the validation sample is representative of the whole registry population after controlling for selection factors, the transportability assumption naturally holds and the final analysis can be applied to the completed data including both samples.

2.2 Model specifications

We propose a class of multivariate extensions of YZ's univariate model, using a similar factorization into the outcome and reporting models, i.e.

$$f(\boldsymbol{Y}_O,\boldsymbol{Y}_R|\mathbf{X},\theta) \propto f_O(\boldsymbol{Y}_O|\mathbf{X},\theta_O)f_R(\boldsymbol{Y}_R|\boldsymbol{Y}_O,\mathbf{X},\theta_R).$$

With $L > 1$, several types of associations exist among multiple therapies, including the associations among different outcomes, Y_{Oi}, Y_{Oj}, associations of reporting of different therapies, Y_{Ri}, Y_{Rj}, and possible dependency of reporting of one treatment, Y_{Ri}, on the actual receipt of the other treatment, Y_{Oj}. Such associations might occur at the individual level, the hospital level, or both. For example, at the individual level, colorectal cancer patients receiving radiation therapy might have chemotherapy reported more completely because radiation therapy is provided in more centralized facilities (e.g. radiation center) than chemotherapy, and the facilities report both chemotherapy and radiation therapy more systematically to the registry than do the hospitals or doctors' offices.

We use multivariate hierarchical probit models to characterize these outcome and reporting processes, allowing us to describe the dependency structure parsimoniously in terms of correlation coefficients of underlying continuous latent variables, and the variation across hospitals through random effects. In addition, Bayesian analysis for probit models can be conducted easily via a simple auxiliary variable Gibbs sampling algorithm (Chib and Greenberg, 1998). Specifically, for the true status of treatment l of the jth person at the ith hospital ($i = 1,\ldots, m$, $j = 1,\ldots, n_i$), let $Y_{Olij} = 1$ if $Z_{Olij} > 0$ and 0 otherwise. The latent variables $\{Z_{Olij}\}$ follow a multivariate random-effects model

$$(Z_{O1ij}, Z_{O2ij}, \ldots, Z_{OLij})^T \sim N(\mathbf{X}_{Oij}\boldsymbol{\beta}_O + \mathbf{W}_{Oij}\boldsymbol{\gamma}_{Oi}, \boldsymbol{\rho}_O). \tag{1}$$

Matrix $\mathbf{X}_{Oij} = \mathrm{diag}(X_{O1ij}, X_{O2ij}, \ldots, X_{OLij})$ contains model covariates, including all or a subset of the variables in $\mathbf{X}_{ij}$ together with any desired transformations or interactions, whereas $\mathbf{W}_{Oij} = \mathrm{diag}\,(W_{O1ij}, W_{O2ij}, \ldots, W_{OLij})$ contains covariates (including at least an intercept) whose coefficients vary across the hospitals. Parameter $\boldsymbol{\beta}_O = (\beta_{O1}^T, \beta_{O2}^T, \ldots, \beta_{OL}^T)^T$ concatenates the fixed-effects coefficients for each treatment, assumed to be common across all hospitals, and parameter $\boldsymbol{\gamma}_{Oi} = (\gamma_{O1i}^T, \gamma_{O2i}^T, \ldots, \gamma_{OLi}^T)^T$ contains random effects specific to hospital i. A correlation matrix $\boldsymbol{\rho}_O = \{\rho_{Oij}\}$ characterizes the correlations among the multiple latent variables for the outcome data that cannot be explained through the common predictors.

Similarly, let Y_{Rlij} and Z_{Rlij} denote the reported value of the lth treatment for patient j at hospital i and the corresponding latent variable, respectively, so that $Y_{Rlij} = 1$ if $Y_{Olij} = 1$ and $Z_{Rlij} > 0$, and 0 otherwise. The corresponding multivariate random-effects model for $\{Z_{Rlij}\}$ is

$$(Z_{R1ij}, \ldots, Z_{RLij})^T \sim N(\mathbf{X}_{Rij}\boldsymbol{\beta}_R + \boldsymbol{\alpha}\mathbf{Y}_{Oij} + \mathbf{W}_{Rij}\boldsymbol{\gamma}_{Ri}, \boldsymbol{\rho}_R), \tag{2}$$

where covariate matrix $\mathbf{X}_{Rij}$, $\mathbf{W}_{Rij}$, and parameters $\boldsymbol{\beta}_R$, $\boldsymbol{\gamma}_{Ri}$, and $\boldsymbol{\rho}_R$ are the analogs of the corresponding covariates and parameters from the outcome latent variable model (1). Parameter $\boldsymbol{\alpha}$ is an $L \times L$ matrix with zeros on the diagonal since no identifiable dependency exists between Y_{Rlij} and Y_{Olij}, while the off-diagonal element α_{ij} models the effect of receipt of treatment i on the reporting of a different treatment j. More general formulations might include interactions between $\mathbf{X}_{Rij}$ and $\mathbf{Y}_{Oij}$, or replace $\boldsymbol{\alpha}$ with hospital-specific random effects $\boldsymbol{\alpha}_i$. We use the simple formulation (2) throughout this paper.

If both outcome and reporting latent variable models include random effects corresponding to the same units, i.e. hospitals, then these effects might be correlated. For example, hospitals with a higher actual rate of treatment might also tend to have better reporting systems. Therefore, we assume a joint multivariate normal distribution for the random effects,

$$\boldsymbol{\gamma}_i = (\boldsymbol{\gamma}_{Oi}^T, \boldsymbol{\gamma}_{Ri}^T)^T \sim \mathrm{iid} N(0, \boldsymbol{\Sigma}).$$

On the other hand, if the validation sample data set is much smaller than the fallible registry data set, we might fit the more parsimonious fixed-effects model for the reporting and the more general mixed-effects model for outcomes, conjecturing that random variation in reporting might influence inferences less than does variation in actual rates of treatments, and fix some α_{ij}'s at zero if there is little data to estimate the corresponding effects.

Finally, we assume an improper uniform or diffuse normal prior distribution for fixed regression coefficients $\boldsymbol{\beta}_O$, $\boldsymbol{\beta}_R$, and $\boldsymbol{\alpha}$, and a proper uniform prior for $\boldsymbol{\rho}_O$ and $\boldsymbol{\rho}_R$ (Gelman et al., 2004, pg. 483–484), letting the posterior inferences be dominated by the observed data. We assume a proper inverse Wishart prior on $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma} \sim IW_r(v,\Lambda)$, where $v \geq r$ and $\Lambda > 0$.

2.3 The missingness mechanism

We assume that the unobserved $\mathbf{Y}_O$ in S_2 are missing at random (MAR) (Rubin, 1987), that is, the probability that an outcome variable is observed depends on the observed data, such as the fully observed $\mathbf{X}$ in the registry, but not on quantities that are missing. Under MAR, the inference can be made solely based on the posterior distributions of the model parameters while ignoring the missingness mechanism.

In our example, inclusion of the patient in the survey depends on the geographical region and year of treatment. Such "planned missingness" for the registry sample excluded from the survey might be MAR and ignorable if the corresponding selection indicators are included in the models. On the other hand, to generalize the outcome model outside the domain represented by S_1, we need more specific assumptions about homogeneity of the reporting process. In the medical records abstraction, because only certain years and regions were included, we need to assume that patients in hospitals of the physician survey are similar to those not selected, conditional on their hospital and patient covariates and observed registry-reported outcomes. Another part of unobserved $\mathbf{Y}_O$ came from the physician survey nonresponses and was shown to be related to known hospital and patient characteristics (Cress et al., 2003). The MAR assumption becomes more plausible as the model is formulated to include more predictors in $\mathbf{X}$ related to nonresponse.

To make inferences regarding comparisons among regions and periods, or among hospitals in different regions and periods, however, it would be preferable to design the survey to represent the entire state through the time span. For example, an annual ongoing quality measurement process might draw a stratified sample representing patients from each of the 10 regions.

2.4 Inferential algorithm

We use Gibbs sampling (Gelfand, Racine-Poon, and Smith, 1990) to draw inferences for the model parameters and impute the missing values. The main steps are sketched here; details for the example with two therapies appear in the Web Appendix.

- **a.** Draw the latent variables $\mathbf{Z}_O$ and $\mathbf{Z}_R$ from truncated multivariate normal distributions.
- **b.** Draw fixed effects coefficients $\boldsymbol{\beta}_O$, $\boldsymbol{\beta}_R$, and $\boldsymbol{\alpha}$ from multivariate normal distributions.
- **c.** Draw random effects $\{\gamma_{Oi}, \gamma_{Ri}\}$ from a multivariate normal distributions of each i.
- **d.** Draw elements of $\boldsymbol{\rho}_O$ and $\boldsymbol{\rho}_R$ using the adaptive rejection Metropolis sampling algorithm (Gilks, Best, and Tan, 1995).
- **e.** Draw missing values of $\mathbf{Y}_O$ from multiple Bernoulli distributions where the probabilities are estimated from the functions of cumulative distributions of L-variate normal distributions.

Excluding the part for the prior distribution, the posterior distribution of the reporting model parameter θ_R involves two parts, that is, $f(\mathbf{Y}_{RS_1}|\mathbf{Y}_{OS_1}, \mathbf{X}_{S_1}, \theta_R)$ and $f(\mathbf{Y}_{RS_2}|\mathbf{X}_{S_2}, \theta_R, \theta_O)$ The latter is the marginal of $f(\mathbf{Y}_{RS_2}, \mathbf{Y}_{OS_2}|\mathbf{X}_{S_2}, \theta_R, \theta_O)$ where $\mathbf{Y}_{OS_2}$ is unobserved but imputed in the Gibbs sampling algorithm from the outcome model in S_2, $f(\mathbf{Y}_{OS_2}|\mathbf{X}_{S_2}, \theta_O)$ YZ considered the fact that S_2 has a much larger size than S_1 so a slight misspecification in either the outcome or reporting model might result in an invalid inference for θ_R from $(\mathbf{Y}_{RS_2}, \mathbf{X}_{S_2})$ that overwhelms the information from the fully observed and hence more direct and valid $(\mathbf{Y}_{RS_1}, \mathbf{Y}_{OS_1}\ \mathbf{X}_{S_1})$. They proposed to base the inference for θ_R solely on the validation sample likelihood $f(\mathbf{Y}_{RS_1} | \mathbf{Y}_{OS_1}, \mathbf{X}_{S_1}, \theta_R)$ and the prior distribution. The corresponding Gibbs sampling algorithm draws θ_R from a conditional distribution that is incompatible with the joint distribution (Gelman, 2004), but this might provide more robust inferences against model misspecification. We adopt their strategy in this paper.

3. Application

3.1 Study sample

The Quality of Cancer Care project combines data from the California Cancer Registry with a survey of physicians to study patterns of care for colorectal cancer, including receipt of adjuvant therapies, across various clinical and demographic factors (Ayanian et al., 2003). From the 10 regional cancer registries in California, we select 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. Our sample includes 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, 99% of the qualified patients' treating physicians were identified and mailed a written survey, asking whether their patients received adjuvant chemotherapy or radiation therapy based on their medical records. The survey cohort included 1956 patients, and physician responses, or direct abstracts from medical records by Registry staff, were obtained for 1450 (74%) of these patients treated at 98 hospitals. Thus, the physician survey/medical records abstraction constitutes the validation sample (S_1) for the California Cancer Registry data (S).

Patients' age, gender, cancer stage at diagnosis, race, marital status, hospital transfer (whether the patient was transferred between diagnosis and treatment), and adjuvant therapies (chemotherapy and radiation therapy) were obtained from the cancer registry. Socioeconomic status was represented by the median income of the patient's census block group. We classified the degree of comorbidity using the Deyo adaptation of the Charlson comorbidity scale (Deyo, Cherkin, and Ciol, 1992), based on conditions identified in hospital records from 18 months before to 6 months after cancer diagnosis. Hospital characteristics included in our models were volume, presence of tumor registry accredited by the American College of Surgeons (ACOS) Commission on Cancer, teaching status, and urban location. The small portion (0.01%–7.02%) of missing data for patient and hospital characteristics was imputed using stochastic regression imputation (Little and Rubin, 2002, Chap. 4).

3.2 Model fitting

Based on national guidelines (National Institute of Health, 1990), all patients included were eligible for adjuvant chemotherapy, and those with rectal cancer were also eligible for adjuvant radiation therapy. As expected, a majority of rectal cancer patients (66.6% from the survey and 55.3% from the registry) had received radiation therapy, as had a substantial number of stage III colon cancer patients (10.5% from the survey and 7.22% from the registry). We analyzed the whole sample, modeling the receipt and reporting of both treatments simultaneously in a bivariate model regardless of guideline eligibility for radiation therapy. For the purpose of comparison, we also analyzed each adjuvant therapy separately using univariate models.

In both imputation methods, the outcome model, but not the reporting model, includes an indicator for being in the region covered by the follow-back survey, and another for being treated in 1996 or 1997. In the bivariate analysis, the reporting model for one treatment includes the receipt of the other as a potential predictor. We included correlated random hospital intercepts in the outcome model with an inverted Wishart prior, i.e. $\mathbf{\Sigma}_2 \sim IW(2, (2I)^{-1})$, but we did not include hospital random effects in the reporting part of the model. The latter stabilizes model estimation due to the much smaller number of hospitals in the physician survey compared to the registry, as well as the highly unbalanced sample size across these hospitals. The priors used in the univariate models were similar to those in the bivariate model except that we adopted a diffuse $IG(1, 1)$ prior for random effects variance of chemotherapy and radiation therapy separately.

Based on time-series plots and sample autocorrelation plots for the model parameters, we concluded that the posterior series had converged after 2500 iterations of the Gibbs chain. The estimates also appeared to be stable under several trials of different initial values of parameters deliberately chosen to be overdispersed. Hence we based our inferences and collected imputations after discarding the first 2500 iterations. We performed 30 imputations so that multiple imputation efficiency (relative to an infinite number of imputations) higher than 97% can be achieved (Rubin, 1987, pg. 114), collecting imputed data sets that were widely separated in the Gibbs chain to minimize serial correlation.

3.3 Main results

Table 1 lists the estimated rates of adjuvant therapies using the survey, registry, and multiply imputed/corrected registry data from the univariate and bivariate models. The multiple-imputation estimates were obtained using the combining rules proposed in Rubin (1987). Rates calculated from imputed data under both methods are substantially larger than those from the registry alone.

Table 2 (in the Web Appendix) shows the parameter estimates for the bivariate model. Chemotherapy was used more often among younger, married, or stage III rectal cancer patients, and less often among 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, while 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 or 1997. For the receipt of radiation therapy, the effect of patients' age, marital status, comorbidity, hospital transfer, ACOS hospital status, as well as year of treatment are similar to those for receipt of chemotherapy. Male patients were more likely to receive radiation therapy. Patients with stage II or III rectal cancer were much more likely to receive radiation therapy than those with stage III colon cancer.

High-volume hospitals reported chemotherapy more completely than others, as did urban hospitals; this might reflect greater investments in data management in these institutions. Treatment of older or married patients was more often underreported. Finally, receipt of radiation therapy is a strong predictor ($\hat{\beta}_{R,radiation} = 0.477$, SE=0.144) of reporting chemotherapy, confirming our conjecture.

There are fewer significant predictors for the reporting of radiotherapy, indicating a more consistent pattern of reporting. In contrast to chemotherapy, which can be administered in ambulatory settings outside hospitals, such as clinics and doctors' offices, radiation therapy is typically administered in centralized radiation centers where data recording and reporting are also likely to be more systematic. Radiation therapy is also more completely reported in the high-volume hospitals. In addition, patients with stage II or III rectal cancer were reported more completely than others for radiotherapy. There is some but not significant evidence

($\hat{\beta}_{R,chemo} = 0.458$, SE=0.305) that reporting of radiotherapy is more complete among patients who had received chemotherapy.

As expected, the two therapies are strongly correlated at the individual level in both treatment and reporting ($\hat{\rho}_O = 0.704$, $\hat{\rho}_R = 0.770$). In addition, there is moderate variation among hospitals in provision of chemotherapy ($\widehat{\mathrm{SD}}_{O,chemo}$=0.488) but less for radiotherapy ($\widehat{\mathrm{SD}}_{O,radiation}$=0.295), and there is little correlation between the two ($\widehat{COR}_{O,chemo,radiation}$=0.104) at the hospital level.

Coefficient estimates from univariate models for each treatment (results not shown) are very similar to those obtained from the bivariate model. Since the univariate reporting model does not explicitly model the effect of receipt of the other treatment, the associated effect appears indirectly through other correlated predictors. In particular, in the reporting model for chemotherapy, the univariate model identifies stage II or III rectal cancer as a significant predictor ($\hat{\beta}_{R,rectal2} = 0.720$, SE=0.201; $\hat{\beta}_{R,rectal3} = 0.458$, SE=0.166). Those patients were more likely to receive radiotherapy, and when the effect of receiving radiotherapy on reporting of chemotherapy is picked up in the bivariate model, the coefficients of the stage variables become smaller and nonsignificant ($\hat{\beta}_{R,rectal2} = 0.440$, SE=0.236; $\hat{\beta}_{R,rectal3} = 0.213$, SE=0.189).

3.4 Model diagnostics

We performed posterior predictive checks (Gelman et al., 2004, Chap. 6) for the model without random reporting effects by duplicating the survey data and simulating true and registry treatment status ($\boldsymbol{Y}_O^{rep}$and$\boldsymbol{Y}_R^{rep}$) in the second copy under the model. Such diagnostics aim to check the lack of fit of the model for a subset of the data. In the first set of checks, we repeated the same hospital numbers in the copy, thus conditioning on both the general parameters and the hospital random effects; this analysis tests whether summaries of the observed data are similar to those that would be obtained if new data were drawn under the model *from the same hospitals* in the survey, and representing a full posterior predictive check. The second set of checks conditions only on the general parameters, so the random effects were drawn from their prior distributions rather than posterior distributions; this analysis tests whether summaries of the observed data are similar to those that would be predicted if *new hospitals* were included in the survey, representing a mixed predictive check (Gelman, Meng, and Stern, 1996, pg. 754). For comparison, we also applied the corresponding diagnostics to the univariate models.

We chose check statistics that summarize the marginal and joint distributions of the measures for patients and hospitals. These included the treatment and registry rates of both therapies, the sensitivities of the registry data, the odds ratios of treatment and registry reports within and between therapies, and the variances and correlations of the treatment and registry rates across the hospitals.

Table 3 (in the Web Appendix) lists the observed-data statistics, their associated 90% posterior predictive intervals (PI), and their associated one-sided posterior *p*-values, that is, $P(Q(\boldsymbol{Y}_{OS_1}^{rep}, \boldsymbol{Y}_{RS_1}^{rep}) < Q(\boldsymbol{Y}_{OS_1}, \boldsymbol{Y}_{RS1}) | \boldsymbol{Y}_{OS_1}, \boldsymbol{Y}_R)$, under the full posterior predictive check. The results obtained under the mixed predictive check are very similar to those under the full posterior predictive check and are therefore omitted here. The odds ratios are significantly underestimated from the univariate models. This is not surprising since that method does not account for the associations between the two therapies except through the common predictors. Conversely, the predictive intervals obtained from the bivariate model contain the observed odds ratios and yield reasonable posterior *p*-values. Both methods tend to overpredict the variance of chemotherapy rates and its correlation with registry rates across hospitals in the survey, that is, VAR ($\bar{Y}_{O1i.}$) and COR($\bar{Y}_{O1i.}$, $\bar{Y}_{R1i.}$), respectively. Posterior predictive checking using models with only fixed effects produced similar results but with slightly less overprediction. Despite that, both methods generally yielded satisfactory predictions for other summary statistics.

3.5 Analysis of multiply imputed data

We performed a simple analysis to illustrate use of the imputed data. As shown in Section 3.2, stage III colon cancer patients were much less likely to receive radiation therapy than rectal cancer patients, consistent with national guidelines that recommend administering this therapy to the latter population. We were interested in investigating the pattern of receipt of radiation therapy for colon cancer patients, controlling for receipt of chemotherapy, by fitting a simple patient-level logistic regression model that omits the survey region indicator, the year of treatment, hospital-level predictors, and random effects. Alternatives include the complete-case approach (fitting only the survey data), using the underreported registry data, and fitting the multiply imputed registry data under both the univariate and bivariate models .

Table 4 (in the Web Appendix) shows the results under various methods. As expected, the coefficients estimated using only the survey data have the largest standard errors because of the much smaller sample size. Conversely, estimates using uncorrected registry data have the smallest standard errors because they do not account for misclassification and hence they overstate the precision of coefficients. Among differences between the estimates from the two imputation methods, most notably, the coefficient estimate for chemotherapy is smallest ($\hat{\beta}_{chemo}$ = 1.079, SE=0.159) in the univariate model but largest ($\hat{\beta}_{chemo}$ = 2.059, SE=0.269) in the bivariate model which better incorporates the associations between the two therapies. This also changes the model fit for other predictors. For example, the univariate model identifies a negative significant association ($\hat{\beta}_{75–84}$ = –0.329, SE=0.149; $\hat{\beta}_{85+}$ = –0.781, SE=0.366) between the two oldest groups (75–84 and > 85 yrs old) and use of radiation therapy. But this effect becomes nonsignificant ($\hat{\beta}_{75–84}$ = –0.115, SE=0.148; $\hat{\beta}_{85+}$ = –0.409, SE=0.377) in the bivariate model because those patients are also less likely to receive chemotherapy (Table 2).

4. Discussion

To correct underreporting of adjuvant therapy in a cancer registry, we extended the multiple imputation approach proposed by YZ to accommodate data on multiple treatments. The extended model captures the associations and dependence among the treatments in the receipt and reporting processes, and hence generates multiply imputed data that are more appropriate for joint analyses involving several treatments. Furthermore, parameter estimates β_R from the reporting model might inform efforts to validate or improve the quality of registry data.

Other administrative systems also suffer from misclassification or misreporting of important quality indicators or indexes, such as measures of diabetes care (Keating et al., 2003) and comorbidity (Klabunde, Harlan, and Warren, 2006). The multiple imputation strategy constitutes a promising tool to tackle this general problem in health services research.

The method can be easily applied to data with more than two variables, such as surgery, chemotherapy, and radiation therapy for cancer patients. An immediate generalization is to consider mixed types of treatment and reporting variables, such as misclassified adjuvant therapy status and misreported comorbidity score which might be regarded as ordinal or continuous. The extension would link the multivariate random effects probit models together with the univariate continuous random effects model and characterize the correlations through the latent Z's.

Our method can be extended in several other directions. First, we might allow both underreporting and overreporting in the registry data. This is similar to the general misclassification problem in epidemiological research which allows both sensitivity and specificity to be less than perfect. The extension can include the same outcome model. The reporting model, however, now includes models for underreporting (where Y_{Ol}=1) and overreporting (where Y_{Ol}=0), corresponding to a second set of latent $\{Z_{Rlij}\}$.

Another extension is to combine information from more than two sources. For example, the Cancer Care Outcomes Research and Surveillance Consortium (CanCORS) (Ayanian et al., 2004) collects information about the enrolled patients from the cancer registry, patient survey, medical records, and Medicare claims. While each source contains information on different aspects of cancer care, all of them provide information on provision of adjuvant therapies and some other overlapping variables. Our imputation strategy provides a means to synthesize information from the various sources. Methods for analyses with such more general data structures are a promising area for future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank John Z. Ayanian, Recai Yucel, and the associate editor for helpful comments and Robert Wolfe for data management. This research was supported by the grants U01-CA93344 from the National Cancer Institute and HS09869 from the Agency for Healthcare Research and Quality.

REFERENCES

Ayanian JZ, Chrischilles EA, Wallace RB, Fletcher RH, Fouad MN, Kiefe CI, Harrington DP, Weeks JC, Kahn KL, Malin JL, Lipscomb J, Potosky AL, Provenzale DT, Sandler RS, Ryn MV, West DW. Understanding cancer treatment and outcomes: the Cancer Care Outcomes Research and Surveillance Consortium. Journal of Clinical Oncology 2004;22:2992–2996. [PubMed: 15284250]

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 Oncoloy 2003;21:1293–1300.

Bickell NA, Chassin MR. Determining the quality of breast cancer care: Do tumor registries measure up? Annals of Internal Medicine 2000;132:705–710. [PubMed: 10787363]

Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceau, CM. Measurement Error in Nonlinear Models: A Modern Perspective. 3rd edition. New York, NY: CRC Press; 2006.

Chib S, Greenberg E. Analysis of multivariate probit models. Biometrika 1998;85:347–361.

Clayton, DG. Models for the analysis of cohort and case-control studies with inaccurately meaured exposures. In: Dwyer, JH.; Feinlab, M.; Lippert, P., et al., editors. Statistical Models for Longitudinal Studies of Health. New York: Oxford University Press; 1992. p. 301-311.

Cress RD, Zaslavsky AM, West DW, Wolf RE, Felter MC, Ayanian JZ. Completeness of information on adjuvant chemotherapies for colorectal cancer in population-based cancer registries. Medical Care 2003;41:1006–1012. [PubMed: 12972840]

Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. Journal of Clinical Epidemiology 1992;45:613–619. [PubMed: 1607900]

Gelfand AE, Racine-Poon A, Smith AFM. Illustration of Bayesian inference in normal data models using Gibbs sampling. Journal of the American Statistical Associations 1990;85:972–985.

Gelman A. Parameterization and Bayesian modeling. Journal of the American Statistical Association 2004;99:537–545.

Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian Data Analysis. 2nd edition. New York, NY: CRC Press; 2004.

Gelman A, Meng XL, Stern HS. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). Statistical Sinica 1996;6:733–807.

Gilks WR, Best NG, Tan KKC. Adaptive rejection Metropolis sampling within Gibbs sampling. Applied Statistics 1995;44:455–472.

Hewitt, M.; Simone, JV. Ensuring Quality Cancer Care. Washington, DC: National Academy Press; 1999.

Keating NL, Landrum MB, Landon BE, Ayanian JZ, Borbas C, Guadagnoli. Measuring the quality of diabetes care using administrative data: is there bias? Health Services Research 2003;38:1529–1545. [PubMed: 14727786]

Klabunde CN, Harlan LC, Warren JL. Data sources for measuring comorbidity. Medical Care 2006;44:921–928. [PubMed: 17001263]

Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2nd edition. New York, NY: Wiley press; 2002.

Malin JL, Kahn KL, Adams J, Kwan L, Laouri M, Ganz PA. Validity of cancer registry data for measuring the quality of breast cancer care. Journal of the National Cancer Institute 2002;94:835–844. [PubMed: 12048271]

National Institute of Health. NIH Consensus Conference: Adjuvant Therapy for Patients with Colon and Rectal Cancer. Journal of the American Medical Association 1990;264:1444–1450. [PubMed: 2202842]

Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York, NY: Wiley press; 1987.

Rubin DB. Multiple imputation after 18+ years. Journal of the American Statistical Association 1996;91:473–489.

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]

Table 1

Adjuvant Therapy Rates %

Sample	Chemo	Radiation	Chemo & Radiation
Survey	73.3 (1.16)	25.4 (1.14)	23.2 (1.11)
Registry (in the survey region)	57.9 (0.79)	22.2 (0.67)	20.0 (0.64)
Registry (statewide)	51.4 (0.45)	19.6 (0.35)	17.0 (0.33)
Imputed registry (univariate models)	61.2 (0.77)	23.1 (0.61)	19.3 (0.44)
Imputed registry (bivariate model)	61.1 (0.74)	23.3 (0.64)	20.2 (0.55)

Note: Inside the parentheses are the SEs.