Published in final edited form as:
Stat Med. 2014 September 20; 33(21): 3710–3724. doi:10.1002/sim.6173.

Combining information from two data sources with misreporting and incompleteness to assess hospice-use among cancer patients: a multiple imputation approach

Yulei He[a,*,†], **Mary Beth Landrum**[b], and **Alan M. Zaslavsky**[b]

[a]Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD 20782, U.S.A

[b]Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A.

Abstract

Combining information from multiple data sources can enhance estimates of health-related measures by using one source to supply information that is lacking in another, assuming the former has accurate and complete data. However, there is little research conducted on combining methods when each source might be imperfect, for example, subject to measurement errors and/or missing data. In a multisite study of hospice-use by late-stage cancer patients, this variable was available from patients' abstracted medical records, which may be considerably underreported because of incomplete acquisition of these records. Therefore, data for Medicare-eligible patients were supplemented with their Medicare claims that contained information on hospice-use, which may also be subject to underreporting yet to a lesser degree. In addition, both sources suffered from missing data because of unit nonresponse from medical record abstraction and sample undercoverage for Medicare claims. We treat the true hospice-use status from these patients as a latent variable and propose to multiply impute it using information from both data sources, borrowing the strength from each. We characterize the complete-data model as a product of an 'outcome' model for the probability of hospice-use and a 'reporting' model for the probability of underreporting from both sources, adjusting for other covariates. Assuming the reports of hospice-use from both sources are missing at random and the underreporting are conditionally independent, we develop a Bayesian multiple imputation algorithm and conduct multiple imputation analyses of patient hospice-use in demographic and clinical subgroups. The proposed approach yields more sensible results than alternative methods in our example. Our model is also related to dual system estimation in population censuses and dual exposure assessment in epidemiology.

Keywords

data augmentation; health services research; measurement error; model diagnostics; multilevel models

[*]Correspondence to: Yulei He, Office of Research and Methodology, National Center for Health Statistics, Centers of Disease Control and Prevention, Hyattsville, MD 20782, U.S.A.
[†]wdq7@cdc.gov

1. Introduction

Combining information from multiple data sources (e.g., surveys, health claims, medical records, and registries) can enhance estimates of health-related measures by using one source to supply information that is lacking in another. For example, health services researchers frequently use health claims data to supplement information available from disease registries such as the linked Surveillance, Epidemiology, and End Results (SEER)-Medicare data. Furthermore, data sources are typically subject to nonsampling errors including missing data due to nonresponse, noncoverage, measurement and/or response errors. If different sources have different limitations, combining information for the same set of variables reported from multiple sources might alleviate these errors and produce improved estimates of these variables.

Schenker and Raghunathan [1] described several examples of combining information from multiple surveys, including the race bridging project that predicted single-race report for census data using multiple race reports from a national survey [2]. In health services research, Yucel and Zaslavsky [3] corrected underreporting of cancer patients' receipt of adjuvant chemotherapy in a statewide registry, using a validation sample from medical records data. He and Zaslavsky [4] extended this approach to multivariate outcomes. A common theme of these research is to combine information on key variable(s) from two sources, one of which is assumed to be accurate and complete (without reporting errors or missing data) for the variable(s) of interest in a validation sub-sample of the population; errors in the other source are corrected using multiple imputation approach [5].

However, there is a lack of methods for studies in which both (or multiple) sources are (all) subject to error. Our research is motivated from a multisite cohort study of care patterns for colorectal and lung cancer patients diagnosed between 2003 and 2005, the Cancer Care Outcomes Research and Surveillance Consortium (CanCORS) sponsored by National Cancer Institute [6]. Patient data were collected from surveys, medical records, Medicare claims, and cancer registries, to provide a rich set of information on a wide variety of topics and issues in the care process. The availability of multiple data sources in CanCORS also allows combining information when necessary.

National guidelines recommend that physicians discuss end-of-life (EOL) care planning with patients who have incurable cancer and a life expectancy of less than 1 year. Benefited from the uniqueness of the CanCORS data, several CanCORS studies describe patterns and quality of EOL care for patients diagnosed in late stages of lung and colorectal cancer (e.g., [7, 8]) including patients' enrollment for hospice services prior to death, a key measure assessed and validated by previous literature (e.g., [9]). For Can- CORS participants, this information was obtained from their medical records, which were abstracted at hospitals and physicians' offices within 15 months of cancer diagnosis. However, because some medical records were not abstracted because of patient nonconsent or provider noncooperation or inaccessibility, hospice-use might be underreported or missing. In addition, medical records from some physicians providing cancer care (e.g., oncologists or radiologists) may miss hospice-use because hospice enrollment involves a change of providers responsible for care.

The misreporting or missingness might also be correlated with the quality of the abstraction process. To address such concerns, Medicare enrollment and claims data for Medicare-eligible enrollees (typically of age ≥ 65 years) within a similar time frame, which is generally recognized as a reliable data source by health services researchers, were obtained as a supplement for analyses.

Table I crosstabulates hospice-use reports from the two data sources for patients who died within 15 months of diagnosis in the analytic sample. The off-diagonal counts (445 for Medicare claims YES/medical records NO and 54 for Medicare claims NO /medical records YES) show considerable inconsistency of reports from the two sources. We conjecture that both sources are subject to underreporting, but overreporting might be relatively unusual and could be neglected. Furthermore, the much larger number for the cell 'Medicare claims YES /medical records NO' than the cell 'medical records YES /Medicare claims NO' (i.e., 445 vs. 54) is consistent with *a priori* expectation that both sources might underreport, but Medicare claims might be more reliable because they are required for payment, not abstracted for specific research purposes such as in CanCORS. Finally, missing data occur for both sources: unit nonresponses from medical records because the abstraction was not implemented for all CanCORS participants and noncoverage from Medicare claims for patients under 65 years old.

A natural analytic strategy is to treat hospice-use as a missing variable (Table II) and impute it, using information from both sources. Some ad hoc imputation procedures might include the 'OR' algorithm, which assigns a 'YES' for an individual if either source reports 'YES', and the 'AND' algorithm, which assigns a 'YES' only if both sources are 'YES'. These procedures, however, lack rigorous statistical justifications and offer no method for imputing missing reports. They also ignore the possible associations between hospice-use and other covariates in the study.

In this paper, we aim to develop a more principled imputation approach. However, missing data methods that handle partially classified contingency tables in the form of Table I (Chapter 13 of [10]) assume no misreports from the two sources. More related research on combining information from two sources assume that one of them can be treated as a gold standard, while the other might be subject to misreporting or missing data [1, 3, 4]. Here, we extend previous research to account for misreports and missing data in both sources. In Section 2, we introduce the notation and modeling strategy. Section 3 presents the analysis of CanCORS data. Section 4 points out connections with related methods and discusses future research topics.

2. Method

Let Y_O be true hospice-use status (1, yes/0, no) and Y_{R1} and Y_{R2} be its reports from medical records and Medicare claims, respectively. As shown in Table II, Y_O is a latent variable (100% missing), and missing values (due to nonresponses for the data sources) can also occur for $Y_{Rl} = (Y_{Rl,obs}, Y_{Rl,mis})$, ($l = 1, 2$). We assume that the mechanisms leading to misreporting and missing responses from each source can be related to some covariates X. For simplicity, we assume that X has no measurement errors and is fully observed. We also treat the linked cases from both sources as a simple random sample from the combined population.

We further assume that missing Y_{Rl}'s are at random (MAR), meaning that the probability of missing reports can only depend on X but not Y_O. Let θ indicate some parameters governing the process of misreporting and nonresponse. In our context, around 90% of the missing Medicare claims data are due to noncoverage from patients younger than 65 years old. This is largely 'missing by design' that satisfies MAR if we include age as a predictor X in the imputation model for hospice-use. On the other hand, the missing cases from the medical records are mainly caused by subject nonconsent and inaccessibility of some records. The rest of missing data resulted from nonmatches in the data linking process (Section 3.1). If the probability of unit nonresponses in medical records and nonmatches in Medicare claims are largely associated with demographic and clinical variables, which are included in X in the modeling process, then the MAR assumption is more plausible.

Under MAR, valid inferences about θ can be made using the observed-data likelihood $P(Y_{R1,obs}, Y_{R2,obs}|X, \theta) = \int P(Y_{R1}, Y_{R2}, Y_O|X, \theta)dY_{R1,mis}dY_{R2,mis}dY_O$, where $P(Y_{R1}, Y_{R2}, Y_O|X, \theta)$ is the complete-data model.

We further consider the following decomposition of the complete-data model

$$P(Y_{R_1}, Y_{R_2}, Y_O|X, \theta) = P_O(Y_O|X, \theta_O)P_R(Y_{R_1}, Y_{R_2}|Y_O, X, \theta_R)$$

Following [3], we refer to $P_O(Y_O|X, \theta_O)$ as the 'outcome' model. It relates hospice-use to covariate X, with regression parameters θ_O that might be of subject-matter interest.

The reporting model (or the measurement error model), $P_R(Y_{R1}, Y_{R2}|Y_O, X, \theta_R)$, characterizes reporting in the two sources given true hospice-use status, covariates, and parameters θ_R. Our reporting model rests on two assumptions:

Assumption 1

Reporting in the two sources is independent conditional on true status and observed covariates:

$$P_R(Y_{R1}, Y_{R2}|Y_O, X, \theta_R) = P_{R1}(Y_{R1}|Y_O, X, \theta_{R1})P_{R2}(Y_{R2}|Y_O, X, \theta_{R2}).$$

Assumption 2

Both sources may be subject to underreporting but not overreporting: $P_{Rl}(Y_{Rl} = 1|Y_O = 0, X, \theta_{Rl}) = 0$.

Assumption 1 is justified if X contains all factors that are predictive of misreporting. However, with enough scientifically relevant covariates, the residual correlation between two reporting systems might be minimal. Section 1 presents arguments for plausibility of Assumption 2 in our application. Generalizations of these assumptions are considered in Section 4.

We propose to multiply impute Y_O to facilitate statistical analyses involving the hospice-use variable. Although originally proposed as a tool that can be used by statistical agencies to handle nonresponse in large-sample public-use household surveys, the multiple imputation framework has been adapted for other statistical contexts over the past 30 years [11]. Relevant examples include latent variables [12] and measurement error problems [13]. See Section 3.3. for more details related to our example.

3. Application

3.1. Data background

The CanCORS Consortium is a collaboration of seven teams of investigators across the nation to evaluate prospectively the quality of cancer care for patients with lung or colorectal cancer. Approximately 10,000 patients newly diagnosed during 2003–2005 were identified from five geographic sites (Northern California, Los Angeles County, central and eastern North Carolina, Iowa, and Alabama), five large health management organizations (Group Health Cooperative, Harvard Pilgrim Healthcare, Henry Ford Health System, Kaiser Permanente Health Insurance, and Kaiser Permanente North West), and 10 Veterans Health Administration hospitals, capturing variation in cancer care across geographical regions and health care systems.

The CanCORS investigators abstracted detailed clinical data from medical records, including cancer-related diagnostic and staging procedures, surgery, chemotherapy, radiation therapy, and EOL care measures [6]. Additional predictors of clinical outcomes included tumor stage, comorbid illnesses, and relevant test results. Medical records for some patients were unavailable because of nonconsent or inaccessibility. As an additional data source, Medicare enrollment and claims data for Medicare enrollees among CanCORS participants (excluding Veterans Health Administration patients) were linked for 89% of CanCORS enrollees aged 65 years and older within the study period. We used inpatient, outpatient, provider, and hospice files to code the EOL care quality indicators including hospice-use.

The analytic sample for the pattern of hospice-use includes 3027 CanCORS patients who died within 15 months of diagnosis (Table I). Among them, 358 patients missed hospice reports from both sources. Under the MAR assumption for reported data, these cases contain no information on the parameters of the outcome and reporting models proposed in Section 2. Thus, the analysis focuses on the remaining 2669 cases with observed data from at least one source. Table III describes patient demographic and clinical characteristics, which might have been associated with the use of EOL care among terminally ill patients as suggested by previous literature [9]. These variables constitute the covariates X in both the outcome and reporting models. They were fully observed and assumed to be without measurement errors. Note that in the reporting model for Medicare claims $P_{R2}(Y_{R2}|Y_O, X, \theta_{R2})$, there exists no less-than-65 age group due to noncoverage, and therefore we could only use the remaining age groups in the covariates.

3.2. Exploratory analyses

We conduct some exploratory analyses initially to establish the basis for model-based imputation. Under our assumptions, hospice-use status might be 1 even if reports from both sources are 0. This phenomenon can be connected with the dual system estimation (DSE) [14] of population counts and other capture–recapture problems. The classical DSE approach conceptualizes that each person in a population is either in or not in the two *lists*. Analogously in our context, a patient's hospice-use status is either reported or not in the two data sources, medical records, and Medicare claims data.

We first treat Table I as a simple DSE example, ignoring the cells containing missing values. Note that the cell with 'NO' from both sources (617) can be viewed as the count of cases excluded from both lists in the DSE framework. The reporting completeness (inclusion probability or capturing rate) for data source 1 (2) can be simply calculated as $E(Y_{R1}|Y_{R2} = 1)$ ($E(Y_{R2}|Y_{R1} = 1)$). For example in Table I, note that the 449(= 395+54) cases who were reported as 'YES' from the medical records data are true positives. Yet only 395 of them were reported as 'YES' from the Medicare claims data. This suggests that the reporting completeness rate for the Medicare claims data is 395/(395 + 54) ≈ 88%. Similarly, the capturing rate for medical records is estimated as 395/(395 + 445) ≈ 44%. In addition, the total number of 'YES' on hospice-use in the sample is estimated as (395+54)(395+445)/395 ≈ 955. That is, among the 617 cases with 'NO' reports from both sources, 445 × 54/395 ≈ 61 cases should be imputed as 'YES' in hospice-use. If we further consider cases with missing reports from either source, an expectation–maximization (EM) algorithm (details not shown) can be developed to fit a model with three parameters including $P(Y_O = 1)$, $P(Y_{R1} = 1|Y_O = 1)$, and $P(Y_{R2} = 1|Y_O = 1)$, the simplest form of Model (1) (Section 3.3.1).

The classic DSE assumes independent reporting between two data sources, homogenous across individuals. A nonhomogeneous population may be divided into several homogeneous post-strata based on demographical variables, or the heterogeneous inclusion probability might be related to covariates using logistic regression models [15]. Our modeling approach (Section 2) is consistent with the latter strategy. However, because some sites have perfect reporting completeness or sparse data, directly including study sites as a fixed-effect covariate for the reporting models would entail the data separation issue: a linear combination of the predictors perfectly predicts the binary outcome, making estimates for the associated parameters unidentifiable [16]. Table AI (Appendix) describes hospice reports stratified by site, showing 100% sample reporting completeness for Medicare claims data from sites HPHC, KPHI, UAB, and UNC and for medical records from site HPHC. On the other hand, removing site as a predictor is not desirable because the heterogeneity of reporting completeness across sites might not be fully attributable to other patient-level characteristics. Use of hospice (the outcome) may also vary across sites, possibly revealing interesting geographical and organizational variations among health care providers. We instead treat site as a random effects covariate, making site effects estimable. At sites with a perfect unadjusted reporting rate, the model-based estimate would be shrunken toward the population average adjusted for other covariates. In addition, the random effects model treats the included sites as a random sample from a population of potential data collection sites, making the inferences more generalizable.

3.3. Model-based imputation

3.3.1. Model specification—We consider Bayesian random effects probit models [17] for both the outcome and reporting processes. Let *i* index site *i* = 1,…, 10 and *j* = 1,…n_i index patients. The complete-data model (1) is

$$\Phi^{-1}(Pr(Y_{Oij}=1|X_{ij}))=Z_{Oij}=X_{ij}^{T}\beta_{O}+\gamma_{Oi} \quad (1.1)$$

$$Y_{R1ij}=Y_{R2ij}=0 \text{if} Y_{Oij}=0 \quad (1.2)$$

$$\Phi^{-1}(Pr(Y_{R1ij}=1|Y_{Oij}=1,X))=Z_{R1ij}=X_{ij}^{T}\beta_{R1}+\gamma_{R1i} \quad (1.3)$$

$$\Phi^{-1}(Pr(Y_{R2ij}=1|Y_{Oij}=1,X))=Z_{R2ij}=X_{ij}^{T}\beta_{R2}+\gamma_{R2i} \quad (1.4)$$

where Z_{Oij}, Z_{R1ij}, and Z_{R2ij} are the normal latent variables, β_O, β_{R1}, and β_{R2} are the fixed-effects parameters, γ_{Oi}, γ_{R1i}, and γ_{R2i} are the random site effects for the outcome and reporting processes, respectively with $\gamma_{Oi}\sim N(0,\sigma_{O}^{2})$, $\gamma_{R1i}\sim N(0,\sigma_{R1}^{2})$, and $\gamma_{R2i}\sim N(0,\sigma_{R2}^{2})$ independently across sites. Following our reasoning in Section 2, Equation (1.1) specifies the outcome model, Equation (1.2) imposes the underreporting assumption, and Equations (1.3) and (1.4) specify the reporting model for the medical records data and Medicare claims data, respectively.

We consider vague priors for parameters given that little is known about the mechanism governing the outcome and reporting processes. We impose flat priors for β's ($p(\beta) \propto 1$). For the between-site variance $\sigma_{O}^{2}, \sigma_{R1}^{2}$ and σ_{R2}^{2}, we lacked solid information and field expertise on how study sites might vary on the hospice-use and reporting completeness and therefore used noninformative or weakly informative priors [18]. Despite the popular use of inverse gamma priors ($IG(a, a)$) for σ^2 because of its conjugacy [19], recent literature (e.g., [18, 20]) has shown the sensitivity of the corresponding posterior inferences to the choices of *a*, especially with a small number of clusters. We confirmed this in our own application: under inverse gamma priors with *a* = 0.001, 0.01, 0.1, 1, the posterior estimate of σ^2 increases considerably as a increases. Gelman [18] recommended using a noninformative uniform prior density for σ, that is, $p(\sigma) \propto 1$, which also implies $p(\sigma^2) \propto (\sigma^2)^{1/2}$. If the number of clusters is small, say below five, a weakly informative half-*t* prior can be used to prevent from obtaining extremely large posterior estimate. Note that our data had 10 sites, offering a reasonable number of clusters for applying the uniform prior. Table V lists posterior estimates of $\sigma_{O}^{2}, \sigma_{R1}^{2}$, and σ_{R2}^{2}. We also implemented the half−*t* priors, and results were similar.

3.3.2. Imputation algorithm—We implemented a data augmentation (DA) algorithm [21] to draw model parameters and impute missing values. The main steps are sketched here, and details on the conditional distributions appear in the Appendix. Some sample R (http://www.r-project.org) code is also attached.

a. Draw latent variables (Z_{Oij}, Z_{R1ij}, Z_{R2ij}) of Model (1) from truncated normal distributions.

b. Draw fixed-effects coefficients β_O and β_R from multivariate normal distributions.

c. Draw random effects $\{\gamma_{Oi}, \gamma_{Ri}\}$ from independent normal distributions for each i.

d. Draw missing values of Y_O from multiple Bernoulli distributions where the probabilities are estimated from the functions of cumulative distributions of the standard normal distribution. Note that $\int P_{Rl}(Y_{Rl}|Y_O = 1, X, \theta_{Rl})dY_{Rl,mis} = \Pi P(Y_{Rl,obs}|Y_O = 1, X, \theta_{Rl})$, $l = 1, 2$. The latter suggests that conditional on $Y_O = 1$ and MAR for missing Y_{Rl}, the likelihood inference on θ_{Rl} does not involve the missing reports. Therefore, it is unnecessary to impute missing Y_{Rl} in the algorithm.

We diagnosed the convergence of the DA algorithm using statistics developed by Gelman and Rubin [22] and concluded that the Gibbs chain converged after 10^5 iterations. The posterior inferences and multiple imputation analyses are based on running the chain for another 10^5 iterations.

3.3.3. Model diagnostics—

We used the posterior predictive checking method [23] for the model diagnostics, comparing the observed reports $Y_{Rl,obs}$ with distributions of their replicates $Y^{rep}_{Rl,obs}$ under Model (1). If Q is a diagnostic statistic (or the discrepancy function as defined in [23]), then the posterior predictive p-value for assessing the model fit in terms of Q is calculated as $P\left(Q(Y_{Rl,obs}, X) > Q\left(Y^{rep}_{Rl,obs}, X\right)\right)$. An extreme p-value (close to 0 or 1) would imply some model misfit.

A natural choice of Q is the average of reports, $Q(Y_{Rl}) = \bar{Y}_{Rl}$, the mean rate of hospice-use estimated from each data source or within certain strata defined by covariates (e.g., for all female patients or patients in certain study sites). These diagnostics (not shown) do not reveal any apparent misfit. We also check the model fit for the joint distribution of Y_{R1} and Y_{R2}. Going back to the DSE framework (Section 3.2), we use some unadjusted estimates of capturing rate for both sources as the discrepancy function (i.e., $Q(Y_{R1}, Y_{R2}) = E(Y_{R1}|Y_{R2} = 1)$ and $E(Y_{R2}|Y_{R1} = 1)$). Here, the term ‘unadjusted’ merely means that they are assessed in a straightforward way as opposed to fitting Model (1) with a extensive set of covariates. These rates can be calculated across the whole sample (e.g., Section 3.2) or within covariate groups (e.g., for each site). In addition, we used the sample variance of these capturing rates as a discrepancy measure for heterogeneity.

Table IV shows the observed capturing rates, medians from their posterior predictions, and posterior predictive p-values for each study site. For example in the GHC site, the capturing rate estimate is $23/(23 + 38) \approx .38$ for Medical records and $23/(23 + 2) \approx .92$ for Medicare claims. The corresponding numerators and denominators can be found in Table 7. The posterior predictive p-values are mostly between 0.1 and 0.9, showing a good fit for the model. Correspondingly, the (medians of) predictions are in general not far different from the observed statistics. For the sites with perfect observed capturing rates (e.g., UAB and UNC from Medicare claims data), their predictions tend to be slightly lower, as expected from shrinkage-to-the-mean effect under the random site effects model. The sample variance of the site-specific capturing rates (shown for the line 'Var' in Table IV) is also well predicted by the model.

3.4. Analytic results

3.4.1. Model estimates—Table V shows the parameter estimates (posterior medians) of β_O and β_R from the outcome and reporting models (Equations (1.1)–(1.4)). Estimates with 95% CIs (not shown) excluding 0 are highlighted with *. Results from the outcome model (β_O from Equation (1.1)) suggest that hospice was used less often by CRC patients than lung cancer patients, more often in the older groups (65–69 years old or >80 years old), more often among late (stage 3 or 4) or missing stage patients, and more often among patients with depression symptoms. Patients who lived longer after diagnosis were more likely to use hospice services. Results from reporting models (β_{R1} and β_{R2} from Equations (1.3) and (1.4)) suggest that medical record data were less complete for hospice-use among patients aged between 65 and 69 years and much less complete among patients with stage missing information. Medicare claims data were less complete for hospice-use among Hispanic patients and more complete among patients with heart failure or diabetes or those who lived longer.

As can be seen from the 'Site random effects variance' in Table V, there was considerable between-site variation for hospice-use and reporting from medical records data but less so for the reporting from Medicare claims. Under Model (1), the posterior median rates for hospice-use range from around 43% to 100% across sites. Those for the reporting completeness range from around 21% to 78% for medical records and from around 77% to 100% for Medicare claims. The much smaller between-site heterogeneity for coverage of Medicare claims than medical records abstraction is well predictable given that the former is collected by the nationally uniform system that pays for the services, while the latter is collected by separate staff groups within each CanCORS site.

The posterior median number of imputed 'YES' case when both sources have negative reports is around 88, substantially more than the estimate (61, Section 3.2) from the simple DSE approach. In addition, the corresponding posterior medians when the reports from medical records (Medicare claims) are negative and Medicare claims (medical records) are missing is around 272 (21). Neither quantity, however, can be estimated using the simple DSE approach that does not account for the missing reports. Therefore, methods that do not capture the heterogeneous reporting probabilities associated with covariates or do not account for missing reports might underestimate the number of false negatives. On the other hand, the posterior medians of the average reporting completeness rate for the medical records and Medicare claims are 46.8% and 85.6%, respectively. They are close to the simple DSE estimates (44.0% and 88.0%). Therefore, simple and model-based approaches agree on the much lower the reporting completeness for medical records than for Medicare claims.

Although the primary analytic goal is to impute the hospice-use status for substantive analyses (Section 3.4.2), the model outputs in Table V might inform us about certain factors correlated with misreporting. Therefore, targeted efforts might be planned in the process of data collection and dissemination to improve the accuracy and reliability of large data systems (e.g., reducing the variation across multiple sites with good coordination).

3.4.2. Post-imputation analyses—We can use the multiply imputed data for substantive analyses involving the hospice-use variable. Because of the 100% missingess rate, we use 50 imputed datasets, selected from every 2000 iterations of the Gibbs chain, to minimize potential autocorrelations and ensure a high relative efficiency [5] for multiple imputation inferences. Model coefficients are estimated from each of the 50 imputed datasets and combined using rules for multiple imputation inference for scalar estimands [5]. Thus, point estimates are means of the estimates from the 50 imputations, and the standard errors combine between-imputation variance because of missing data with sampling (within-imputation) variability conditional on imputed data.

We consider two analyses. One is to describe the hospice-use rate for the analytic sample. The other is a logistic regression model predicting hospice-use for lung cancer patients. These patients are more likely to be diagnosed at late stage and to die more quickly after diagnosis. Therefore, hospice-use might be a more relevant issue for them and their family members. We exclude from the predictors the time from diagnosis till death and study sites to make the inferences more predictive for patients seeking EOL care and therefore more generalizable.

For comparative purposes, we consider several alternative missing data strategies. The ‘OR’ algorithm assigns true status ‘YES’ if either source reports ‘YES’, ‘NO’ if both sources report ‘NO’, and ‘MISSING’ if one source reports ‘NO’ and the other reports ‘MISSING’. The ‘AND’ algorithm assigns true status ‘YES’ if both sources report ‘YES’, ‘NO’ if neither source reports ‘MISSING’ and at least one source reports ‘NO’, and ‘MISSING’ if at least one source reports ‘MISSING’.We also analyze the data using either source alone.

Table VI shows estimates from the two aforementioned post-imputation analyses. The first row lists marginal estimates of the hospice-use rate, and other rows list the logistic regression coefficients. Except for the analysis using ‘OR’ algorithm, all other methods yield substantially lower estimates for the rate of hospice-use than that from multiple imputation. This is because these methods treat reported ‘NO’ as true negatives and therefore tend to underestimate the rate under the assumption of independent underreporting. The rate estimate from the ‘OR’ algorithm is higher because it removes cases with missing reports from either source and therefore deflates the denominator. The logistic regression analysis results from the multiply imputed data suggest that older patients (60–65 or 80 years or older), patients in late stage (III/IV) or with stages missing, and patients with depression are more likely to use hospice. The directions and magnitudes of these coefficients from other missing data methods can be considerably different. For example, the analysis using the ‘OR’ algorithm shows patients older than 65 years old are less likely to use hospice than those younger than 65 years old. This less obvious result might be caused by solely using the misreported medical records data for patients younger than 65 years old. In addition, as opposed to the multiple imputation approach, the other missing data methods rarely show strong associations between late stage and hospice-use, which is less plausible because we might expect more use of hospice by patients with more predictable mortality.

Because the alternative methods do not appropriately account for measurement errors in the response variable, the resulting coefficients are susceptible to bias [24], which might lead to misleading scientific conclusions. For example, using medical records only, which contain substantial measurement errors, might be less capable of detecting true effects. On the other hand, using Medicare claims only, which might be more accurate yet are limited to the sample older than 65 years old, might produce association estimates that cannot be generalizable to younger patients. The multiple imputation analysis overcomes the weakness of the two data sources. It yields overall larger standard errors than the alternative approaches that overstate the precision of the coefficients because they fail to account for the misclassification of the outcome (hospice-use) in an appropriate way.

In multiple imputation analysis, an important statistic is the rate of missing information, which is approximately the ratio of between-imputation to total variance and shows the increase of variance associated with missing data under the model. In the logistic regression analysis, the estimated rate of missing information is substantial, ranging largely between 40% and 70% for various estimates (results not shown). But they are still considerably lower than the rate of imputation (100%). This range suggests that using mismeasured reports and covariates is as informative as having observed around 30–60% (=(1–70%, 1–40%)) of the true hospice-use data for the analyses, supporting the utility of Model (1) for imputation [25].

4. Discussion

In our application, a binary measure of hospice-use is available from two data sources, yet it is subject to misclassification and missing data in both datasets, errors commonly encountered in population research. We multiply impute the true status of this measure, using models that capture both the outcome and reporting processes (Equations (1.1)–(1.4)). The subjective-matter analyses yield conclusions more sensible than those of ad hoc approaches that fail to account for these errors. Although Medicare claims data are shown to be more accurate than medical records in our case, the former is limited to patients older than 65 years old. Therefore, a major advantage of the proposed approach is to analyze hospice-use across the entire age spectrum in a more reliable way.

Some major limitations of the proposed approach stem from the assumptions (Section 2.1), including MAR for the reported data, conditional independence of reporting, and underreporting for both sources. We discuss possible generalizations that can lead to future research topics. First, we might generalize the missingness mechanism to not MAR [5] by assuming that, for example, the medical records for patients who did not use hospice were more likely to be missing. However, not MAR models are usually weakly identifiable [26] and are especially likely to be so in our application because true hospice-use is 100% missing. A more feasible strategy might be to perform sensitivity analysis under possible specifications of models for missingness mechanisms that rely on strong subject-matter expertise on how medical records data might be missing during the abstracting process.

Our topic is related to dual exposure assessment in the measurement error problems literature [27]. There, two assessments are used to determine whether subjects were exposed to putative binary risk factors in case-control studies. For example, Drews *et al*. [28] used patient interviews and medical records to provide two assessments of a variety of exposures (e.g., maternal anemia during pregnancy) for a case-control study on sudden infant death syndrome. The objective is to estimate exposure-outcome associations (Section 5.3–5.4 of [29]). In our context, the analog to exposure is hospice-use status, reported by two data sources. Unlike typical case-control studies, there is no analog to the disease outcome measure in our setting, although hospice-use could be either the exposure (predictor) or outcome (response) in post-imputation analyses. Our model factorization corresponds to the 'exposure given outcome model' in Equation (5.14) of [29].

Gustafson's [29] variations on dual exposure assessment models include relaxing the conditional independence assumption and allowing both overreporting and underreporting. For example, we might consider a more general specification of the reporting model $P_R(Y_{R1}, Y_{R2}|Y_O, X, \theta_R)$ relaxing both assumptions, with complete-data model

$$g(Pr(Y_O=1|X))=X^T\beta_0 \; g(Pr(Y_{R1}=1|Y_O, X))=\beta_1 Y_O+X^T\beta_2 \; g(Pr(Y_{R2}=1|Y_O, Y_{R1}, X))=\beta_3 Y_O+\beta_4 Y_{R1}+X^T\beta_5 \quad (2)$$

where g is a link function for the binary outcome Y_O, X is the vector of covariates including the intercept term, and β's are the coefficients. Thus, the misreporting of two data sources can go either direction, and they can still be correlated after controlling for covariates.

However, Model (2) might be weakly identified. Although with an increasing number of covariates X included in the data, there seem to be enough degrees of freedom for estimating the associated parameters, practical experience with such estimation has been largely negative (Section 5.4.2 of [29]). See also comments from Section 15.3.2.1 of [27] in a similar context with unknown response misclassifications in logistic regression models. They argued that the practical nonidentifiability of these models might be due to the fact that the misclassification probabilities are only weakly identified by the data if both overreporting and underreporting are allowed. Our own application of Model (2) to the hospice-use data show similar lack of identifiability (results not shown). Further methodological investigations on the issues of model nonidentifiability is beyond the scope of this paper.

On the other hand, identifiability of more general misreporting models might be improved if data are available from more than two sources. With $L > 2$ sources, the complete-data model can be factorized as

$$P(Y_{R1}, Y_{R2}, \ldots, Y_{RL}, Y_O|X, \theta)=P_O(Y_O|X, \theta_O)P_R(Y_{R1,} Y_{R2}, \ldots, Y_{RL}|Y_O, X, \theta_R)$$

The main challenge is to model the multivariate reported outcomes $P_R(Y_{R1}, Y_{R2}, \ldots, Y_{RL}|Y_O, X, \theta_R)$ (i.e., the reporting model). With L increasing, the data might contain more flexibility for allowing both overreporting and underreporting, as well as residual correlations among reporting sources after adjusting for covariates. Intuitively, binary data Y_{Rl}'s constitute a 2^L table, which opens more degrees of freedom for model parameters. With $L = 3$, the modeling framework resembles triple system estimation, which augments census and post-enumeration survey data (both of which are included in DSE) with administrative-list data to improve estimates of population count [30]. Similarly, CanCORS patients or their surrogates were also surveyed on whether they received hospice care. Patient survey data can be treated as the third source for this measure, which might include both overreporting and underreporting due to recall bias. Further, modeling efforts to impute hospice-use will be necessary in such a complicated data structure. Extensions can also be considered for mismeasured variables with distributions other than binary (e.g., ordinal, nominal, or continuous).

Increased efforts on data collection and assembly offer great opportunities to use information from multiple data sources for scientific investigations. The imputation strategy provides an effective means to synthesize information, closely related to advances in survey sampling and measurement error problems. Methods under more general modeling assumptions or for more complicated data structures deserve further research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The findings and conclusions in this study are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention. The authors acknowledged the journal editors and reviewers, Meena Khare and Jennifer Madans, for their valuable input. The work of the CanCORS Consortium was supported by grants from the National Cancer Institute (NCI) to the Statistical Coordinating Center (U01 CA093344) and the NCI-supported Primary Data Collection and Research Centers (Dana-Farber Cancer Institute/Cancer Research Network U01 CA093332, Harvard Medical School/Northern California Cancer Center U01 CA093324, RAND/UCLA U01 CA093348, University of Alabama at Birmingham U01 CA093329, University of Iowa U01 CA093339, and University of North Carolina U01 CA093326) and by the Department of Veterans Affairs grant to the Durham VA Medical Center CRS 02-164.

Appendix

Let $\theta=(\beta_O, \{\gamma_{Oi}\}, \sigma_O^2, \beta_{R1}, \{\gamma_{R1i}\}, \sigma_{R1}^2, \beta_{R2}, \{\gamma_{R2i}\}, \sigma_{R2}^2)$. The imputation step of the DA algorithm:

- Set $Y_O = 1$ if $Y_{R1} = 1$ or $Y_{R2} = 1$ by the assumption of underreporting
- When $Y_{R1} = 0$ and $Y_{R2} = 0$, by the Bayes theorem,

$$\begin{aligned}
&P(Y_O=1|Y_{R1}=0, Y_{R2}=0, X, \theta)\\
&=\frac{P(Y_{R1}=0|Y_O=1, X, \beta_{R1}, \{\gamma_{R1i}\})P(Y_{R2}=0|Y_O=1, X, \beta_{R2}, \{\gamma_{R2i}\})P(Y_O=1|X, \beta_O, \{\gamma_{Oi}\})}{P(Y_{R1}=0|Y_O=1, X, \beta_{R1}, \{\gamma_{R1i}\})P(Y_{R2}=0|Y_O=1, X, \beta_{R2}, \{\gamma_{R2i}\})P(Y_O=1|X, \beta_O, \{\gamma_{Oi}\})+P(Y_O=0|X, \beta_O, \{\gamma_{Oi}\})}\\
&=\frac{\Phi(-X^T\beta_{R1}-\gamma_{R1})\Phi(-X^T\beta_{R2}-\gamma_{R2})\Phi(X^T\beta_O+\gamma_O)}{\Phi(-X^T\beta_{R1}-\gamma_{R1})\Phi(-X^T\beta_{R2}-\gamma_{R2})\Phi(X^T\beta_O+\gamma_O)+\Phi(-X^T\beta_O-\gamma_O)}
\end{aligned}$$

draw Y_O from a Bernoulli distribution with this probability.

- When $Y_{R1} = 0$ and Y_{R2} is missing, draw Y_O from a Bernoulli distribution with the following probability

$$
\begin{aligned}
&P(Y_O=1|Y_{R1}=0, X, \theta)\\
&\quad=\frac{P(Y_{R1}=0|Y_O=1, X, \beta_{R1}, \{\gamma_{R1i}\})P(Y_O=1|X, \beta_O, \{\gamma_{Oi}\})}{P(Y_{R1}=0|Y_O=1, X, \beta_{R1}, \{\gamma_{R1i}\})P(Y_O=1|X, \beta_O, \{\gamma_{Oi}\})+P(Y_O=0|X, \beta_O, \{\gamma_{Oi}\})}\\
&\quad=\frac{\Phi(-X^T\beta_{R1}-\gamma_{R1})\Phi(X^T\beta_O+\gamma_O)}{\Phi(-X^T\beta_{R1}-\gamma_{R1})\Phi(X^T\beta_O+\gamma_O)+\Phi(-X^T\beta_O-\gamma_O)}
\end{aligned}
$$

- Similarly when Y_{R1} is missing and Y_{R2} is 0, draw Y_O from a Bernoulli distribution with the following probability

$$
P(Y_O=1|Y_{R2}=0, X, \theta)=\frac{\Phi(-X^T\beta_{R2}-\gamma_{R2})\Phi(X^T\beta_O+\gamma_O)}{\Phi(-X^T\beta_{R2}-\gamma_{R2})\Phi(X^T\beta_O+\gamma_O)+\Phi(-X^T\beta_O-\gamma_O)}
$$

The posterior step of the DA algorithm draws a new value of θ conditional on imputed Y_O via the auxiliary variable Gibbs sampling algorithm for probit models [31]. To describe the algorithm in a general way, suppose data contains m sites and, in site i, there are n_i patients. The observed data consist of $\{Y_{R1ij}\}$ and $\{Y_{R2ij}\}$, and patient-level covariates X_{ij}. Unless denoted specifically (e.g., for σ_O^2), the posterior distributions are standard, and we omit the details. The Gibbs sampler is as follows:

Step 1. Draw latent variables $\{Z_{Oij}\}$ from truncated univariate standard normal distributions with mean $X_{ij}\beta_O + \gamma_{Oi}$ with the signs of latents depending on Y_{Oij}, that is, $Z_{Oij} > 0$ iff $Y_{Oij} = 1$.

Step 2. For the cases with $Y_{Oij} = 1$, draw latent variables $\{Z_{Rlij}\}$ from truncated univariate standard normal distributions with mean $X_{ij}\beta_{Rl} + \gamma_{Rli}$ with the signs of latents depending on Y_{Rlij}, that is, $Z_{Rlij} > 0$ iff $Y_{Rlij} = 1$ ($l = 1, 2$).

Step 3. Draw

$$
\beta_O \sim N\left(\left(\sum_{i,j} X_{ij}^T X_{ij}\right)^{-1}\left(\sum_{i,j} X_{ij}^T (Z_{Oij}-\gamma_{Oi})\right), \left(\sum_{i,j} X_{ij}^T X_{ij}\right)^{-1}\right).
$$

Step 4. Draw

$$
\beta_{Rl} \sim N\left(\left(\sum_{i,j,Y_{Oij}=1} X_{ij}^T X_{ij}\right)^{-1}\left(\sum_{i,j,Y_{Oij}=1} X_{ij}^T (Z_{Rlij}-\gamma_{(Rli)})\right), \left(\sum_{i,j,Y_{Oij}=1} X_{ij}^T X_{ij}\right)^{-1}\right).
$$

Step 5. Let S be the design matrix for random effects γ_{Oi}. Draw random effects $\{\gamma_{Oi}\}$ from a multivariate normal with covariance matrix

$\Omega_{\gamma_O} = (diag\{n_i\} + 1/\sigma_O^2 I_m)^{-1}$ and mean vector $\mu_{\gamma O} = \Omega_{\gamma O} (S^T (Z_O - X\beta_O))$, where I_m is the identity matrix with dimension m and $diag\{n_i\}$ is a $m \times m$ diagonal matrix with the i-th element as n_i.

Step 5. Let $S_{Y_O} = 1$ be the submatrix of S for the cases with $Y_O = 1$. Draw random effects $\gamma_{(RI)}$ from a multivariate normal with covariance matrix

Table AI

Frequency tables of hospice-use reports from medical records and medicare claims, by study sites.

	GHC	
	Medicare claims yes	Medicare claims no
Medical records yes	23	2
Medical records no	38	22
	HPHC	
	Medicare claims yes	Medicare claims no
Medical records yes	2	0
Medical records no	0	0
	HFHS	
	Medicare claims yes	Medicare claims no
Medical records yes	7	3
Medical records no	2	11
	KPHI	
	Medicare claims yes	Medicare claims no
Medical records yes	1	0
Medical records no	4	8
	KPNW	
	Medicare claims yes	Medicare claims no
Medical records yes	41	7
Medical records no	20	23
	NCCC	
	Medicare claims yes	Medicare claims no
Medical records yes	43	12
Medical records no	130	179
	UAB	
	Medicare claims yes	Medicare claims no
Medical records yes	20	0
Medical records no	82	101
	UCLA	
	Medicare claims yes	Medicare claims no
Medical records yes	62	9
Medical records no	78	122
	IOWA	
	Medicare claims yes	Medicare claims no
Medical records yes	190	21
Medical records no	70	112
	UNC	
	Medicare claims yes	Medicare claims no

Medical records yes	6	0
Medical records no	21	39

Note: The subsample consists of 2669 CanCORS patients who died within 15 months of diagnosis and had at least one observed hospice-use record from two sources. An example for calculating unadjusted site-specific capturing rate for two data sources: For GHC, the estimate of Medicare Claims is 23/(23 + 2) ≈ 92%; the estimate of medical records is 23/(23 + 38) ≈ 38% (see also the footnote of Table IV).

$\Omega_{\gamma_{Rl}} = \left(S^T_{Y_O=1} S_{Y_O=1} + 1/\sigma^2_{Rl} I_m\right)^{-1}$ and mean vector $\mu_{\gamma_O} = \Omega_{\gamma_O}\left(S^T_{Y_O=1}(Z_{Rl} - X_{Y_O=1}\beta_{Rl})\right)$.

Step 6. Draw σ^2_O from $IG((m-1)/2, \sum_i \gamma^T_{Oi}\gamma_{Oi}/2)$. Note that the conditional distribution of $\sigma^2_O, p(\sigma^2_O|\text{Others}) \propto p(\{\gamma_{Oi}\}|\sigma^2_O)p(\sigma^2_O)$. As in [18], a uniform prior on σ_O implies that $p(\sigma^2_O) \sim (\sigma^2_O)^{-\frac{1}{2}}$. Therefore,

$$p(\sigma^2_O|\text{Others}) \propto (\sigma^2_O)^{-\frac{m}{2}} exp\left(-\frac{1}{2\sigma^2_O}\sum_i \gamma^T_{Oi}\gamma_{Oi}\right)(\sigma^2_O)^{-\frac{1}{2}} = (\sigma^2_O)^{-\frac{m+1}{2}} exp\left(-\frac{1}{2\sigma^2_O}\sum_i \gamma^T_{Oi}\gamma_{Oi}\right)$$

. On the other hand, for a random variable $X \sim IG(a, b)$, $p(X|a, b) \propto X^{-(a+1)} exp(-b/X)$. Therefore, $p(\sigma^2_O|\text{Others}) \sim IG(a=(m-1)/2, b=\sum_i \gamma^T_{Oi}\gamma_{Oi}/2)$. In addition, note that $p(\sigma^2_O|\text{Others}) \sim \text{Inv} - \chi^2(\nu=m-1, s^2=\sum_i \gamma^T_{Oi}\gamma_{Oi}/(m-1))$ by the connection of the inverse gamma between scaled inverse chi-square distribution with DOF ν and scale s [19].

Step 7. Draw σ^2_{Rl} from $IG((m-1)/2, \sum_i \gamma^T_{Rli}\gamma_{Rli}/2)$, where the conditional distributions can be derived similarly as in Step 6.

References

1. Schenker N, Raghunathan TE. Combining information from multiple surveys to enhance estimation of measures of health. Statistics in Medicine. 2007; 26:1802–1811. [PubMed: 17278184]
2. Schenker N, Parker JD. From single-race reporting to multiple-race reporting: using imputation methods to bridge the transition. Statistics in Medicine. 2003; 22:1571–1587. [PubMed: 12704616]
3. Yucel RM, Zaslavsky AM. Imputation of binary treatment variables with measurement error in administrative data. Journal of American Statistical Association. 2005; 100:1123–1132.
4. He Y, Zaslavsky AM. Combining information from cancer registry and medical records data to improve analyses of adjuvant cancer therapies. Biometrics. 2009; 65:946–952. [PubMed: 19210743]
5. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.
6. 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]
7. Huskamp HA, Keating NL, Malin JL, Zaslavsky AM,Weeks JC, Earle CC, Teno JM, Virnig BA, Kahn KL, He Y, Ayanian JZ. Discussions with physicians about hospice among patients with metastatic lung cancer. Archives of Internal Medicine. 2009; 169:954–962. [PubMed: 19468089]
8. Mack JW, Cronin A, Taback N, Huskamp HA, Keating NL, Malin JL, Earle CC, Weeks JC. End-of-Life care discussions among patients with advanced cancer: a cohort study. Annals of Internal Medicine. 2012; 156:204–210. [PubMed: 22312140]

9. Earle CC, Landrum MB, Souza JM, Neville BA, Weeks JC, Ayanian JZ. Aggressiveness of cancer care near the end of life: is it a quality-of-care issue. Journal of Clinical Oncology. 2008; 26:3860–3866. [PubMed: 18688053]
10. Little, RJA.; Rubin, DB. Statistical Analysis of Missing Data. New York: Wiley; 2002.
11. Reiter JP, Raghunathan TE. The multiple adaptations of multiple imputation. Journal of the American Statistical Association. 2007; 102:1462–1471.
12. Mislevy RJ. Randomized-based inference about latent variables from complex samples. Psychometrika. 1991; 56:177–196.
13. Cole SR, Chu H, Greenland S. Multiple imputation for measurement-error correction. International Journal of Epidemiology. 2006; 35:1074–1081. [PubMed: 16709616]
14. Sekar C, Deming EW. On a method of estimating birth and death rates and the extent of registration. Journal of the American Statistical Association. 1949; 44:101–115.
15. Alho JM, Mulry MH, Wurdeman K, Kim J. Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. Journal of the American Statistical Association. 1993; 88:1130–1136.
16. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika. 1984; 71:1–10.
17. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association. 1993; 88:669–679.
18. Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis. 2006; 1:515–533.
19. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian Data Analysis. 2nd edn.. New York, NY: CRC Press; 2004.
20. Spiegelhalter, DJ.; Abrams, KR.; Myles, JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester: Wiley; 2004. Section 5.7.3.
21. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association. 1987; 82:528–550.
22. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical Science. 1992; 7:457–472.
23. Gelman A, Meng XL, Stern HS. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). Statistical Sinica. 1996; 6:733–807.
24. Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. Biometrika. 1999; 86:843–855.
25. Harel O, Miglioretti D. Missing information as a diagnostic tool for latent class analysis. Journal of Data Science. 2007; 5:269–288.
26. Molenberghs, G.; Kenward, MG. Missing Data in Clinical Studies. West Sussex: Wiley; 2007.
27. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceau, CM. Measurement Error in Nonlinear Models: A Modern Perspective. 3rd edn.. New York, NY: CRC Press; 2006.
28. Drews CD, Flanders WD, Kosinski AS. Use of two data sources to estimate odds-ratios in case-control studies. Epidemiology. 1993; 4:327–355. [PubMed: 8347743]
29. Gustafson, P. Measurement Error and Misclassification in Statistics and Epidemiology. Boca Raton, FL: CRC Press; 2004.
30. Zaslavsky A, Wolfgang GS. Triple-system modeling of census, post-enumeration survey, and administrative-list data. Journal of Business and Economic Statistics. 1993; 11:279–288.
31. Chib S, Greenberg E. Analysis of multivariate probit models. Biometrika. 1998; 85:347–361.

Table I

Hospice-use reports from medical records and medicare claims, a subsample of CanCORS data.

		Medicare claims		
		Yes	**No**	**Missing**
	Yes	395	54	260
Medical	No	445	617	646
Records	Missing	136	116	358

Note: The subsample consists of 3027 CanCORS patients who died within 15 months of diagnosis (Section 3.1).

Table II

Combining information from two data sources in a missing-data analysis framework.

True hospice-use status, Y_O	Medical records report, Y_{R1}	Medicare claims report, Y_{R2}	Covariates, X
?	1	1	…
?	1	0	…
?	0	?	…
?	0	0	…
?	0	1	…
?	?	1	…
?	?	0	…
?	…	…	…

Note: 1, yes (patient had hospice-use); 0, no; ‘?’, missing.

Table III

Distribution of covariates from medical records and medicare claims, a subsample of CanCORS data.

Variables	Levels (%)
Cancer type	Lung (78%), CRC (22%)
Gender	Male (56%), Female (44%)
Marital status	Married (56%), Unmarried (44%)
Age	<65 (20%), 65–69 (22%), 70–74 (18%), 75–79 (16%), 80+(24%)
Education	<High school (25%), high school (57%), college (18%)
Income	<20K (34%), 20–39K (36%), 40–59K (17%), >60K (13%)
Race	NH-White (74%), NH-Black (11%), Hispanic (6%), NH-Asian (4%), Others (5%)
Cancer stage	I (9%), II (12%), III (28%), IV (46%), unknown (5%)
Comorbidity	Myocardial infarction (21%)
	Chronic heart failure (15%)
	Stroke (18%)
	Lung disease (35%)
	Diabetes (25%)
	Depression (23%)
Time between diagnosis and death	Mean = 123 days, standard deviation = 132 days
Study site	GHC (5.1%), HPHC (.4%), HFHS (2.0%), KPHI (1.6%), KPNW (5.8%), NCCC (22%), UAB (13%), UCLA (22%), IOWA (23%), UNC (5.2%)

Note: The subsample consists of 2669 CanCORS patients who died within 15 months of diagnosis and had at least one observed hospice-use record from two sources (Section 3.1).

CRC, colorectal cancer; GHC, Group Health Cooperative; HPHC, Harvard Pilgrim Healthcare; HFHS, Henry Ford Health System; KPHI, Kaiser Permanente Health Insurance; KPNW, Kaiser Permanente North West; NCCC, Northern California Cancer Center; UAB, University of Alabama; UCLA, University of California at Los Angeles; IOWA, University of Iowa; UNC, University of North Carolina.

Table IV

Posterior predictive checking results from the medical records and medicare claims, a subsample of CanCORS data.

	Medical records			Medicare claims		
Study site	**Observed capturing rate**	**Median from predictions**	***ppp***	**Observed capturing rate**	**Median from predictions**	***ppp***
GHC	.38	.31	.78	.92	.89	.64
HPHC	1	1	.41	1	1	.16
HFHS	.78	.67	.73	.70	.83	.18
KPHI	.20	.38	.48	1	1	.20
KPNW	.67	.67	.48	.85	.85	.51
NCCC	.25	.26	.43	.78	.83	.25
UAB	.20	.22	.30	1	.89	.92
UCLA	.44	.41	.73	.87	.83	.79
IOWA	.73	.71	.66	.90	.88	.72
UNC	.22	.23	.46	1	1	.44
Var	.08	.07	.21	.01	.01	.54

Note: The subsample consists of 2669 CanCORS patients who died within 15 months of diagnosis and had at least one observed hospice-use record from two sources. The posterior predictive checks are based on 105 replicates. *ppp*, posterior predictive *p*-values. For example in the GHC site, the capturing rate is 23/(23 + 38) = .38 for medical records and 23/(23 + 2) = .92 for Medicare claims. The corresponding numerators and denominators can be found in Table AI.

Table V

Coefficient estimates from outcome and reporting models using the medical records and medicare claims, a subsample of CanCORS data.

Covariates	Outcome model	Reporting model for medical records	Reporting model for Medicare claims
Colorectal cancer	−.58*	.16	−.41
65–69 years	.45*	−.39*	NA
70–74 years	.31	−.23	.00
75–79 years	.35	−.25	−.11
80 years or older	.54*	−.29	.10
Female	.09	−.09	.31
Black	−.08	−.18	−.01
Hispanic	.30	.03	−.70*
Asian	.01	.03	−.23
Other race	−.28	.19	−.40
Less than high school	.17	.11	−.08
High school	.04	.22	−.23
Less than 20k	.08	−.21	−.02
20–39k	.09	−.28	−.01
40–59k	.19	−.18	.16
Married	−.10	.03	.15
Stage 2	.12	−.43	−.42
Stage 3	.40*	−.36	−.07
Stage 4	.81*	−.36	−.31
Stage missing	.49*	−.61*	.03
MI	−.05	−.07	−.18
CHF	−.07	.09	.54*
Stroke	.09	.03	−.23
Lung disease	−.09	−.09	.16
Diabetes	−.02	.06	.37*
Depression	.22*	−.12	−.11
Days till death	.30*	−.06	.24*
Site random effects variance	.10	.48	.05

Note The subsample consists of 2669 CanCORS patients who died within 15 months of diagnosis and had at least one hospice-use record from two sources. Outcome model: Equation (1.1), reporting model for medical records: Equation (1.2), and reporting model for Medicare claims: Equation (1.3). Posterior medians for β_O and β_R's are based from 10^5 iterations of the Gibbs chain. Estimates with 95% CIs (not shown) excluding 0 are highlighted with *. The reference categories are lung cancer, younger than 65 years old for the outcome model and reporting model for medical records/younger than 70 years old for the reporting model for Medicare claims, White, college, more than 60K, stage I. The variable 'Days till death' is included in the model using a standardized *Z*-score.

Table VI

Two post-imputation analysis results from alternative missing data methods using the medical records and medicare claims, a subsample of CanCORS data.

Variables	MI	OR	AND	Medical records	Medicare claims
Hospice-use	**62.8%**	**67.6%**	**39.0%**	**29.3%**	**55.4%**
65–69 years	.56*	−1.51*	−.05	−.11	NA
70–74 years	.26	−1.61*	.08	−.03	.10
75–79 years	.36	−1.52*	.13	.04	.09
80 years or older	.68*	−1.22*	.26	.08	.42*
Female	.12	.27*	.20	−.03	.24*
Black	−.31	−.28	−.69*	−.89*	−.20
Hispanic	.31	.11	−.63	−.55*	−.17
Asian	−.36	−.42	−1.54*	−.62*	−.66*
Other race	−.58	−.58*	−1.12*	−.50*	−.83*
Less than high school	.11	.00	.33	.33*	.02
High school	−.03	−.10	.25	.42*	−.17
Less than 20K	.23	−.16	−.08	.01	.18
20–39K	.32	−.09	−.18	−.05	.22
40–59K	.44*	.14	−.10	.12	.37
Married	−.23	−.16	−.05	.02	−.14
Stage 2	.49	.34	.20	−.04	.30
Stage 3	.48*	.35	.23	.13	.29
Stage 4	.90*	.48*	.21	.21	.34
Stage missing	.78*	.68*	−.10	−.22	.44
MI	−.02	−.18	−.16	−.02	−.11
CHF	−.27	−.04	−.04	.06	−.06
Stroke	.08	−.09	.04	.17	−.10
Lung disease	−.13	−.11	−.10	−.18*	−.03
Diabetes	.03	.09	.30*	.04	.28*
Depression	.33*	.18	.03	.00	.15

Note: The subsample consists of 2669 CanCORS patients who died within 15 months of diagnosis and had at least one observed hospice-use record from two sources. The first row (‘hospice use’) lists marginal estimates of the hospice-use rate. Following rows list the logistic regression coefficients (Section 3.4.2). MI, multiple imputation. Standard errors are not shown. Logistic regression coefficients that are significant at 10% level are highlighted. The reference categories are lung cancer, younger than 65 years old for MI, OR, AND, and Medical records/younger than 70 years old for Medicare Claims, White, college, more than 60K, stage I.