Title

Classification accuracy of claims-based methods for identifying providers failing to meet performance targets.

Journal

Statistics in medicine, 34(1)

ISSN

0277-6715

Authors

Hubbard, Rebecca A
Benjamin-Johnson, Rhondee
Onega, Tracy
et al.

Publication Date

2015

DOI

10.1002/sim.6318

Peer reviewed

Published in final edited form as:
Stat Med. 2015 January 15; 34(1): 93–105. doi:10.1002/sim.6318.

Classification accuracy of claims-based methods for identifying providers failing to meet performance targets

Rebecca A. Hubbard[1,2], **Rhondee Benjamin-Johnson**[3], **Tracy Onega**[4], **Rebecca Smith-Bindman**[5], **Weiwei Zhu**[1], and **Joshua J. Fenton**[6]

[1]Group Health Research Institute, Seattle, WA

[2]Department of Biostatistics, University of Washington, Seattle, WA

[3]The Lewin Group, Falls Church, VA

[4]Department of Community and Family Medicine, Norris Cotton Cancer Center, Dartmouth Medical School, Lebanon, NH

[5]Departments of Radiology, Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA

[6]Departments of Family and Community Medicine and the Center for Healthcare Research and Policy, University of California, Davis, Sacramento, CA

Summary

Quality assessment is critical for healthcare reform but data sources are lacking for measurement of many important healthcare outcomes. With over 49 million people covered by Medicare as of 2010, Medicare claims data offer a potentially valuable source that could be used in targeted health care quality improvement efforts. However, little is known about the operating characteristics of provider profiling methods using claims-based outcome measures that may estimate provider performance with error. Motivated by the example of screening mammography performance, we compared approaches to identifying providers failing to meet guideline targets using Medicare claims data. We used data from the Breast Cancer Surveillance Consortium and linked Medicare claims to compare claims-based and clinical estimates of cancer detection rate. We then demonstrated the performance of claim-based estimates across a broad range of operating characteristics using simulation studies. We found that identification of poor performing providers was extremely sensitive to algorithm specificity, with no approach identifying more than 65% of poorly performing providers when claims-based measures had specificity of 0.995 or less. We conclude that claims have the potential to contribute important information on healthcare outcomes to quality improvement efforts. However, to achieve this potential, development of highly accurate claims-based outcome measures should remain a priority.

Keywords

breast cancer; hierarchical models; Medicare; provider profiling

1. Introduction

Quality assessment and improvement using quantitative measures of provider performance has become a high profile activity and is emphasized by the Patient Protection and Affordable Care Act of 2010 as a key element of healthcare reform. One of the challenges to this effort is lack of available data on key health outcome measures. Screening mammography is one area of healthcare in which provider performance has been assessed for some time. The Mammography Quality Standards Act of 1992 (MQSA) mandated that facilities providing screening mammography maintain standardized records of patient outcomes. The policy goal of this requirement was that mammography facilities would use these metrics to monitor and improve the quality of mammography interpretation. However, while these data are available for review by the individual facility, they are not available for public reporting or benchmarking. Thus, even in the case of a healthcare service that is carefully regulated, additional data sources are needed for public reporting and evaluation.

Medicare claims are a potentially rich source for evaluating performance of medical providers. With over 49 million people covered by Medicare [1], these data represent a vast source of information on health care provider performance. Medicare data are already being used to provide healthcare quality information through the Center for Medicare & Medicaid Services' (CMS) Hospital Compare website [2]. As efforts to develop and validate claims-based measures for provider performance outcomes continue, Medicare claims will gain additional value as a potential tool for quality assessment and improvement. However, no existing research has investigated classification accuracy associated with using claims-based measures to classify providers as succeeding or failing to meet performance targets.

Previous literature has developed statistical methods for provider profiling and investigated their performance. Specifically, several alternative methods have been proposed for identifying providers with outlying performance. Hierarchical models that incorporate information about the distribution of performance across providers have been favored by some because they stabilize estimates from providers with smaller patient volumes and decrease the risk of penalizing small providers due to high variability in their performance estimates [3, 4]. Conversely, by shrinking provider estimates towards the population mean, hierarchical methods tend to have poorer sensitivity for identifying providers with extreme performance values and have been criticized on this basis [5, 6]. Fixed effects approaches including indirect standardization may be favored if identifying extreme outliers is a high priority.

The statistical properties of alternative profiling methods also depend on the goal of the profiling activity. Possible objectives of the analysis include obtaining good estimates of individual provider performance measures, comparing the relative performance of providers via ranks, or generating a reliable estimate of the distribution of provider performance. As discussed by Shen and Louis [7], no single statistical method will be optimal for all three of these goals. In the context of screening mammography, estimation of provider performance measures is mandated by the MQSA and benchmarks for acceptable performance have been developed and are a focus of quality assessment and improvement activities. Estimation of provider performance and evaluation of performance estimates relative to existing

benchmarks is thus of primary interest. More broadly, evaluation of performance estimates relative to fixed benchmarks is of interest because the majority of programs offering financial incentives for high quality care award incentives to providers who exceed fixed thresholds [8]. However, accurately identifying providers at the extremes of the performance distribution is also of great clinical importance and has been undertaken in many prior studies [5]. Providers with outlying poor performance are of interest for intervention such as re-training, while outliers with excellent performance may receive bonuses or other incentives.

Using Medicare claims-based quality measures for provider profiling may compound the challenges to estimation of provider performance by introducing an additional source of error. While methods based on clinical performance measures are subject to error due to random variability, claims-based approaches may introduce error due to misclassification if the measure fails to identify clinical events for some patients and erroneously identifies events for others. Although several prior clinical studies have investigated agreement of clinical and claims-based provider performance estimates [8, 9], to our knowledge, no previous study has evaluated the effect of operating characteristics of claims-based measures on alternative statistical approaches to profiling.

This investigation was motivated by the need for standard methods and data sources for evaluating screening mammography performance. The rate at which mammography detects breast cancers that are present at the time of the screening mammogram (the cancer detection rate) is one example of a widely used metric for screening mammography performance. Performance based on this metric has been demonstrated to vary widely between radiologists [10, 11], and there has been interest in identifying poor performing providers. Claims from over 8.5 million women receiving mammograms paid for by Medicare annually [12] coupled with established thresholds for minimally acceptable mammography provider performance may allow for public evaluation of provider performance on a broad scale. We recently developed and validated claims-based algorithms to measure mammography facility interpretive performance including the breast cancer detection rate [13]. While these algorithms have good sensitivity and specificity for identifying individual events, the implications of using algorithms such as this for provider profiling are unknown.

The broad objective of this study was to assess the performance of claims-based approaches to provider profiling using the breast cancer detection algorithm as a motivating example. Our specific objectives were: (1) to compare alternative statistical methods commonly used for profiling when applied to performance estimates based on claims and (2) to evaluate the performance of claims-based provider performance estimates as a function of the operating characteristics of the claims-based algorithm used for outcome ascertainment. In Section 2, we introduce notation and methods for conducting provider profiling using claims-based algorithms and describe an illustrative approach to estimating breast cancer detection rates using Medicare claims data. We then describe simulation studies conducted to characterize performance of claims-based algorithms for estimating provider performance. We also introduce data from the Breast Cancer Surveillance Consortium (BCSC) and matched Medicare claims that were used to estimate breast cancer detection rates using claims and

clinical outcomes. Section 3 provides results of simulation studies and analyses of the Medicare-linked BCSC data. Finally, in Section 4, we summarize these results and draw conclusions on the contexts in which provider profiling may be feasible using claims data.

2. Methods

2.1 Notation and definitions

In estimation of provider performance, we assume there are N providers, with n_i patients observed for the ith provider and a total of y_i events observed across the n_i patients. We assume that y_i, the clinical outcome measure, is unobserved because claims do not directly capture outcomes of interest. We focus on binary outcomes, although similar considerations would apply to continuous outcomes. We assume that each provider has an underlying, true performance measure, θ_i, arising from a possibly unknown distribution with mean μ and variance σ^2. Our objective is to classify providers as to whether their performance measures fail to meet some performance benchmark. Without loss of generality, we assume a provider fails to meet the guideline target and is a true poor performer if $\theta_i < k$.

2.2 Claims-based outcome measures

We assume the existence of an algorithm based on claims data that can be used to predict the outcome of interest, y_i. Let z_{ij} represent a binary indicator of whether the jth patient of the ith provider experienced an event and $\tilde{z}_{ij}$ represent a binary classification derived from the claims-based algorithm indicating whether claims identified an event as having occurred. The claims-based measure of the number of events for provider i is thus given by $\tilde{y}_i = \sum_{j=1}^{n_i} \tilde{z}_{ij}$. We characterize the claims-based algorithm in terms of its sensitivity, $S = P(\tilde{z}_{ij}=1 | z_{ij}=1)$, and specificity, $P = P(\tilde{z}_{ij}=0 | z_{ij}=0)$. Claims-based performance estimates are based on outcomes, $\tilde{y}_i$, obtained by applying an algorithm to claims data, while clinically-based performance measures are based directly on y_i. It can also be seen that $\tilde{y}_i = \tilde{y}_i^S + \tilde{y}_i^P$, where $\tilde{y}_i^S$ is binomially distributed with mean S and sample size y_i and $\tilde{y}_i^P$ is binomially distributed with mean (1-P) and sample size n_i-y_i. Since $E(\tilde{y}_i | y_i) = S y_i + (1-P)(n_i - y_i)$, $\tilde{y}_i$ and y_i differ in expectation except in the case of an algorithm with perfect sensitivity and specificity. Indeed, in the case of an algorithm with perfect sensitivity and specificity, clinical and claims-based measures are identical. Holding specificity constant, as algorithmic sensitivity decreases, y_i will tend to be underestimated and, holding sensitivity constant, as specificity decreases y_i will tend to be overestimated. We can correct this by using a bias-adjusted estimator that explicitly accounts for the known bias in $\tilde{y}_i$. We define $\tilde{y}_i^* = \tilde{y}_i + (1-S) n_i \hat{\theta} - (1-P) n_i \left(1 - \hat{\theta}\right)$, where $\hat{\theta}$ is an estimate of the population mean, to be the bias-adjusted claims-based number of events observed for the ith provider.

Any of the three outcome measures described above, y_i,$\tilde{y}_i$, or $\tilde{y}_i^*$ can be used in standard statistical methods for provider profiling. Below we illustrate approaches using y_i. However, statistical methods based on $\tilde{y}_i$ or $\tilde{y}_i^*$ are analogous.

2.3 Methods for estimating provider performance

2.3.1 Fixed effects estimates of provider performance—A standard approach to provider profiling is to estimate θ_i using the maximum likelihood estimator (MLE). In the simple setting of a binomial outcome with no patient or provider characteristics to be adjusted for as described in Section 2.1, the MLE is simply the sample mean, y_i/n_i. In many cases, patient-characteristics will be strongly associated with the probability of an event, and adjustment for these characteristics will be important for obtaining unbiased estimates of provider performance. In this case, a generalized linear model of the form

$$g\left(E\left(z_{ij}\right)\right)=X_{ij}\beta+\gamma_i, \quad (1)$$

can be used, where g(.) is a link function relating the expectation of an event to patient- and provider-specific characteristics, X_{ij} is a vector of covariates for the *j*th patient of the *i*th provider, and γ_i is a provider-specific fixed effect. Predicted provider performance measures can then be derived either by calculating $\hat{\theta}_i=\mathrm{g}^{-1}\left(X^*\hat{\beta}+\hat{\gamma}_i\right)$ for some set of covariates, *X**, with the same vector of covariates used for all providers or by computing a standardized version of $\hat{\theta}_i$, estimated by taking a weighted average of $g^{-1}\left(X\hat{\beta}+\hat{\gamma}_i\right)$ for all values of *X* with weights proportional to the representation of each value of *X* in the population of interest. Upper and lower confidence limits can also be constructed for our fixed effects estimates using standard methods.

2.3.2 Hierarchical Bayesian estimates of provider performance—Although straightforward, using a fixed effects approach for classifying providers has been criticized for failing to account for instability of estimates for providers with small patient volumes. Performance estimates from providers with small patient volumes will be unstable and hence extreme performance values may be observed by chance. In order to stabilize performance estimates, a number of authors have proposed using hierarchical Bayesian methods to estimate provider performance [3-5, 14-18]. In addition to stabilizing estimates from providers with small sample sizes, Bayesian methods are also appealing because they allow for direct estimation of the probability that a provider's true performance score falls below a performance threshold.

The first level of the Bayesian hierarchical model takes the form given in equation (1). However, we further assume a second-level model in which provider-specific effects arise from a common distribution, $\gamma_i \sim G(\alpha)$, where α is a vector of hyperparameters. A prior distribution for the hyperparameters, $\pi(\alpha)$, is also assumed. Based on this model, the posterior distribution for θ_i can be derived either conditional on a specific vector of covariates which is used in common for all providers or by standardizing across the distribution of covariate vectors as described under the fixed effects approach. For some choices of link functions and prior distributions the posterior distribution may be available in closed form. However, this will not be the case in general, and a Markov Chain Monte Carlo approach for simulating from the posterior distribution will typically be required.

An empirical Bayes approach to estimating provider performance is an alternative to the fully Bayesian approach described above. The empirical Bayes approach may be preferred

over a fully Bayesian framework because it avoids the necessity of specifying a prior distribution for the hyperparameters and can lead to computationally efficient, closed form estimators in some cases. In the simple case of a binary outcome with no covariate adjustment and assuming a conjugate beta distribution with parameters a and b for θ_i, the posterior distribution for θ_i is available in closed form as Beta($y_i + a, n_i + b$). In the empirical Bayes framework, a and b are obtained using estimates of the parameters of the marginal distribution of the observed data. In the case of binomial data with beta distributed provider-specific means, this marginal distribution is beta-binomial. MLEs for a and b are not explicitly available but can easily be obtained through numerical maximization of the beta-binomial marginal likelihood. Alternatively, method of moments estimators are available in closed form [19].

Once the posterior distribution for θ_i has been obtained, point estimates and interval estimates can be directly derived from the posterior. Alternatively, in the case of classification relative to a fixed threshold it may be of interest to estimate the posterior probability that a provider falls above or below a target performance level.

2.3.3 Classification using fixed effects or hierarchical Bayesian estimates—

Using the ML and Bayesian provider performance estimation approaches described above, we will evaluate four methods for identifying providers performing below specified benchmarks:

1. Classification based on ML fixed effects estimates. Providers with $\hat{\theta}_i < k$ are classified as failing to meet performance benchmarks.
2. Classification based on ML fixed effects confidence intervals. Providers with $\hat{\theta}_i^{1-\alpha/2} < k$ are classified as failing to meet performance benchmarks, where $\hat{\theta}_i^{1-\alpha/2}$ represents the upper $\alpha/2$ confidence interval limit for $\hat{\theta}_i$.
3. Classification based on Bayesian posterior means. Providers with $E(\theta_i \mid y_i) < k$ are classified as failing to meet performance benchmarks.
4. Classification based on posterior probabilities. Providers with posterior probabilities of failing to meet performance targets greater than some threshold probability, $P(\theta_i < k | y_i) > p^*$, are classified as failing to meet performance benchmarks.

We contrast approaches 1 and 3 based on provider performance point estimates with approaches 2 and 4 which incorporate uncertainty in our estimates based on sampling error. Additionally, approaches 1 and 2 are based on fixed effects estimates which may be unstable for providers with small volumes while approaches 3 and 4 use a hierarchical Bayesian model to stabilize estimates for small volume providers by shrinking them towards the population mean. Choice of the probability threshold for method 4, p^*, can be based on the relative costs of failing to identify a truly poor performing provider compared to erroneously labeling a provider with acceptable performance as poor [16, 18]. In numerical examples presented below, we chose to use a probability threshold of 75%, representing a case in

which the cost of falsely labeling a provider as a poor performer is three times greater than the cost of failing to identify a truly poor performing provider.

As an alternative to evaluating performance relative to a fixed threshold, we can also explore the relative ordering of providers based on either fixed effects or hierarchical Bayesian point estimates using receiver operating characteristic (ROC) curves. By using ROC curves we are able to compare the sensitivity and specificity of classification for a range of thresholds rather than focusing on classification accuracy relative to a fixed benchmark. The ROC analysis also allows us to evaluate the accuracy of provider rankings rather than focusing on absolute performance estimates.

2.4 Claims-based algorithms for breast cancer detection rate

We use a claims-based algorithm for breast cancer detection as an example claims-based approach to identifying outcomes. Like many outcomes of interest for quality assessment and improvement, breast cancer detection is a rare event, with a prevalence of approximately 0.5% [20]. In the context of screening mammography interpretive performance, the objective is to achieve a high cancer detection rate and providers with very low cancer detection rates would be flagged as poor performers.

We have previously developed a simple claims-based algorithm for estimating the breast cancer detection rate. The breast cancer detection rate algorithm uses procedure codes for imaging and biopsy subsequent to the mammogram as well as diagnosis codes for invasive breast cancer and carcinoma *in situ* to identify detected cancers. This algorithm had a sensitivity of 94% and specificity of 99.9% [13]. In simulation studies below, we illustrate the performance of profiling methods for identifying providers failing to meet guideline thresholds when outcomes are estimated from claims using algorithms with operating characteristics similar to this existing algorithm. In the context of screening mammography, the objective of profiling is to identify providers with cancer detection rates below targets. Previous work has proposed a benchmark of 0.2% as a minimally acceptable performance target for cancer detection rate [21].

2.5 BCSC and Medicare data

We used data on screening mammography to compare the agreement of performance estimates based on claims and clinical outcomes using data from the National Cancer Institute-funded Breast Cancer Surveillance Consortium (BCSC) (http://breastscreening.cancer.gov). The BCSC links information on women who receive a mammogram at a participating facility to regional cancer registries and pathology databases to determine breast cancer outcomes. BCSC facilities submit prospectively collected patient and mammography data to regional registries, which link the data to breast cancer outcomes ascertained from cancer registries. Mammography data include radiologist information on the purpose for the examination (screening or diagnostic) and interpretations (normal or abnormal). The BCSC has established standard definitions for key variables and multiple levels of data quality control and monitoring [22]. BCSC sites have received institutional review board approval for active or passive consenting processes or a waiver of consent to enroll participants, link data, and perform analytic studies. All procedures are Health

Insurance Portability and Accountability Act compliant, and BCSC sites have received a Federal Certificate of Confidentiality to protect the identities of patients, physicians, and facilities.

Medicare claims from 1998 to 2006 were linked with BCSC mammography data derived from regional mammography registries in four states (North Carolina; San Francisco Bay Area, California; New Hampshire; and Vermont). We used data from Medicare claims files (the Carrier Claims, Outpatient, and Inpatient files) and the Medicare denominator file, which provides demographic, enrollment, and vital status data.

We identified a sample of screening mammograms performed in 2003-2005 appearing in both Medicare claims and BCSC data with the same date of service using a validated claims-based algorithm based upon Healthcare Common Procedure Coding System (HCPCS) mammography codes and International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes for breast cancer [23]. We selected mammograms for women age 68 years and older with continuous enrollment in fee-for-service Medicare (parts A and B) for twelve months after mammography and three years prior to mammography, required by the claims based algorithms, enabling both prospective assessment of outpatient claims that might indicate incident breast cancer following abnormal screening mammography, and retrospective assessment for claims indicating prevalent breast cancer.

Our clinical breast cancer detection outcome was defined as a screening mammogram with a positive result based on the radiologist's assessments and recommendations followed by a diagnosis of invasive cancer or ductal carcinoma *in situ* based on cancer registry data within one year. For each mammogram in the linked BCSC-Medicare data we defined both a clinical breast cancer detection outcome and a claims-based outcome derived by applying the previously developed algorithm described in Section 2.4.

We compared clinical and claims-based estimates of facility-level breast cancer detection rates using ML fixed effects and posterior means estimated via an empirical Bayes approach in terms of Spearman's correlation. We also compared the concordance of classification of facilities as failing to meet guideline thresholds based on either clinical or claims-based estimates of provider performance using the four methods described in Section 2.3.3. For claims-based estimates we computed both unadjusted and bias-adjusted measures. Concordance was estimated using the kappa statistic.

Facilities with less than 50 mammograms observed over the three-year study period were excluded from analysis. We investigated alternative minimum volume criteria up to a minimum of 1000 mammograms for facility but results were not qualitatively different. Facilities were classified as failing to meet guideline thresholds using the previously proposed standards for mammography performance of 0.2% for cancer detection rate [21].

2.6 Simulation study design

In addition to comparing the observed agreement between performance estimates based on BCSC clinical data and Medicare claims, we conducted simulation studies to evaluate the performance of claims-based algorithms for classifying providers as failing to meet

guideline thresholds across a range of values for sensitivity and specificity of the claims-based algorithm for identifying events. These simulation studies were undertaken because comparing the observed performance of clinical and claims-based approaches provides information on the concordance of these two measures but is not informative with respect to the underlying true performance of the provider. In order to understand performance of clinical and claims-based estimates relative to true provider performance, we have conducted a simulation study which allows us to know the underlying, true value of θ_i.

For each simulation scenario, we simulated 1000 samples consisting of an average of 1000 providers. In our simulation studies, patient volume was generated using the distribution of volumes observed for mammography facilities. However, in general, providers can be considered to be facilities or physicians. For each provider we simulated a patient volume, n_i, and a true performance measure value, θ_i. We assumed that provider volumes were gamma distributed and that performance measures were beta distributed across providers. Conditional on a provider's true performance and volume, we then simulated a true number of events, y_i, from a binomial distribution with mean θ_i. Conditional on y_i, we sampled from binomial distributions with sample size y_i and mean S to obtain $\tilde{y}_i^S$ and with sample size ($n_i - y$) and mean ($1 - P$) to obtain $\tilde{y}_i^P$. The observed number of events based on claims was $\tilde{y}_i = \tilde{y}_i^S + \tilde{y}_i^P$. Note that in the setting where $S = 1$ and $P = 1$, $\tilde{y}_i = y_i$, that is, the clinical and claims-based measures are identical. Therefore, this setting provides information on the performance of the clinical measure and can be contrasted with the performance of claims-based measures as S and P deviate from 1.

Parameters for the distribution describing provider volume was chosen based on the distribution of 3,687 mammography outpatient facilities with reported Medicare claims data for 2010 that were included as part of the CMS Hospital Compare Program. The distribution data used in this study was supported by the Centers for Medicare & Medicaid Services (contract HHSM-500-2008-00020I/Task Order 0002) as part of its Hospital Outpatient Quality Reporting program. We used BCSC data on screening mammograms performed in 2010 to select parameters for the beta distribution for facility breast cancer detection rates used in simulation studies. Based on these data, in simulations reported below provider volumes were assumed Gamma(0.29, 1835.67) distributed. Simulated facilities with volume less than 50 mammograms were removed from the sample. Cancer detection rates were assumed Beta(1.36, 372.69) distributed.

We first evaluated performance by computing the sensitivity and specificity of classification into a "poor" performance category based on failure to meet the guideline threshold of 0.2% using unadjusted and bias-adjusted claims-based outcomes and each of the four classification approaches described in Section 2.3.3. We then compared discrimination of adjusted and unadjusted ML and posterior mean estimates without requiring a known fixed threshold using receiver operating characteristics (ROC) curves.

3. Results

3.1 Agreement of BCSC and Medicare performance estimates

We identified a sample of 134,330 screening mammograms performed between 2003 and 2005 that were included in both Medicare claims and BCSC records. Data were from 106 mammography facilities with a range of 52 to 5,925 mammograms per facility. Based on the ML fixed effects approach, BCSC clinical data produced an estimate of 0.56% for the cancer detection rate, while claims data resulted in an estimate of 0.59%. Using a proposed benchmark of 0.2% screen-detected cancers, ML point estimates from clinical data indicated that 19 facilities (18%) were below the cancer detection rate target, while unadjusted claims data identified 18 (17%) such poor performing facilities, and adjusted claims data flagged 21 (20%) poor performing facilities. Bayesian estimates based on either clinical or claims-based outcome measures failed to identify any poor performing facilities. Fixed effects ML point estimates and hierarchical Bayesian posterior means both displayed substantial uncertainty (Figure 1). Bayesian estimates also showed notable shrinkage towards the population mean, diminishing the overall range of variability observed in estimates of provider performance.

ML estimates of facility-level cancer detection rates from the BCSC and Medicare claims were strongly correlated ($R = 0.962$), as were Bayesian posterior means ($R = 0.937$). Clinical and claims-based classifications of facilities as failing to meet the 0.2% benchmark also showed strong agreement with kappa in excess of 0.87 for both the point estimate and confidence interval-based approaches (Table 1). Using either BCSC clinical data or Medicare claims data, no facilities were classified as poor providers by either the Bayesian posterior mean or posterior probability approaches. Kappa statistics are thus omitted for these approaches.

3.2 Simulation study results

3.2.1 Comparison of alternative methods—Across all algorithm operating characteristics investigated, classification based on point estimates was more sensitive and less specific than either the confidence interval or posterior probability approaches, which incorporate uncertainty (Figures 2 and 3). For instance, for a claims-based algorithm with sensitivity of 0.9 and specificity of 1.0, similar to the operating characteristics of an existing cancer detection rate algorithm, the ML and Bayesian point estimate approaches had sensitivity of 0.858 and 0.563, respectively, and specificity of 0.692 and 0.890. The confidence interval and posterior probability approaches had sensitivity of only 0.093 and 0.377, respectively, and specificity of 0.999 and 0.963. Fixed effects estimates also had greater sensitivity and poorer specificity than Bayesian methods. Incorporating the bias adjustment substantially improved the sensitivity of all four approaches, but at the cost of decreased specificity. Bias-adjusted approaches also had more stable performance across the range of algorithm operating characteristics investigated than did unadjusted approaches.

Comparing the discrimination of the fixed effects ML and hierarchical Bayesian approaches without requiring a fixed threshold, we found little difference between bias-adjusted and unadjusted methods (Figure 4). The hierarchical Bayesian approach outperformed the ML

approach in the low specificity range because fixed effects estimates are unable to distinguish between the relatively large population of providers with no observed events, all of whom were classified as poor performers. No method was able to simultaneously achieve both good sensitivity and specificity for identifying poor performers.

3.2.2 Classification performance as a function of algorithm sensitivity and specificity—In the case of perfect algorithmic sensitivity and specificity, corresponding to a clinical outcome measure, the fixed effects ML point estimate achieved a sensitivity of 0.834 and specificity of 0.728 for classifying providers as failing to meet the guideline target for cancer detection rate (Figures 2 and 3). Other measures performed more poorly with the ML confidence interval approach performing the most poorly with classification sensitivity of only 0.082 and specificity of 1.0. As algorithm specificity decreased, the performance of the unadjusted approaches declined. For instance, at algorithmic sensitivity of 1 and specificity of 0.995 the ML fixed effects approach achieved a sensitivity of only 0.164. Correspondingly, classification specificity increased to near 1 for all unadjusted methods. Performance was relatively insensitive to algorithm sensitivity. Across a range of sensitivities from 0.7 to 1.0, classification sensitivity increased by only about 8% for the ML fixed effects method. Changes in classification sensitivity were similar for other approaches.

4. Discussion

We investigated the performance of provider profiling methods using claims-based approaches for measuring outcomes. In comparisons of provider performance estimates based on BCSC clinical data and Medicare claims data, we found that claims-based estimates of cancer detection rates corresponded well with estimates based on information from the radiology practice and cancer registry data. Classification of providers as poor performers also agreed closely when clinical or claims-based outcomes were used. In simulation studies, we found that claims-based provider performance estimates based on the ML fixed effects approach achieved fair performance when algorithmic specificity was high. In the context of a rare outcome like cancer detection rate, a highly specific algorithm is required because even minor deviations from perfect algorithmic specificity compromise the ability of claims-based approaches to identify providers failing to meet targets.

Both fixed effects and hierarchical Bayesian estimates incorporating uncertainty via confidence intervals or posterior probabilities had poorer sensitivity than classification based on point estimates because they are by design more conservative. The choice of whether or not to incorporate uncertainty should be based on the relative impact of failing to identify a truly poorly performing provider vs. erroneously classifying a truly acceptably performing provider as failing to meet guideline targets. Previous work has noted that hierarchical Bayesian methods may result in failing to identify truly outlying providers, especially for providers with small sample sizes [5, 6]. The relative drawback of such approaches depends on the purpose of the profiling effort and the importance of identifying as many providers failing to meet targets as possible. In the case of the posterior probability approach, the relative importance of the two types of classification errors can be tuned by varying p^*, the posterior probability required before a provider is classified as failing to meet the target. Previous work has placed the choice of posterior probability required for classification of a

provider as failing to meet targets in a decision theoretic context [16, 18]. Such considerations may be useful in selecting a posterior probability threshold for this purpose. Because confidence interval and hypothesis testing-based approaches to provider classification may result in undesirably high rates of misclassification of truly acceptably performing providers, thresholds that correct for the number of comparisons, i.e. number of providers, may be preferred. Jones et al. [24] have proposed the use of the false discovery rate to limit this type of misclassification.

One previous investigation explored the effect of errors in outcome ascertainment on provider profiling [25]. In the context of provider performance estimates based on claims data, bias may arise through imperfect sensitivity and specificity of the claims-based algorithm for event ascertainment and through variability in coding practices across providers which may result in differential performance of the algorithm for some providers. Previous work [25] proposed a method using auxiliary data measured without error to correct for this bias. However, such auxiliary data are not generally available. It is also important to note that systematic error due to imperfect outcome ascertainment is only one component of the challenge to claims-based provider profiling. Random variability also plays a critical role, particularly for rare outcomes. In our simulation study, we found that performance of provider profiling methods was relatively constant across a range of values for algorithm sensitivity and specificity. Thus random variability appears to be the dominant factor in provider misclassification for algorithms with excellent operating characteristics such as the cancer detection algorithm used as the motivating example for this study as well as for those with more modest operating characteristics. We also found relatively little improvement in classification accuracy after applying a bias-adjustment factor to provider outcome counts, corroborating the finding that error in outcome ascertainment plays less of a role in classification accuracy than random error does. Past research has emphasized the importance of quantifying random error when determining whether provider profiling will be feasible in a given context [26, 27]. Our results reinforce this idea by demonstrating that although clinical and claims-based measures may agree closely, performance of both approaches may be quite poor when evaluated relative to true provider performance as demonstrated in our simulation studies which incorporated both random error and systematic error due to imperfect claims-based algorithm operating characteristics.

Previous studies of profiling for mammography performance measures have focused on the importance of covariate adjustment [28]. In the context of Medicare claims, it may be possible to adjust for a limited set of patient characteristics such as age, race, and measures of comorbidity. Both the fixed effects and hierarchical Bayesian methods discussed here can accommodate covariate adjustment. Challenges associated with using Medicare claims for profiling will persist regardless of whether case-mix adjustment is carried out or not. Indeed, incorporating covariates may exacerbate issues of variability by introducing an additional source of error. Previous research on mammography provider performance that incorporated covariate effects found that no providers could be classified as failing to meet standards for identifying breast cancers due to the rarity of this outcome and resultant substantial uncertainty in performance estimates [28].

An inherent limitation of using Medicare claims for provider profiling is that performance estimates can only be obtained for the population of patients over age 65. To the extent that performance in this population differs from performance in a younger population, Medicare-based estimates may be biased relative to performance estimates for the total population of patients. This limitation could be mitigated by using other claims databases that include patients of all ages or by restricting inference to the population age 65 years and older.

Our simulation studies suggest that cancer detection algorithms hold promise for evaluation of provider performance. However, because of sampling variability in facility-specific estimates of cancer detection rate due to low outcome prevalence, methods incorporating uncertainty performed poorly. Performance of Bayesian approaches to classification in the rare outcome setting could be improved if substantial prior information were available. Classification based on the MLE has reasonable sensitivity but should be interpreted cautiously as our simulation studies demonstrated that, for an algorithm with operating characteristics similar to an existing cancer detection rate algorithm, about 30% of providers exceeding the cancer detection rate threshold will be erroneously classified as poor performers. As new algorithms for performance measures emerge our simulation study results will provide a guide to the expected classification accuracy of profiling approaches using these measures.

Our results indicate that Medicare claims can be used for estimating provider performance measures, but only when the operating characteristics of the claims-based algorithm for event ascertainment are known to be good. Using the example of the cancer detection rate algorithm, we were able to identify poorly performing providers with a sensitivity of 86% and specificity of 70%. This specificity is likely insufficient for the purposes of public reporting where the cost of erroneously classifying a good performer as poor is high. However, for the purposes of providing feedback to mammography facilities for quality improvement this may be adequate. For rare outcomes, such as breast cancer detection, a highly specific claims-based algorithm is required in order to achieve reasonable estimation of provider performance. Given the vast quantity of data on provider performance available via Medicare claims data and the importance of quality assessment and improvement for healthcare reform, development of high quality claims-based outcome measures should remain a priority.

Acknowledgments

We thank the BCSC investigators, participating women, mammography facilities, and radiologists for the data they have provided for this study. A list of the BCSC investigators and procedures for requesting BCSC data for research purposes are provided at: http://breastscreening.cancer.gov/. The Research Data Assistance Center (ResDAC) provides free assistance to academic and non-profit researchers interested in using Medicare, Medicaid, SCHIP, and Medicare Current Beneficiary Survey (MCBS) data for research. Primary funding for ResDAC comes from a CMS research contract. ResDAC is a consortium of faculty and staff from the University of Minnesota, Boston University, Dartmouth Medical School, and the Morehouse School of Medicine. ResDAC offers a number of services for researchers with all levels of experience using or planning to use CMS data. Services include technical data assistance, information on available data resources, and training. Procedures for requesting Medicare data for research purposes are provided at: http://www.resdac.org/cms-data/request/cms-data-request-center.

Funding sources

This work was supported by the National Cancer Institute-funded grant R21CA158510, the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C), and the Centers for Medicare &

Medicaid Services (contract HHSM-500-2008-00020I/Task Order 0002). The collection of cancer data used in this study was supported in part by several state public health departments and cancer registries throughout the U.S. For a full description of these sources, please see: http://www.breastscreening.cancer.gov/work/acknowledgement.html. The content is solely the responsibility of the authors and does not represent the official views of the US Department of Health and Human Services (DHHS), the Centers for Medicare & Medicaid Services (CMS), the National Cancer Institute (NCI), or the National Institutes of Health (NIH).

References

1. [Accessed February 21, 2014] United States Census. Census 2010: Health Insurance. Available at: http://www.census.gov/hhes/www/hlthins/data/incpovhlth/2010/highlights.html
2. Center for Medicare and Medicaid Services. [Accessed February 21, 2014] Hospital Compare. Available at: http://www.hospitalcompare.hhs.gov/
3. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. Annals of Internal Medicine. 1997; 127:764–768. [PubMed: 9382395]
4. Normand SLT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: Issues and applications. Journal of the American Statistical Association. 1997; 92:803–814.
5. Austin PC, Alter DA, Tu JV. The use of fixed- and random-effects models for classifying hospitals as mortality outliers: A Monte Carlo assessment. Medical Decision Making. 2003; 23:526–539. [PubMed: 14672113]
6. Racz MJ, Sedransk J. Bayesian and Frequentist Methods for Provider Profiling Using Risk-Adjusted Assessments of Medical Outcomes. Journal of the American Statistical Association. 2010; 105:48–58.
7. Shen W, Louis TA. Triple-goal estimates in two-stage hierarchical models. Journal of the Royal Statistical Society - Series B. 1998; 60:455–471.
8. Mack MJ, Herbert M, Prince S, Dewey TM, Magee MJ, Edgerton JR. Does reporting of coronary artery bypass grafting from administrative databases accurately reflect actual clinical outcomes? Journal of Thoracic and Cardiovascular Surgery. 2005; 129:1309–1317. [PubMed: 15942571]
9. Shahian DM, Silverstein T, Lovett AF, Wolf RE, Normand SL. Comparison of clinical and administrative data sources for hospital coronary artery bypass graft surgery report cards. Circulation. 2007; 115:1518–1527. [PubMed: 17353447]
10. Beam CA, Lavde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. Archives of Internal Medicine. 1996; 156:209–213. [PubMed: 8546556]
11. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, Yankaskas BC, Kerlikowske K, Onega T, Rosenberg RD, Sickles EA, Buist DS. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology. 2009; 253:641–651. [PubMed: 19864507]
12. Lee DW, Stang PE, Goldberg GA, Haberman M. Resource use and cost of diagnostic workup of women with suspected breast cancer. Breast Journal. 2009; 15:85–92. [PubMed: 19120378]
13. Fenton JJ, Onega T, Zhu W, Balch S, Smith-Bindman R, Henderson L, Sprague B, Kerlikowske K, Hubbard RA. Validation of a Medicare Claims-based Algorithm for Identifying Breast Cancers Detected at Screening Mammography. Medical Care. 2013 In press.
14. Austin PC. A comparison of Bayesian methods for profiling hospital performance. Medical Decision Making. 2002; 22:163–172. [PubMed: 11958498]
15. Austin PC. The reliability and validity of Bayesian measures for hospital profiling: a Monte Carlo assessment. Journal of Statistical Planning and Inference. 2005; 128:109–122.
16. Austin PC. Bayes rules for optimally using Bayesian hierarchical regression models in provider profiling to identify high-mortality hospitals. BMC Medical Research Methodology. 2008; 8:1–11. [PubMed: 18215293]
17. Austin PC, Brunner LJ. Optimal Bayesian probability levels for hospital report cards. Health Services and Outcomes Research Methodology. 2008; 8:80–97.
18. Lin RH, Louis TA, Paddock SM, Ridgeway G. Loss Function Based Ranking in Two-Stage, Hierarchical Models. Bayesian Analysis. 2006; 1:915–946. [PubMed: 20607112]

19. Carlin, BP.; Louis, TA. Bayes and empirical Bayes methods for data analysis. 2nd edn. Chapman & Hall; New York: 2000.
20. Rosenberg RD, Yankaskas BC, Abraham LA, Sickles EA, Lehman CD, Geller BM, Carney PA, Kerlikowske K, Buist DS, Weaver DL, Barlow WE, Ballard-Barbash R. Performance Benchmarks for Screening Mammography. Radiology. 2006; 241:55–66. [PubMed: 16990671]
21. Carney PA, Sickles EA, Monsees BS, Bassett LW, Brenner RJ, Feig SA, Smith RA, Rosenberg RD, Bogart TA, Browning S, Barry JW, Kelly MM, Tran KA, Miglioretti DL. Identifying minimally acceptable interpretive performance criteria for screening mammography. Radiology. 2010; 255:354–361. [PubMed: 20413750]
22. Ballard-Barbash R, Taplin SH, Yankaskas BC, Ernster VL, Rosenberg RD, Carney PA, Barlow WE, Geller BM, Kerlikowske K, Edwards BK, Lynch CF, Urban N, Chrvala CA, Key CR, Poplack SP, Worden JK, Kessler LG. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. AJR Am J Roentgenol. 1997; 169:1001–1008. [PubMed: 9308451]
23. Fenton JJ, Zhu W, Balch S, Smith-Bindman R, Fishman P, Hubbard RA. Distinguishing Screening from Diagnostic Mammograms using Medicare Claims Data. Medical Care. 2012 Epub ahead of print. doi: 10.1097/MLR.0b013e318269e0f5.
24. Jones HE, Ohlssen DI, Spiegelhalter DJ. Use of the false discovery rate when comparing multiple health care providers. Journal of Clinical Epidemiology. 2008; 61:232–240. [PubMed: 18226745]
25. Roy J, Mor V. The effect of provider-level ascertainment bias on profiling nursing homes. Statistics in Medicine. 2005; 24:3609–3629. [PubMed: 16158404]
26. Adams JL, Mehrotra A, Thomas JW, McGlynn EA. Physician cost profiling--reliability and risk of misclassification. New England Journal of Medicine. 2010; 362:1014–1021. [PubMed: 20237347]
27. Adams, J. The Reliability of Provider Profiling: A Tutorial. RAND Corporation; Santa Monica, CA: 2009.
28. Woodard DB, Gelfand AE, Barlow WE, Elmore JG. Performance assessment for radiologists interpreting screening mammography. Statistics in Medicine. 2007; 26:1532–1551. [PubMed: 16847870]

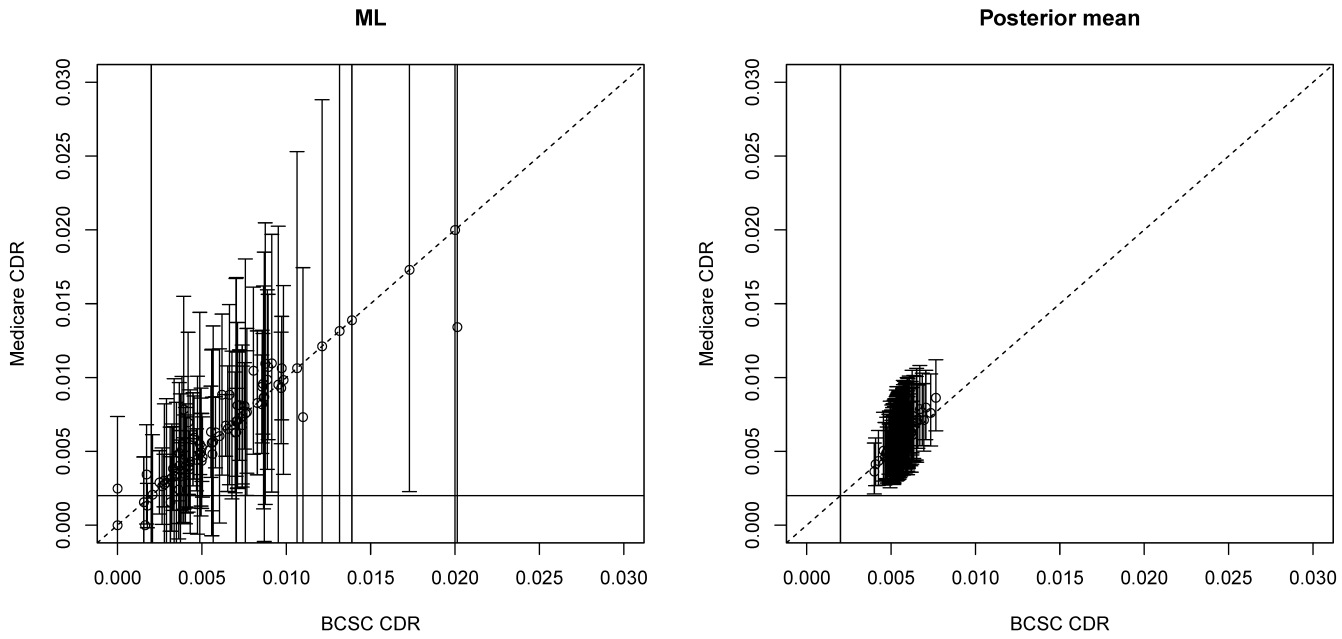

Figure 1.

Fixed effects maximum likelihood estimates and 95% confidence intervals (left) and Bayesian posterior mean estimates and 95% credible intervals (right) based on a Medicare claims-based algorithm for cancer detection rate vs BCSC clinical data for 106 mammography facilities. Horizontal and vertical lines represent thresholds for poor performance. Dotted line represents perfect agreement between BCSC and Medicare performance estimates.

BCSC=Breast Cancer Surveillance Consortium; CDR=Cancer detection rate.

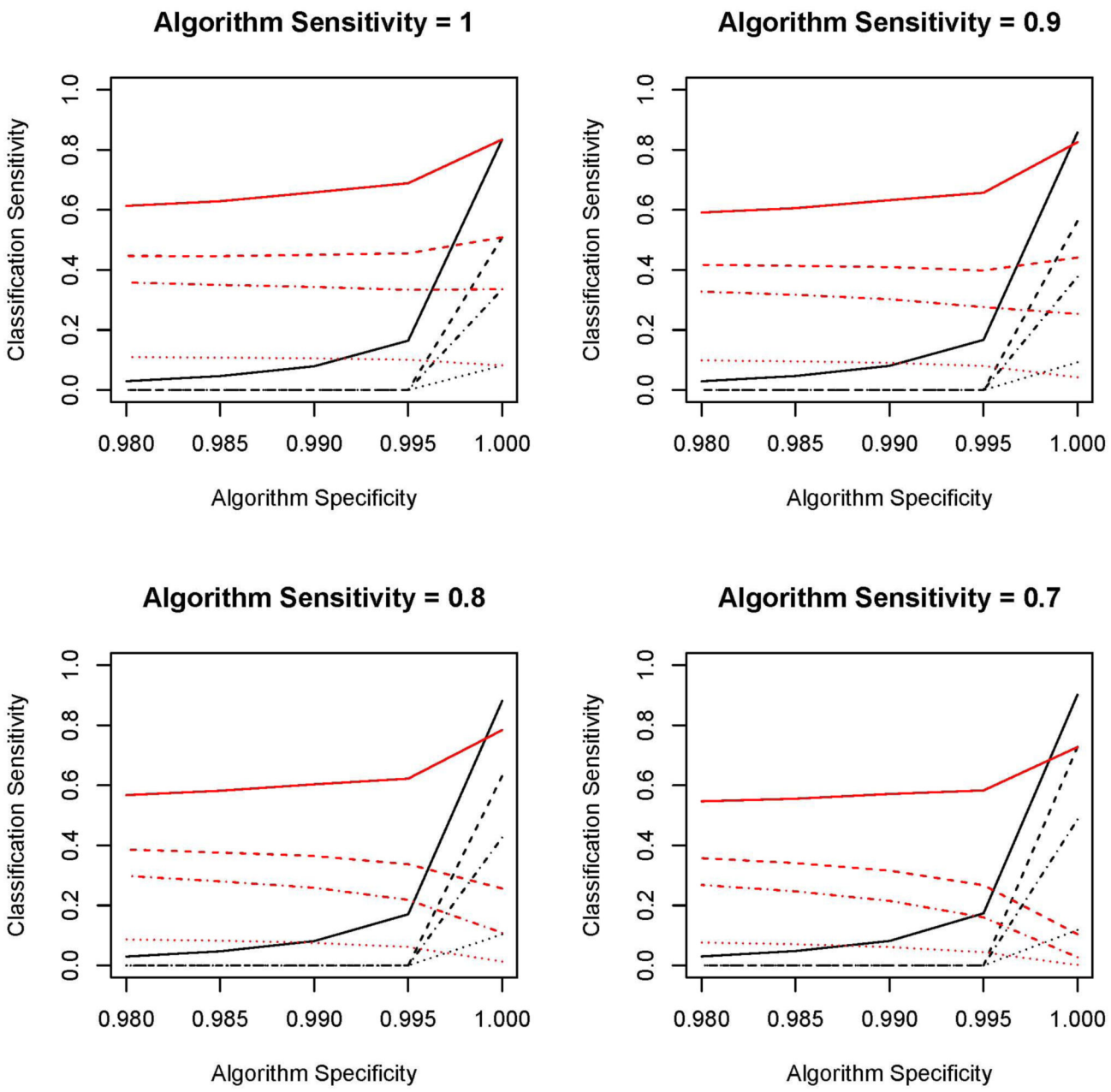

Figure 2.

Simulation results for sensitivity of classification of a provider as failing to meet guideline thresholds vs specificity of the claims-based algorithm used to obtain performance estimates. Reported for four levels of sensitivity of the claims-based algorithm. Black lines represent unadjusted estimates and red lines represent bias-corrected estimates. Solid = maximum likelihood, dashed = Bayesian posterior mean, dotted = maximum likelihood 95% confidence interval, dashed-and-dotted = Bayesian posterior probability.

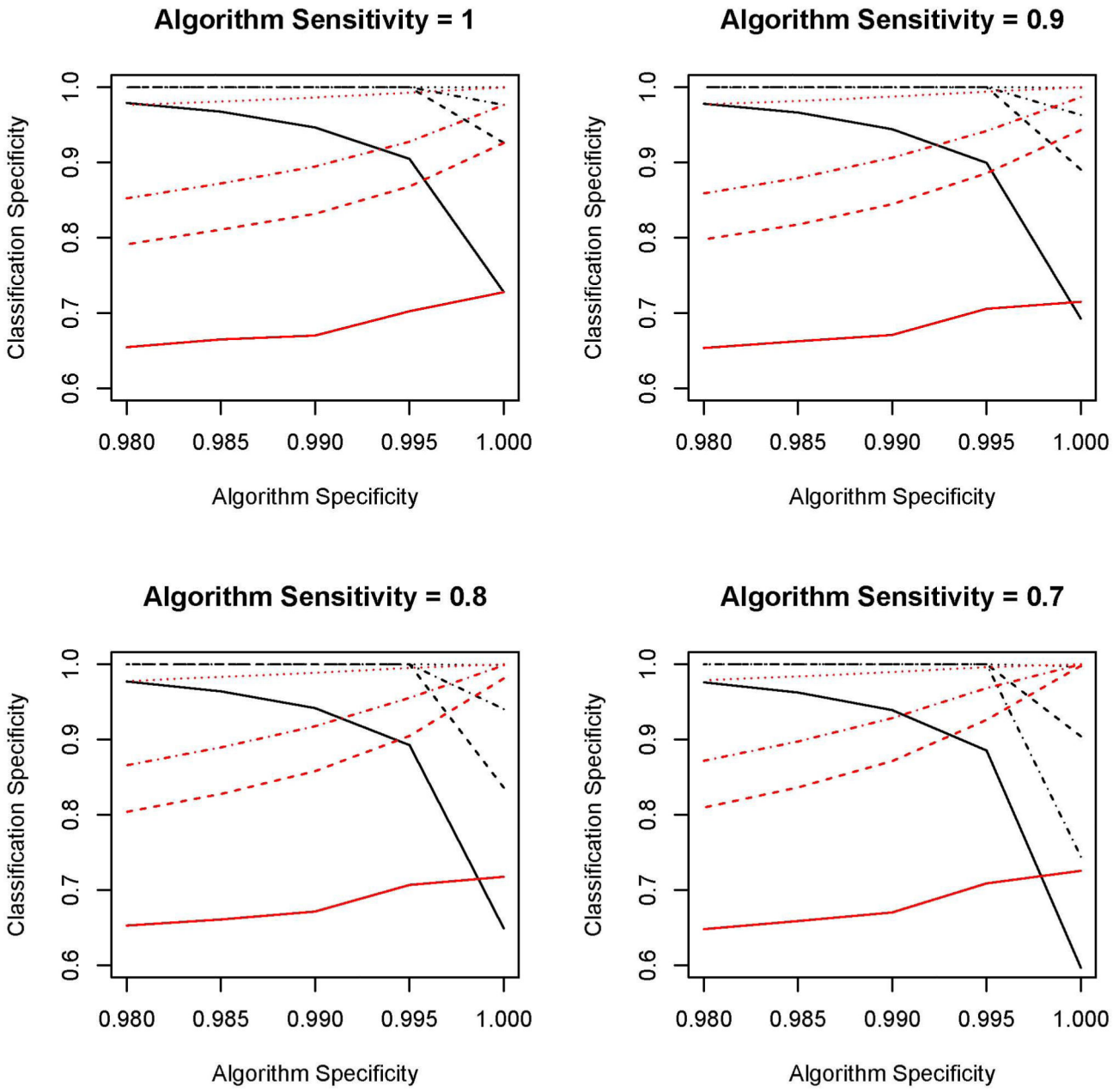

Figure 3.

Simulation results for specificity of classification of a provider as failing to meet guideline thresholds vs specificity of the claims-based algorithm used to obtain performance estimates. Reported for four levels of sensitivity of the claims-based algorithm. Black lines represent unadjusted estimates and red lines represent bias-corrected estimates. Solid = maximum likelihood, dashed = Bayesian posterior mean, dotted = maximum likelihood 95% confidence interval, dashed-and-dotted = Bayesian posterior probability.

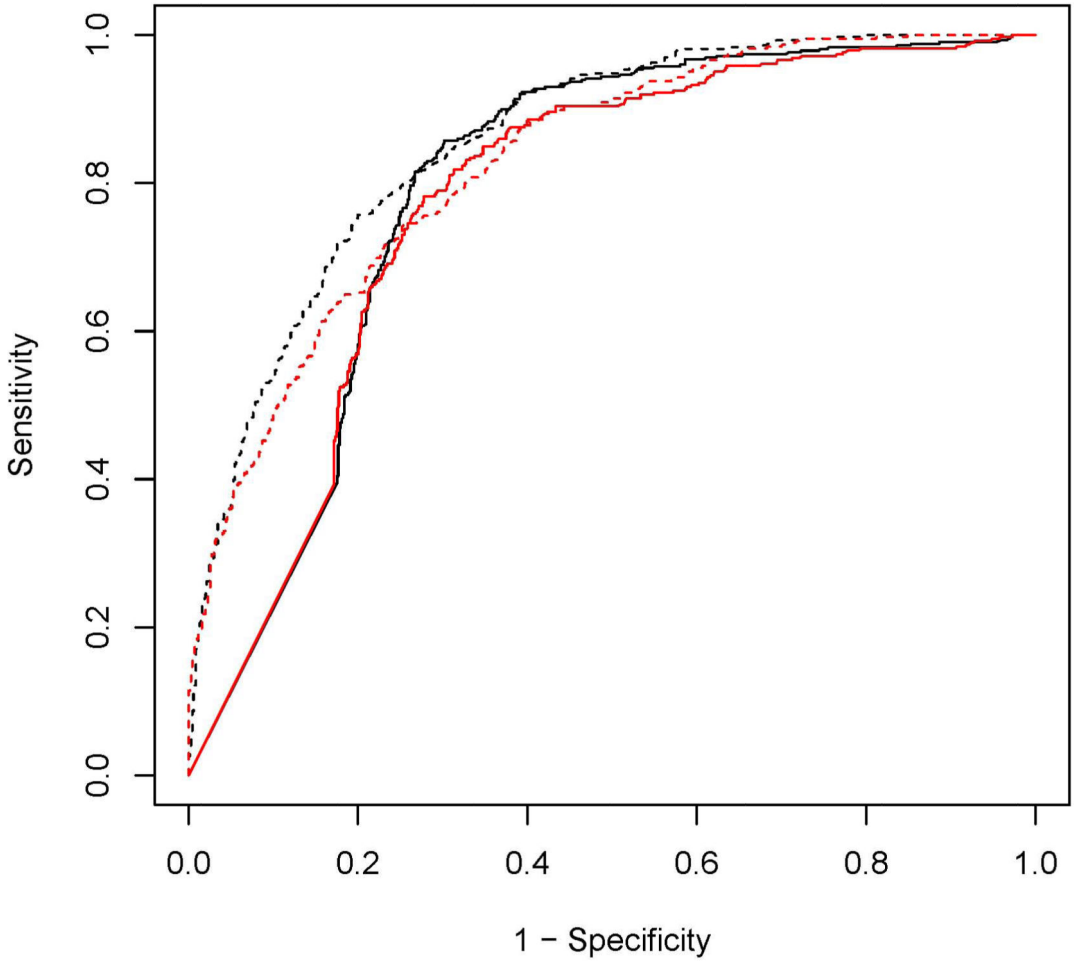

Figure 4.

Receiver operating characteristic curves for unadjusted (black) and adjusted (red) maximum likelihood (solid) and posterior mean (dashed) estimates based on claims-based algorithms with sensitivity and specificity similar to an existing algorithm for cancer detection rate (S = 0.940, P = 0.999).

Table 1

Correlation of BCSC clinical measures and Medicare claims-based measures of cancer detection rate (CDR) estimated using fixed effects maximum likelihood (ML) or Bayesian hierarchical estimation methods. Kappa was computed by comparing agreement of classification of facilities as failing to meet a benchmark of 0.2% cancer detection rate.

	Correlation	Unadjusted Kappa	Adjusted Kappa
ML	0.962	0.902	0.877
ML confidence interval	--	0.958	0.918
Posterior mean	0.937	--	--
Posterior probability	--	--	-