Published in final edited form as:

Stat Med. 2008 May 10; 27(10): 1745–1761. doi:10.1002/sim.3186.

Joint modeling of sensitivity and specificity

Gavino Puggioni^{1,*}, Alan E. Gelfand¹, and Joann G. Elmore²

¹Department of Statistical Science, Duke University, Durham, NC, U.S.A.

²School of Medicine, University of Washington, Seattle, WA, U.S.A.

SUMMARY

Sensitivity and specificity are two customary performance measures associated with medical diagnostic tests. Typically, they are modeled independently as a function of risk factors using logistic regression, which provides estimated functions for these probabilities. Change in these probabilities across levels of risk factors is of primary interest and the indirect relationship is often displayed using a receiver operating characteristic curve. We refer to this as analysis of 'first-order' behavior. Here, we consider what we refer to as 'second-order' behavior where we examine the stochastic dependence between the (random) estimates of sensitivity and specificity. To do so, we argue that a model for the four cell probabilities that determine the joint distribution of screening test result and outcome result is needed. Such a modeling induces sensitivity and specificity as functions of these cell probabilities. In turn, this raises the issue of a coherent specification for these cell probabilities, given risk factors, i.e. a specification that ensures that all probabilities calculated under it fall between 0 and 1. This leads to the question of how to provide models that are coherent and mechanistically appropriate as well as computationally feasible to fit, particularly with large data sets. The goal of this article is to illuminate these issues both algebraically and through analysis of a real data set.

Keywords

sensitivity; specificity; hierarchical models; radiologist variability

1. INTRODUCTION

Two commonly used measures of performance of diagnostic medical procedures are sensitivity and specificity. Customarily they are modeled directly. That is, we explain $\Pr(\text{positive screening result} | \text{occurrence of outcome})$ and $\Pr(\text{negative screening result} | \text{non-occurrence of outcome})$ using binary regression models, introducing risk factors associated with the particular outcome. We obtain estimated functional forms for these probabilities and, as we change the levels of risk factors, we can see how these probabilities change, how they co-vary as functions of the risk factors. Often such dependence is depicted through receiver operating characteristic curves (see, e.g. Pepe [1]), which assist in revealing the nature of the indirect co-variation.

Under a Bayesian perspective, these probabilities are random variables with a posterior distribution for each, at each set of risk factor levels. Then, we could refer to the

investigation of the previous paragraph as 'first order', i.e. we examine say posterior means and the functional dependence between these means.

Our contribution here is to introduce a 'second-order' analysis in this setting. More precisely, at any set of risk factor levels, there is a 2×2 table of screening result by outcome result with entries that are probabilities which sum to 1. In fact, suppose we model these random cell probabilities as functions of the risk factors, we obtain dependent random variables. Since sensitivity and specificity are functions of these cell probabilities, they are dependent as well. It is this stochastic dependence that we seek to learn about here. In particular, we illuminate the nature of this dependence under different stochastic models for the random cell probabilities. Apart from quantifying this stochastic dependence, it may be of interest in order to learn how predictable one is given the other.

The key point is that in order to study this dependence, we need to proceed from a model for the joint cell probabilities. If we build a model for sensitivity and then, separately, a model for specificity, they become independent *by assumption*. Moreover, though we illustrate within the Bayesian framework, the same argument applies to a classical analysis. To study the stochastic dependence between an estimated sensitivity and an estimated specificity, we need to induce them from a model for the joint cell probabilities.

There are many possible ways to jointly model the four cell probabilities. Evidently, only three specifications are needed. Regardless, any such model will be a (nonlinear) reparametrization of any other. However, this does not imply that similar dependence structure will be induced. In its simplest form, if we add a model for the probability of the outcome to the models for sensitivity and specificity we uniquely determine the cell probabilities but impose independence for the latter two probabilities. In fact, if for a given vector of risk factors, X , we denote the three random variables by $Se(X)$, $Sp(X)$, and $O(X)$, then the joint density for these variables takes the form $f_1(Se(X)) f_2(Sp(X) f_3(O(X)))$.

Hence, an important issue that underlies what we are exploring is *coherence*. Clearly, we cannot write down three arbitrary functions of the cell probabilities and then model them. This will not ensure that the resulting four cell probabilities are non-negative and sum to 1. And, if this is not guaranteed, then probabilities computed from these cell probabilities need not be non-negative and less than 1. Hence, a specification is coherent if it does ensure this. Unfortunately, the natural coherent specification can be challenging to fit; an alternative version proves more tractable. We discuss these issues in detail below.

Within the range of medical screening procedures, we are motivated by screening mammography. Screening mammography examinations are used by radiologists to provide a preliminary assessment regarding the presence or absence of breast cancer in healthy, asymptomatic women and therefore help determine whether a patient requires more advanced examination. More accurate diagnosis of this class of tumors requires other procedures such as diagnostic mammograms, biopsy, or ultrasound.

Despite attempts to standardize radiologist performance [2] and to limit the level of subjectivity involved, it is well established that there is a considerable variability in radiologist performance in reading/interpreting film (see [3–7]). There is also a growing literature that attempts to explain [8, 9] and quantify the extent of variability (see [10, 9]). Strategies to compare physicians in a more fair fashion, by taking into account the differences in case mix have been proposed (see [11–13]).

In particular, Woodard *et al.* [14] propose a Bayesian approach to model radiologist performance rates. They focus on the probability of correctly detected cancers at the screening level (sensitivity) and the probability of correctly identified non-cancer cases at

the screening level (specificity), given a collection of patient and radiologist characteristics. As a result, assessment of clinical performance for particular patient types becomes available. The work of Woodard *et al.* employs a retrospective data set to model sensitivity and specificity. In other words, the two performance measures are modeled separately, conditioning on the occurrence (or absence) of cancers. Evidently, if one were to look at the four joint probabilities for screening outcome (+/−) and cancer outcome (presence/absence) as random, these probabilities are dependent, in fact negatively associated since their sum is 1. Then, the induced sensitivity and specificity would be random and would be expected to be dependent as well. Intuitively, why should the unknown (random) probability of recall given cancer present be independent of the unknown (random) probability of recall given cancer absent.

Hence, in the spirit of the above discussion, our contribution is to develop a joint model for sensitivity and specificity through hierarchical specifications that uniquely and coherently determine the cell probabilities in the foregoing 2×2 table. Using two different models we will be able to assess the nature of the dependence between these two performance measures. In particular, we do this for a large screening data set taken from three registries that are part of the Breast Cancer Surveillance Consortium (BCSC), see [15].

The potential clinical relevance of this work is to encourage those who study performance of screening tests to think jointly with regard to explaining screening result and disease outcome. Understanding the behavior of the joint process enables enhanced insight into the joint behavior of the induced sensitivity and specificity.

Thus, the format of the paper is as follows. In Section 2, we briefly review the data set used to infer about the foregoing dependence. In Section 3, we discuss coherent modeling for the joint distribution of screening outcome and disease outcome. In Section 4, we briefly discuss computational issues associated with fitting the models in Section 3. In Section 5, we analyze the data set under two different models presented in Section 3, in particular, with regard to the dependence between sensitivity and specificity.

2. THE DATA SET

The data set we use consists of screening mammograms dated from 1 January 1996 to 31 December 2002. In particular, three registries that are members of the BCSC have been selected: the Colorado Mammography Program, the New Hampshire Mammography Network, and the Breast Cancer Surveillance Project at Group Health in the state of Washington. Each BCSC registry and the statistical coordinating center for the consortium have received IRB approval for either active or passive consenting processes or a waiver of consent to enroll participants, link data, and perform analytic studies. All procedures are HIPAA compliant and all registries and the statistical coordinating center have received a Federal Certificate of Confidentiality that protects the identities of research subjects.

For each examination, standardized information is collected on the women's risk factors for breast cancer, the radiologist identity, the radiologist's assessment on the examination, and the outcome of breast cancer diagnosis within one year of the examination. Radiologists provide an assessment for each examination using the BI-RADS® categories [2]. This classification features outcome categories for screening examinations labeled from 0 to 5. In order to define binary outcomes we consider as positive all the mammograms labeled as 0 (needs additional evaluation), 3 (probably benign but with recommended immediate work-up), 4 (suspicious abnormality), and 5 (highly suggestive of malignancy). Negative mammograms are those with an assessment of 1 (normal), 2 (benign), or 3 (probably benign) with no immediate work-up required. In the context of screening mammography, a positive mammogram is referred to as a *recall* and so, in the sequel, we suggestively write R

=1,0 to indicate recall/no recall result. Similarly, we write $C=1,0$ to indicate cancer/no cancer outcome.

We chose to include only the patient risk factors for breast cancer considered by Woodard *et al.* [14]. In this regard, previous studies, for instance, [16, 17], have already shown that factors such as woman's age, breast density, hormone therapy, family history, and menopausal status have an impact on sensitivity and/or specificity. We have a total of 448 083 mammography examinations after the exclusion of the patients with incomplete or missing information. The positive examinations (assessments), number 46 677 (10.4 per cent), are referred to as recalls because these women are called back for further testing. The number of cancer outcomes is 2358 (0.52 per cent incidence rate). More details and an indication of the overall empirical sensitivity and specificity can be found in Table I.

Data from a radiologist survey was used in order to model radiologist random effects level. More precisely, a total of 181 radiologists were contacted via a mail survey to provide data about their clinical practice, demographic attributes, academic affiliation, years of experience, concern about malpractice, etc. The rate of response, 139 out of 181, is quite high for surveys of physicians. However, as in Woodard *et al.* [14] we only include those radiologists who read more than 480 mammograms. The radiologist characteristics we employed are the same as in Woodard *et al.* 14 and are reported in Table II.

3. MODELING

3.1. Modeling the joint distribution

Woodard *et al.* propose a two-stage hierarchical generalized linear model to model separately sensitivity $P(R_{ij}=1|C_{ij}=1)$ and specificity $P(R_{ij}=0|C_{ij}=0)$. In particular, they set

$$\text{logit}\left(\Pr(R_{ij}=1|C_{ij}=1)\right) = \mathbf{X}'_{ij}\beta_i^{(1)} + \tau_i^{(1)} \quad (1)$$

$$\text{logit}\left(\Pr(R_{ij}=0|C_{ij}=0)\right) = \mathbf{X}'_{ij}\beta_i^{(0)} + \tau_i^{(0)} \quad (2)$$

where $R_{ij}=1$ and $C_{ij}=1$ indicate recall and cancer for mammogram (i, j) , $i=1, \dots, I$ indexes the radiologist, $j=1, \dots, n_i^{(k)}$, $k \in \{0, 1\}$ the patient associated with the i th radiologist, and $\mathbf{X}'_{ij}$ the patient design vector. The second stage involves modeling the random effect for the i th radiologist as

$$\tau_i^{(k)} \stackrel{iid}{\sim} \mathcal{N}(\mathbf{W}_i\gamma^{(k)}, \sigma^{2(k)}), \quad k \in \{0, 1\} \quad (3)$$

Then, $\mathbf{W}_i$ is the vector containing the radiologist characteristics.

From (1) to (3), it is evident that sensitivity and specificity are random variables. As noted in the Introduction, the retrospective modeling in (1) and (2) implicitly assumes that sensitivity and specificity are independent. We consider modeling that enables us to capture anticipated dependence between them. We do this by inducing a joint distribution for inference about sensitivity and specificity rather than two marginal distributions. Such an analysis is apart from studying how, say, the posterior mean sensitivity and posterior mean specificity co-vary as functions over different levels of the risk factors (X 's and W 's).

We obtain a joint model for sensitivity and specificity induced by the specification of a joint model for C_{ij} and R_{ij} , i.e. for

$$\Pr(C_{ij}=h, R_{ij}=k), \quad h, k \in \{0, 1\} \quad (4)$$

Here, we note that if we model $\Pr(C_{ij}=h)$ in addition to (1) and (2), we do induce a model for (4) but with imposed independence of sensitivity and specificity. However, there are many ways to build a model for (4). Under modeling that allows dependence, it will be interesting to see how dependence varies across levels of risk factors and across radiologists.

Perhaps the most natural approach to model the cells of the joint distribution in (4) is to use a multinomial logit

$$\begin{aligned} &\Pr(C_{ij}=1, R_{ij}=1) \\ &\Pr(C_{ij}=1, R_{ij}=0) \\ &\Pr(C_{ij}=0, R_{ij}=1) \end{aligned} \quad (5)$$

relative to the baseline probability $\Pr(C_{ij}=0, R_{ij}=0)$. However, such models with two stage specifications analogous to (1)–(3) are very demanding to fit; the resulting likelihood is a highly nonlinear form and Markov chain Monte Carlo (MCMC) model fitting is usually poorly behaved. This is not a *Bayesian computation* problem. Obtaining MLE's and associated standard errors is, arguably, an even more difficult computational challenge.

Instead, we consider two different specifications (which we denote as Models 1 and 2) that uniquely determine the distribution in (4). Their advantage is that each provides three separate, easy-to-fit submodels. In each case, putting the three together allows us to calculate sensitivity and specificity. We refer to the first as the 'chronological model', in accordance with the assumption that cancer outcome is observed after screening outcome. Simpler versions than the model below have been discussed in Paliwal *et al.* [17]. We have Model 1:

$$\begin{aligned} &\Pr(C_{ij}=1|R_{ij}=1) \\ &\Pr(C_{ij}=1|R_{ij}=0) \\ &\Pr(R_{ij}=1) \end{aligned} \quad (6)$$

Simple algebra shows that these three probabilities uniquely determine the probabilities in (4). See Section 3.2 for details.

An alternative specification, less intuitive but also straightforward to fit, is Model 2:

$$\begin{aligned} &\Pr(C_{ij}=0|R_{ij}=0) \\ &\Pr(R_{ij}=0 | (R_{ij}=0) \cup (R_{ij}=1, C_{ij}=0)) \\ &\Pr((R_{ij}=0) \cup (R_{ij}=1, C_{ij}=0)) \end{aligned} \quad (7)$$

For each of the different submodels in either (6) or (7) we fit a two-stage logistic regression similar to the ones described above, i.e.

$$\text{logit}(\Pr(\cdot)) = \mathbf{X}'_i \beta_i^{(k)} + \tau_i^{(k)} \quad (8)$$

$$\tau_i^{(k)} \stackrel{iid}{\sim} \mathcal{N}(\mathbf{W}_{iy^{(k)}} \gamma^{(k)}, \sigma^{2(k)}) \quad (9)$$

where k is an indicator of the submodel we are estimating.

For each of the models in either (6) or (7) the likelihood is the following:

$$L(\beta^{(k)}, \{\tau_i^{(k)}\}, \gamma^{(k)}, \sigma^{2(k)}; \cdot) = \prod_{i=1}^I \left[\prod_{j=1}^{n_i^{(k)}} p_{ij}(\beta^{(k)}, \tau_i^{(k)})^{r_{ij}^{(k)}} \times (1 - p_{ij}(\beta^{(k)}, \tau_i^{(k)})^{1-r_{ij}^{(k)}} \right] \times \mathcal{N}(\tau_i^{(k)}; \mathbf{W}'_i \gamma^{(k)}, \sigma^{2(k)})$$

$$p_{ij}(\beta^{(k)}, \tau_i^{(k)}) = \Pr(\cdot | \beta^{(k)}, \tau_i^{(k)}) = \frac{\exp(\mathbf{X}'_{ij} \beta^{(k)} + \tau_i^{(k)})}{1 + \exp(\mathbf{X}'_{ij} \beta^{(k)} + \tau_i^{(k)})}$$

where $r_{ij}^{(k)}$ indicate the outcome for the (i, j) th observation of the event with the probability that we estimate with model k . Care is needed for prior specification. In fact, as noted, for instance, in [18], flat priors in a logistic model with random effects induce identifiability problems and improper posterior distributions. Thus, as in [14], we used proper but fairly weak, independent normal priors with mean 0 and variance 10 for all the regression coefficients (at both patient and radiologist levels). Previous studies [8, 14, 17] showed that such parameters were contained in confidence intervals not wider than $(-3, 3)$; hence, the choice of a variance of 10 should not be very informative. For the parameter σ^2 we use an inverse-Gamma prior with shape 2 and scale 3, also weak in the sense that it provides an infinite prior variance for σ^2 . In conjunction with our analysis of the roughly 450 000 mammograms (Tables I–III) in Section 5, we ran a small prior robustness study relative to the above choices, revealing essentially no sensitivity in our specificity and sensitivity findings in that section.

3.2. Obtaining sensitivity and specificity

After the estimation of each single submodel we construct the nonlinear expressions for sensitivity and specificity. In particular, under (6) we obtain

$$\begin{aligned} \text{Se}_1 &= P(R=1|C=1) = \frac{P(C=1|R=1)P(R)}{P(C=1|R=1)P(R) + P(C=1|R=0)(1-P(R))} \\ \text{Sp}_1 &= P(R=0|C=0) = \frac{(1-P(C=1|R=0))(1-P(R=1))}{(1-P(C=1|R=0))(1-P(R=1)) + (1-P(C=1|R=1))P(R=1)} \end{aligned}$$

and for (7) we have

$$\begin{aligned} \text{Se}_2 &= P(R=1|C=1) = \frac{P(C=1, R=1)}{P(C=1, R=1) + P(C=1, R=0)} \\ \text{Se}_2 &= P(R=0|C=0) = \frac{P(C=0, R=0)}{P(C=0, R=0) + P(C=0, R=1)} \end{aligned}$$

where

$$\begin{aligned}
P(C=1, R=1) &= 1 - P(R=0 \cup (R=1, C=0)) \\
P(C=1, R=0) &= (1 - P(C=0|R=0)) \times P(R=0|R=0 \cup (R=1, C=0)) \times (R=0 \cup (R=1, C=0)) \\
P(C=0, R=0) &= P(C=0|R=0) \times P(R=0|R=0 \cup (R=1, C=0)) \\
P(C=0, R=1) &= (1 - P(R=0|R=0 \cup (R=1, C=0))) \times P(R=0 \cup (R=1, C=0))
\end{aligned}$$

3.3. Coherence

While there are many ways to specify three logistic regression models to determine the probabilities in (4), attention should be devoted to checking the probabilistic coherence of such models. For instance, let us consider the following specification using the two marginal probabilities and one of the cell probabilities, i.e.

$$\begin{aligned}
&\Pr(C=1) \\
&\Pr(R=1) \\
&\Pr(C=h, R=k) \quad h, k \in \{0, 1\}
\end{aligned} \tag{10}$$

Let us fix $h=0$ and $k=1$ and assume distinct two-stage logistic specifications of forms (8) and (9) for each of these probabilities. The expressions for sensitivity (Se_3) and specificity (Sp_3) as a function of (10) become

$$Se_3 = \frac{\Pr(R=1) - \Pr(C=0, R=1)}{\Pr(C=1)} \tag{11}$$

$$Sp_3 = \frac{1 - \Pr(C=1) - \Pr(C=0, R=1)}{1 - \Pr(C=1)} \tag{12}$$

Clearly, such a specification is incoherent; the distinct models for the numerator probabilities in (11) and (12) cannot ensure that the numerators themselves are positive. Moreover, the denominator, if small enough can make these expressions greater than 1. The same incoherence could arise using, say, two marginal probabilities and the odds ratio.

Also, suppose we return to a multinomial logit model where we specify *individual* logistic models of the type in (8) and (9) for each of

$$\Pr(C_{ij}=1, R_{ij}=1)$$

$$\Pr(C_{ij}=1, R_{ij}=0)$$

$$\Pr(C_{ij}=0, R_{ij}=0)$$

Unlike the multinomial logit model in (5), this is also not a coherent specification; nothing controls the sum of these three probabilities.

In summary, the safest way to avoid coherence concerns is to jointly model the probabilities in (5). Unfortunately, such models are difficult to fit. The model in (6) presents both a coherent and easy-to-fit specification.

4. EFFICIENT MCMC MODEL FITTING

Both the hierarchical models discussed above can be fitted using Gibbs sampling. Closed-form full conditional distributions for the patient regression coefficients and for the radiologist random effect are not available. The standard Metropolis–Hastings (M–H) algorithm, a popular choice in such instances, displayed some convergence problems. With approximately 15 000 observations, Woodard *et al.* [14] employed the adaptive rejection sampling (ARS) algorithm [19]. Although it produces exact sampling from full conditional distributions, we do not recommend its implementation with large data sets such as the roughly 450 000 mammograms we work with. It is very slow and cumbersome.

We recommend the Langevin algorithm as an alternative to the basic random walk–H [20, 21]. It provides a ‘directed’ M–H algorithm, proposing values that take advantage of gradient information in the full conditional distributions that need to be sampled. In particular, we create the random walk-like transition, $x^{(t+1)} = x^{(t)} + (\nu^2/2) \nabla \log f(x^{(t)}) + \nu \varepsilon_t$ where $\varepsilon_t \sim N_p(0, I_p)$ and ν^2 is a tuning parameter. In the literature this sampler is known as the Metropolis-adjusted Langevin algorithm (MALA). Given a current $x^{(t)}$, we accept a new Y_t with probability

$$\frac{f(Y_t) \exp\left\{\left\|Y_t - x^{(t)} - \frac{\nu^2}{2} \nabla \log f(x^{(t)})\right\|^2 / 2\nu^2\right\}}{f(x^{(t)}) \exp\left\{\left\|x^{(t)} - Y_t - \frac{\nu^2}{2} \nabla \log f(x^{(t)})\right\|^2 / 2\nu^2\right\}} \wedge 1 \quad (13)$$

The addition of the gradient of $\log f$ is relevant because it improves the movement toward the modes of f without the need to construct an entire envelope as in ARS. In practice, MALA is no more difficult to implement than the usual M–H updates since the required gradients are straightforward to calculate and to code. We use it in order to update patient coefficients and the random effects coefficients for each model, that is, respectively, $\beta^{(k)}$ and $\tau_i^{(k)}$.

5. RESULTS OF THE ANALYSIS

In this section, we offer an application of the methodology outlined above. That is, we focus on second-order behavior under both Models 1 and 2. Under each model, at each risk factor profile, for each radiologist, there is a joint posterior distribution for sensitivity and specificity, altogether more than 240 000 distributions. Every patient can be studied using the posterior samples from the model fitting but we can only present some illustrative summaries.

First, we examine how the two models compare on first-order analysis. In particular, we obtain the posterior sensitivity and specificity for a hypothetical mammography exam with a particular profile of patient risk factors. In fact, we consider a ‘typical’ patient, i.e. one with the most common levels of the risk factors. Such a reference female has no history of breast biopsy or surgery or family history of breast cancer, has age between 50 and 59, is post-menopausal, is not using hormone replacement therapy, has density breast classification 2, and has no reported symptoms. In Figure 1(a) the posterior mean sensitivity for each of the 120 radiologists is shown under each of the models. In Figure 1(b), we have an analogous plot for specificity. The models agree quite well with regard to specificity but a little less so

with regard to sensitivity. The latter is perhaps expected due to the sparsity of observed cancers even in our large data set. Agreement is anticipated since both models are explaining the same joint cell probabilities. Although these probabilities (and the induced sensitivities and specificities) arise as different parametrizations under the two models, the same class of logistic hierarchical specifications is used in each case.

It is perhaps more illuminating to look at whole posteriors for this typical patient. We do so in Figures 2 (for sensitivity) and Figure 3 (for specificity) for nine randomly chosen radiologists. We note three points: (i) these posteriors do not appear to be very model sensitive; (ii) the posteriors for specificity are much tighter than those for sensitivity, expected again since there are few cancer cases; and (iii) agreement is stronger for radiologists seeing more cases (number of cases is indicated above each panel), again as expected.

Turning to second-order behavior, we look at the joint posterior for sensitivity and specificity under each of the models, still for the typical patient. For the same nine radiologists, we present this in Figure 4 for Model 1 and in Figure 5 for Model 2. We see variability across radiologists and, also, some differences between the models. In fact, under Model 1, across all 120 radiologists the correlations range from -0.692 to 0.211 . Under Model 2, across all 120 radiologists the correlations range from -0.305 to 0.225 . We can conclude that second-order behavior appears to be more model sensitive than first-order behavior. Figure 6 focuses solely on the correlations associated with these joint distributions summarizing the correlations for the 120 radiologists under each model with histograms. Model 1 produces generally smaller (more negative) correlations than Model 2. However, of related interest is that we have another manifestation of radiologist variability.

Table I reveals that there are more than 1000 different risk factor profiles. To glimpse the effect of risk factor profile on second-order behavior we identified a high-risk and a low-risk patient. The high-risk patient presents the most likely risk factor levels with regard to a positive screen, and the low-risk patient presents the most unlikely levels. Table III summarizes, again using the same nine radiologists. We see a considerable variation, again across models, again across radiologists and, in addition, across the selected patient types.

6. SUMMARY AND DISCUSSION

Treating unknown probabilities in a collection of 2×2 tables as random, the induced sensitivity and specificity become random variables. We have discussed the importance of coherence in specifying models for these cell probabilities and proposed two models that ensure coherent cell probabilities. We have distinguished the familiar first-order analysis of sensitivity and specificity from second-order analysis. With regard to first-order inference similar behavior is seen under both models but with regard to second-order behavior we have found that: (i) there is predictability of sensitivity from specificity and *vice versa*; (ii) the association tends to be negative; (iii) it is more strongly negative under Model 1; and (iv) it varies widely across radiologist/patient combinations and can be extreme (as strong as roughly -0.7).

Our approach can be extended to screening procedures having two diagnostic stages—a first assessment and then a follow-up assessment after further diagnostic work. We would now obtain a collection of $2 \times 2 \times 2$ tables. Respecting the chronology, we can now define various notions of sensitivity and specificity; coherence in modeling is still a central notion; under any coherent specification, again, stochastic dependence between them can be studied. Similarly, settings with two (in fact multiple) screening procedures or settings with two (or more) radiologists reading the same film can be examined. Finally, there may be interest in the temporal evolution of sensitivity and specificity for a radiologist. Introducing a temporal

component to our modeling would allow us to learn if (and how) dependence between them is changing over time.

Acknowledgments

Data collection for this work was supported by an NCI-funded Breast Cancer Surveillance Consortium co-operative agreement (U01CA86076, U01CA63731, U01CA86082, U01CA63736) and by public service grant R01 HS-010591 from the Agency for HealthCare Research and Quality and National Cancer Institute grants R01 CA107623 and K05 CA 104699. We thank the BCSC investigators, participating 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/>.

Contract/grant sponsor: NCI-funded Breast Cancer Surveillance Consortium Co-operative Agreement; contract/grant numbers: U01CA86076, U01CA63731, U01CA86082, U01CA63736 Contract/grant sponsor: Agency for HealthCare Research; contract/grant number: R01 HS-010591 Contract/grant sponsor: Quality and National Cancer Institute; contract/grant numbers: R01 CA107623, K05 CA 104699

APPENDIX A: PRIOR AND FULL CONDITIONAL DISTRIBUTIONS

Again, each of the three-component models in either (6) or (7) is specified using a model as in (8) and (9). Each can be fitted separately (perhaps, in parallel) with a likelihood as given below (9). Again, the priors are

$$\beta \sim N(\beta_0, \sigma_\beta^2 I_p), \quad \beta_0=0, \quad \sigma_\beta^2=10$$

$$\gamma \sim N(\gamma_0, \sigma_\gamma^2 I_q), \quad \gamma_0=0, \quad \sigma_\gamma^2=10$$

with

$$(\tau_i | \sigma^2) \sim N(\mathbf{W}_i' \gamma, \sigma^2)$$

$$\sigma^2 \sim \mathcal{IG}(a, b)$$

Indicating with $\mathcal{D}$ the data we are using in a given submodel we have the following full conditional distributions:

$$[\beta | \tau, \gamma, \sigma^2, \mathcal{D}] \propto N(\beta_0, \sigma_\beta^2 I_p) \times \prod_{i=1}^I \left[\prod_{j=1}^{n_i} p_{ij}(\beta, \tau_i)^{r_{ij}} (1 - p_{ij}(\beta, \tau_i)^{1-r_{ij}}) \right]$$

$$[\tau_i | \beta, \gamma, \sigma^2, \mathcal{D}] \propto N(0, \sigma^2) \left[\prod_{j=1}^{n_i} p_{ij}(\beta, \tau_i)^{r_{ij}} (1 - p_{ij}(\beta, \tau_i)^{1-r_{ij}}) \right]$$

$$[\gamma|\beta, \tau_i, \sigma^2, \mathcal{D}] = \mathcal{N}\left(\mathbf{v}\left(\frac{\gamma_0}{\sigma_y^2} + \frac{\mathbf{W}'\boldsymbol{\tau}}{\sigma^2}\right), \mathbf{v}\right), \quad \mathbf{v} = \left(\frac{\mathbf{I}_q}{\sigma_y^2} + \frac{\mathbf{W}'\mathbf{W}}{\sigma^2}\right)^{-1}$$

$$[\sigma^2|\beta, \tau_i, \gamma, \mathcal{D}] = \mathcal{IG}\left(a + \frac{I}{2}, b + \frac{I}{2} + \frac{1}{2} \sum_{i=1}^I (\tau_i - \mathbf{W}_i'\boldsymbol{\gamma})^2\right)$$

An iteration of the Gibbs sampler proceeds with a MALA update of β , then of the τ_i , as described in Section 4. We then update γ with a normal draw followed by an update of σ^2 with an inverse Gamma draw. For both of the models the MCMC is well behaved. The individual models are run for roughly 6000 iterations to burn-in, followed by roughly 12 000 iterations, no thinning required, to achieve a posterior sample used to obtain the posterior summaries.

REFERENCES

1. Pepe, MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press; Oxford: 2003.
2. American College of Radiology. Breast Imaging Reporting and Data System (BI – RADS® Atlas). American College of Radiology; Reston, VA: 2003.
3. Mammography Quality Standard Act. Publication L; October. 1992 p. 102-539.
4. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International variation in screening mammography interpretations in community-based programs. *Journal of the National Cancer Institute*. 2002; 95:1384–1393. [PubMed: 13130114]
5. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretation of mammograms. *The New England Journal of Medicine*. 1994; 331:1493–1499. [PubMed: 7969300]
6. Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. *Journal of the American Medical Association*. 1993; 270:2444–2450. [PubMed: 8230621]
7. Smith-Bindman R, Chu PW, Miglioretti DL, Sickles EA, Blanks R, Ballard-Barbash E, et al. Comparison of screening mammography in the United States and in the United Kingdom. *Journal of the American Medical Association*. 2003; 290:2129–2137. [PubMed: 14570948]
8. Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi CJ, Cutter G, Hendrick RE, Elmore JG. Accuracy of screening mammography interpretation by characteristics of radiologists. *Journal of the National Cancer Institute*. 2004; 96:1840–1850. [PubMed: 15601640]
9. Elmore JG, Miglioretti DL, Reisch LM, Barton MB, Kreuter W, Christiansen CL, Fletcher SW. Screening mammograms by community radiologists: variability in false positive rates. *Journal of the National Cancer Institute*. 2002; 94:1373–1380. [PubMed: 12237283]
10. Christiansen CL, Wang F, Barton MB, Kreuter W, Elmore JG, Gelfand AE, Fletcher SW. Predicting the cumulative risk of false-positive mammograms. *Journal of the National Cancer Institute*. 2000; 92:1656–1666.
11. Iezzoni LI, Greensberg LG. Widespread assessment of risk-adjusted outcomes: lessons from local initiatives. *Joint Commission Journal on Quality Improvement*. 1994; 20:305–316. [PubMed: 8087195]
12. Landon BE, Iezzoni LI, Ash AS, Shwartz M, Daley J, Hughes JS, Mackiernan YD. Judging hospitals by severity-adjusted mortality rates: the case of CABG surgery. *Inquiry*. 1994; 33:155–166. [PubMed: 8675279]

13. Salem-Schatz S, Morre G, Rucker M, Pearson SD. The case for case-mix adjustment in practice profiling: when good apples look bad. *Journal of the American Medical Association*. 1994; 272:871–874. [PubMed: 8078165]
14. Woodard DB, Gelfand AE, Barlow WE, Elmore JG. Performance assessment for radiologists interpreting screening mammography. *Statistics in Medicine*. 2007; 26(7):1532–1551. [PubMed: 16847870]
15. Ballard-Barbash R, Taplin SH, Yankaskas BC, Ernster VL, Roseberg RD, Carney PA, Barlow WE, Geller K, Kerlikowske MB, 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. *American Journal of Roentgenology*. 1997; 169:1001–1008. [PubMed: 9308451]
16. Carney PA, Miglioretti DL, Yankaskas BC, Kerlikowske K, Rosenberg R, Rutter CM, Geller BM, Abraham LA, Taplin SH, Dignan M, Cutter G, Ballard-Barbash R. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Annals of Internal Medicine*. 2003; 138:168–175. [PubMed: 12558355]
17. Paliwal P, Gelfand AE, Abraham L, Barlow WE, Elmore JG. Examining accuracy of screening mammography using an event order model. *Statistics in Medicine*. 2006; 25:267–283. [PubMed: 16381074]
18. Gelfand AE, Sahu S. Gibbs sampling, identifiability and improper priors in generalized linear mixed models. *Journal of the American Statistical Association*. 1999; 94:247–253.
19. Gilks WR, Wild P. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*. 1992; 41:337–448.
20. Besag JE. Discussion of “Markov chains for exploring posterior distributions”. *Annals of Statistics*. 1994; 22:1734–1741.
21. Stramer O, Tweedie RL. Langevin-type models II: self targeting candidates for Hastings–Metropolis algorithms. *Methodology and Computing in Applied Probability*. 1999; 1:307–328.

Figure 1. Posterior mean sensitivity (a) and specificity (b) for a reference patient (see text for details) across radiologists.

Figure 2.
Posterior density of sensitivity for nine randomly selected radiologists.

Figure 3.
Posterior density of specificity for nine randomly selected radiologists.

Figure 4. Joint distribution of sensitivity and specificity under Model 1 for nine randomly selected radiologists.

Figure 5. Joint distribution of sensitivity and specificity under Model 2 for nine randomly selected radiologists.

Figure 6. Histogram of correlation between sensitivity and specificity under Model 1 (a) and Model 2 (b) across all radiologists.

Table I

Summary of patient characteristics and associated cancer outcomes and screening mammography assessments.

Variable	Cancer group	Total assessment	Subsequent assessment	Positive assessment
Age	40–49	145147	446	17319
	50–59	141063	691	15203
	60–69	83193	569	7809
	70 and older	78680	652	6346
Family history of breast cancer	No	379366	1880	39509
	Yes	69717	478	7168
History of breast biopsy/surgery	No	364210	1757	36309
	Yes	83873	601	10368
Breast cancer symptoms	No symptom	432134	2148	44085
	Lump/nipple discharge/other	15949	210	2592
Menopause/hormone therapy (HT)	Pre-menopausal	125471	451	15448
	Post-menopausal, no HT	156073	938	17046
	Post-menopausal, HT	146325	844	12504
	Post-menopausal, unknown HT	20214	125	1679
Previous mammogram	No	18617	115	3256
	Yes	429466	2243	43421
Breast parenchymal density	Almost entirely fatty	48198	139	2286
	Scattered fibroglandular tissue	195097	920	17869
	Heterogenously dense	167892	1081	22026
	Extremely dense	36896	218	4496
Total		448083	2358	46677

Table II

Summary of radiologist characteristics and associated performance.

Variable	Group	Total	Sensitivity	Specificity
Number of years interpreting mammograms	<10	26	86.0	88.8
	10–19	56	82.8	89.4
	20+	38	78.8	91.6
Number of mammograms interpreted per year	≤1000	31	73.3	92.6
	1001–2000	45	82.7	89.1
	>2000	44	84.2	89.9
Academic affiliation	Primary appointment	6	78.0	93.9
	Adjunct appointment	12	83.6	89.4
	None	102	82.5	89.8
Total		120	82.4	90.0

Table III

Correlation between sensitivity and specificity for nine randomly selected radiologists.

Patient	Rad 1	Rad 2	Rad 3	Rad 4	Rad 5	Rad 6	Rad 7	Rad 8	Rad 9
<i>Model 1</i>									
Typical	0.178	-0.142	-0.067	-0.337	-0.297	0.185	-0.118	-0.076	-0.144
High risk	0.023	-0.340	-0.025	-0.266	-0.460	-0.173	-0.179	0.249	-0.525
Low risk	-0.091	-0.180	0.025	-0.067	-0.213	0.074	-0.254	0.048	-0.383
<i>Model 2</i>									
Typical	-0.254	-0.231	-0.065	0.097	-0.339	-0.012	-0.379	-0.189	-0.508
High risk	0.125	0.040	-0.088	-0.131	-0.029	0.192	0.021	0.184	-0.282
Low risk	-0.016	-0.052	0.0954	0.093	-0.078	0.033	-0.217	-0.030	-0.005