Published in final edited form as:

Stat Med. 2006 January 30; 25(2): 267–283. doi:10.1002/sim.2220.

Examining accuracy of screening mammography using an event order model

Prashni Paliwal^{1,*†}, Alan E. Gelfand², Linn Abraham³, William Barlow³, and Joann G. Elmore³

¹Women's Health Research, Yale University, P.O. Box 208091, New Haven, CT 06520-8091, U.S.A.

²Institute of Statistics and Decision Sciences, P.O. Box 90251, Duke University, NC 27708-0251, U.S.A.

³Center for Health Studies, Group Health Cooperative, 1730 Minor Ave, Seattle, WA 98101, U.S.A.

SUMMARY

Screening mammography is a widely used method for breast cancer detection. For each mammogram we propose a performance model based on order of outcomes. That is, we envision an initial assessment, a follow up assessment if the initial one is positive and, eventually, a determination of whether cancer was present or not. A model can be built at each stage reflecting effects due to patient characteristics, to the facility where mammogram was performed and to the radiologist reading the mammogram. Since assessment is not perfectly associated with outcome, familiar rates of agreement and disagreement are of interest. These rates can be investigated at various levels of risk factors of interest. The approach is illustrated with screening mammography data from the Group Health Cooperative in Seattle, WA. A Bayesian framework is adopted for inference and an analysis of the data set is presented.

Keywords

breast cancer; false positive/negative; logistic regression; model selection; sensitivity; specificity

1. INTRODUCTION

Screening mammography is a widely used method for breast cancer detection in the United States. In fact, more than 70 per cent of women aged 40 years or older reported having at least one mammogram within the past two years [1]. The U.S. Preventive Services recommends that females have annual mammograms beginning at age 40 [2]. While mammography is invaluable in enabling early treatment of cancer, thereby increasing the chance of patient survival, mammogram assessment is not perfectly associated with true cancer state. Hence, a variety of rates measuring agreement or disagreement are of interest. These rates can be prospective or retrospective and include standard notions such as sensitivity, specificity, false positive rate, false negative rate, positive predictive value and

negative predictive value. These rates can be investigated at various levels of aggregation according to which patient characteristics we wish to control for and at what level.

Here we propose a general approach for examining such rates based upon modelling which reflects the order of events associated with each screening mammogram. (See Reference [3] in this regard.). In particular, we assume that within a facility at the time the screening mammogram is taken there is a first event which is an initial assessment by the radiologist that is classified as positive or negative. For those receiving a positive assessment, further work-up is usually carried out leading to a second event which results in a final assessment that, again, is either positive or negative. For those receiving a negative initial assessment, the final assessment is automatically negative. Finally, using a window of twelve months post-mammogram, we arrive at a third event which is an outcome determination as to whether or not breast cancer is present.[‡] Hence, we envision prospective data collection and, associated with each mammogram, is a three-dimensional response vector, recording the results of the three stages. Risk factors for each female are recorded at the time of the screening mammogram. These include characteristics associated with the woman receiving the mammogram, an indicator of previous mammography, an indicator of the facility at which mammogram was taken and also an indicator of which radiologist interpreted the film. The modelling is built in event order. The effects of the patient characteristics, the facility and the radiologists are expected to be different for initial assessment as opposed to cancer outcome. In particular, variability in mammographic interpretation is well established in the literature [4–11].

As a result, for the i th mammogram, we have the data X_i^I, X_i^F, Y_i and $\mathbf{R}_i$ where $X_i^I=1$ or 0 according to whether the initial assessment is positive or negative, $X_i^F=1$ or 0 according to whether the final assessment is positive or negative, $Y_i=1$ or 0 according to whether the breast cancer is present or not and $\mathbf{R}_i$ is a vector of covariates (patient characteristics and previous mammography indicator), facility dummy variables, and radiologist dummy variables. We model the joint distribution of X_i^I, X_i^F and Y_i given $\mathbf{R}_i$. Accuracy or performance measures take the form of rates, i.e. the probability of an event defined through specification of levels of X_i^I, X_i^F, Y_i and $\mathbf{R}_i$. From this joint distribution, by appropriate conditioning, any rate of interest can be studied. In fact, with a parametric model for the joint distribution of (X_i^I, X_i^F, Y_i) , any rate becomes a parametric function.

We model the joint distribution through the event order, i.e.

$$f(X_i^I, X_i^F, Y_i | \mathbf{R}_i) = f(X_i^I | \mathbf{R}_i) \cdot f(X_i^F | X_i^I, \mathbf{R}_i) \cdot f(Y_i | X_i^I, X_i^F, \mathbf{R}_i) \quad (1)$$

Each density on the right-hand side of (1) is that of binary variable which will be modelled using logistic regression. In the sequel we adopt a Bayesian framework for inference with regard to such models. The Bayesian framework is attractive on at least two accounts. First, the rates of interest can have complicated forms. Inference about such rates using likelihood approaches requires virtually intractable delta method approximation to assess variability. Also, for many of the models we explore, we introduce random effects associated with individual radiologists. Inference for the resulting ensemble of mixed effects models within a Bayesian setting avoids possibly inappropriate asymptotics and yields straightforward interpretation. Also, marginalization over random effects is straightforward, enabling inference averaged over facilities and/or radiologists with suitable adjustment of variability.

[‡]Determination in this fashion is a consequence of definitions by the Breast Cancer Surveillance Consortium. A portion of their data is used in our analysis. Also, note that, strictly speaking, the chronology may be such that cancer is present before the initial assessment. However, our event order assumes that outcome associated with a particular mammogram is only determined *after* the mammogram.

The motivating data set is extracted from the records of the Group Health Cooperative in Seattle, Washington. Full details of the data set including definitions of covariates as well as assessments are presented in the next section. The data set is large enough (roughly 78 000 mammograms) to provide a sufficient number of positive mammograms (roughly 11 per cent) and cancers (roughly 0.7 per cent) to enable fitting of models as described above. In fact, (1) suggests that we can fit the modelling stages separately using, at each stage, specifications of increasing complexity. This exposes the need for model selection which can be implemented stagewise and which we explore below.

Thus the format of the paper is as follows. In Section 2, we describe in detail the database. In Section 3 we present the collection of models for each stage. Section 4 provides a discussion of model comparison and the selection of one model as our focus. The analysis of the data under this model is presented in Section 5. In Section 6 we discuss and illustrate computation of rates. We conclude with a summary of our findings which motivates future data collection and analysis to better explain this multistage process.

2. DESCRIPTION OF DATA SET

This analysis uses Group Health Cooperative (GHC) breast cancer surveillance data from 1 January 1996 to 31 December 1998. GHC is a non-profit consumer-governed health maintenance organization in the Pacific Northwest [12]. The data were collected primarily through the formal breast cancer screening program that began in 1986, but also includes radiology done outside the formal program [12, 13]. The program automatically sends invitations to women aged 50 and above to initiate screening. Additionally, women aged 40–49 are invited to participate if they are considered to be at high risk for breast cancer. Baseline demographic data, health history, screening history, and risk factor information are ascertained by a self-administered questionnaire. The visit includes a clinical breast examination, instruction in breast self-examination, and a bilateral screening mammogram obtained from one of six facilities. The interval between screening mammograms depends on age and other risk factors. Reminder letters are sent to the screening program participants when a mammogram is due.

This study focuses only on screening mammography. A screening mammogram was defined to be a mammogram designated as ‘screening’ by the radiologist. We exclude exams that are indicated as being unilateral; this often indicates that the woman may have had a previous unilateral mastectomy or that the exam was done for diagnostic purposes. Exams from women with a previous history of breast cancer or reporting breast implants at the time of the exam were excluded. In addition, we excluded exams preceded by any breast imaging within the prior nine months since this may indicate that such an exam was really not a screening mammogram. We note that many women receive clinical breast exams in conjunction with screening mammography. However, such physical exams are not consistently recorded and may have followed a positive mammogram, so we cannot infer about their performance. We also note that, across patients, outcomes for first screening mammogram will tend to show more positives than those for subsequent screening exams. Within our data set, we can only work at the mammogram level, introducing ‘presence of previous mammogram’ as an indicator variable.

Breast cancer cases were identified from the cancer registry or pathology data. Invasive breast cancer and ductal carcinoma *in situ* (DCIS) cancer cases were included and lobular carcinoma *in situ* (LCIS) was excluded. § The follow-up period to assess cancer status was the minimum of one year or until the next screening exam, if the screening exam was more than 9 months after the initial mammogram. All the cancer cases with the follow-up period

are included whether screen detected or not. Diagnostic exams or screens less than 270 days do not terminate the follow-up period.

To determine the initial assessment of the screening mammogram we used the BI-RADS™ assessment assigned by the radiologist [14]. An assessment of 1 or 2 indicates a normal or benign finding. An assessment of 3 or 0 signifies a finding which is likely benign or the need for additional imaging, respectively. An assessment of 4 or 5 indicates an abnormality or malignancy. Standard definitions developed by the Breast Cancer Surveillance Consortium using BI-RADS™ categories [15] were used [11]. A screening mammogram was considered positive if interpreted as BI-RADS™ category 0, 4, 5, or 3 with a recommendation for immediate work-up (denoted as 3⁺). A mammogram was considered negative if interpreted as BI-RADS™ category 1, 2 or a 3 without any recommendations or with only a recommendation for normal or short-interval follow-up (i.e. 6 months) (3⁻). If additional imaging was done on the same day as the screening mammogram, the original BI-RADS™ was recorded to 0 (needs additional assessment). If a woman had different assessments for each breast, the higher assessment level was used according to the following hierarchy (1 < 2 < 3 < 0 < 4 < 5). We recognize that a binary classification which declares positive only for BI-RADS™ category 4 or 5 is simpler. However, it is the case that most cancers come from 3⁺ and 0. An analysis using all the BI-RADS™ categories in an ordinal model would, perhaps, be most satisfying but is beyond the scope of this paper.

Additional variables related to breast cancer risk and mammogram accuracy examined in this analysis were age, breast symptoms at the time of the exam, family history of breast cancer, history of breast biopsy or surgery, menopausal status, hormone replacement therapy (HRT), breast density, and presence of a previous mammogram.

Age, breast symptoms, family history of breast cancer and history of breast biopsy or surgery were computed by compiling self-reported information from the woman. Menopausal status was a function of age and self-report. If a woman was aged 55 or older she was assumed to be post-menopausal. Post-menopausal women were further divided into those who used HRT and those who did not. Previous mammography history was computed using both self-report and prior radiology data. Breast density was determined by the radiologist using a four-point scale.

Many of the variables were naturally dichotomous. Age was categorized into 5-year groups and breast density had four categories based on the BI-RADS™ coding system: (1) almost entirely fatty, (2) scattered fibroglandular densities, (3) heterogeneously dense and (4) extremely dense. Menopausal/HRT status was created using the following categories: (1) pre-menopausal, (2) post-menopausal without HRT, (3) post-menopausal with HRT and (4) post-menopausal with use of HRT unknown. Summaries of these variables can be found in Table I.

Data is from six mammography facilities. An individual radiologist may read mammogram in more than one facility. However, a radiologist must read at least 480 mammograms per year to be considered eligible to interpret mammograms at an accredited mammography facility in U.S. [16]. As a result, in our analysis we included radiologists that have read 480

§DCIS is classified as a breast carcinoma and is grouped with invasive breast cancer as an outcome. About 20 per cent of breast carcinoma is DCIS [11]. DCIS is reported to SEER cancer registries along with invasive disease and is commonly included in analyses of breast cancer in the U.S. In many ways treatment for DCIS is the same as invasive breast cancer, i.e. mastectomy in many cases. It is true that it often does not progress to invasive disease—in part, because of treatment that may prevent that progression or because it never would have progressed. Unfortunately, since it is not possible to tell which DCIS will progress it is treated in an aggressive manner with respect to treatment. Because DCIS is reportable to cancer registries and because treatment for DCIS is similar to invasive cancer we include as an outcome.

or more mammograms resulting in a total of 77 912 mammograms which are used to fit all of the models in Section 4.

Initial assessment is based solely on first reading of the mammogram. BI-RADS™ classification 1 is by far the most frequent accounting for 76 per cent of these assessments. Classifications 0 and 2 add 11 and 10 per cent, respectively. Using the above definitions, roughly 11.4 per cent of initial assessments are positive. Table II provides a three way table summarizing joint counts for initial assessment, final assessment¶ and cancer outcome. Marginalizing over outcome we find that initial assessment and final assessment differ for $1884 + 19 = 1903$ mammograms (<3 per cent). In other words ignoring covariates, $1903/8900 = 0.214$ is the chance that an initial positive will be revised on final assessment. Twenty-eight radiologists were included in our study. They are listed in decreasing order of number of mammograms interpreted in, e.g. Figures 1(a) and 3(a). Radiologist 1 read 6658 mammogram, radiologist 28 read 649 mammograms.¶ The proportion of mammograms taken at each facility range from roughly 30 per cent at facility 1 to 7 per cent at facility 6. The actual counts are attached to Figure 1(b). Note that a radiologist may read films at more than one facility so radiologists are not nested within facility.

3. MODELLING APPROACH AND MODEL SPECIFICATION

We focus the model development here on the two event setting ignoring the X_i^F . The approach for the three event setting is similar. However, since X_i^F differs from X_i^I for only 1903 of 77 912 mammograms (2.4 per cent), the latter model proves challenging to fit for this data set. Suppressing the superscript I , we model $P(X_i = x, Y_i = y | \mathbf{R}_i)$ as

$$P(X_i = x | \mathbf{R}_i).P(Y_i = y | X_i = x, \mathbf{R}_i) \quad (2)$$

We offer five first event logistic regression models letting $\mathbf{Z}_i$ denote the risk factors from $\mathbf{R}_i$ used in explaining X_i and eight second event logistic regression models letting $\mathbf{W}_i$ denote the risk factors from $\mathbf{R}_i$ used in explaining $Y_i | X_i$. In the model fitting we considered all the main effects for the variables in Table I. We also examined several covariate interaction effects finding no additional significances.

The models are formulated as follows. At the first stage, model 1a is a fixed effects specification setting $\text{logit}(\text{Pr}(X_i = 1)) = \mathbf{Z}_i^T \alpha$. Model 1b introduces an effect ψ_k for each of the

K facilities. It sets $\text{logit}(\text{Pr}(X_i = 1)) = \mathbf{Z}_i^T \alpha + \sum_{k=1}^K v_{ik} \psi_k$. The v_{ik} are dummy variables, i.e. $v_{ik} = 1$ or 0 according to whether or not the i th mammogram is read in k th facility. Model 1c introduces a set of random effects δ_l , one for each of the L radiologists. It sets

$\text{logit}(\text{Pr}(X_i = 1)) = \mathbf{Z}_i^T \alpha + \sum_{l=1}^L u_{il} \delta_l$. The u_{il} are dummy variables, i.e. $u_{il} = 1$ or 0 according to whether or not i th radiologist read the i th mammogram. Model 1d is a combination of model 1b and 1c. It has effects for facilities as well as radiologists. It sets

$\text{logit}(\text{Pr}(X_i = 1)) = \mathbf{Z}_i^T \alpha + \sum_{l=1}^L u_{il} \delta_l + \sum_{k=1}^K v_{ik} \psi_k$. Model 1e captures the idea that there is anticipated to be some interaction between breast density classification and the radiologist assessment. From Table I, taking breast density classification 2 (scattered fibroglandular tissue) which is the most common, as the reference group, it introduces radiologist random

¶Final assessment is based upon additional imaging information (not biopsy) that could include more mammography, ultrasound and other non-invasive techniques.

¶Note that since a radiologist may read mammograms outside of GHC, our counts do not provide a complete marker of the clinical volume for each radiologist during the study period.

effects for each of the other three classifications. Whereas model 1c permits only radiologist level adjustments, model 1e enables adjustment reflecting both radiologist and breast density classification. In particular, it sets

$\text{logit}(\Pr(X_i=1)) = \mathbf{Z}_i^T \alpha + \sum_{l=1}^L u_{il} \delta_{2l} + \sum_{l=1}^L u_{il} a_{i1} \delta_{1l} + \sum_{l=1}^L u_{il} a_{i3} \delta_{3l} + \sum_{l=1}^L u_{il} a_{i4} \delta_{4l}$, where $a_{i1} = 1$ or 0 according to whether density for i th mammogram is 1 or not, $a_{i3} = 1$ or 0 according to whether density for i th mammogram is 3 or not, $a_{i4} = 1$ or 0 according to whether density for i th mammogram is 4 or not and u_{il} as in model 1b. We could extend model 1e to include facility effects but since model 1e does not improve upon model 1c there seems to be no reason to look further for this stage.

At the second stage, model 2a is a fixed effects specification adjusting the logistic form additively in the level of X_i . That is, $\text{logit}(\Pr(Y_i=1 | X_i=x_i)) = \mathbf{W}_i^T \beta + \gamma x_i$. Model 2b is again a fixed effects model which assigns a different coefficient vector according to whether $X_i = 0$ or 1, setting $\text{logit}(\Pr(Y_i=1 | X_i=0)) = \mathbf{W}_i^T \beta_1$ and $\text{logit}(\Pr(Y_i=1 | X_i=1)) = \mathbf{W}_i^T \beta_2$. Model 2c extends model 2a in the same way as the model 1b extends model 1a using facility effects. These will of course be different from those in the first stage. The form of model 2c is

$\text{logit}(\Pr(Y_i=1 | X_i=x_i)) = \mathbf{W}_i^T \beta + \gamma x_i + \sum_{k=1}^K v_{ik} \eta_k$. Finally model 2d extends 2b in the same fashion as model 2c extends model 2a. It sets $\text{logit}(\Pr(Y_i=1 | X_i=0)) = \mathbf{W}_i^T \beta_1 + \sum_{k=1}^K v_{ik} \eta_{1k}$ and sets $\text{logit}(\Pr(Y_i=1 | X_i=1)) = \mathbf{W}_i^T \beta_2 + \sum_{k=1}^K v_{ik} \eta_{2k}$.

Model 2e extends model 2a in the same way as the model 1c extends model 1a using radiologists random effects, again different from those in the first stage. The form of model

2e is $\text{logit}(\Pr(Y_i=1 | X_i=x_i)) = \mathbf{W}_i^T \beta + \gamma x_i + \sum_{l=1}^L u_{il} \phi_l$. Next, model 2f extends 2b in the same fashion as model 2e extends model 2a. Now we have a set of radiologist random effects given $X_i = 0$ and a different set given $X_i = 1$. Specifically, model 2f sets

$\text{logit}(\Pr(Y_i=1 | X_i=0)) = \mathbf{W}_i^T \beta_1 + \sum_{l=1}^L u_{il} \phi_{1l}$ and sets $\text{logit}(\Pr(Y_i=1 | X_i=1)) = \mathbf{W}_i^T \beta_2 + \sum_{l=1}^L u_{il} \phi_{2l}$. Model 2g includes facility effects as well as radiologist random effects. It is

$\text{logit}(\Pr(Y_i=1 | X_i=x_i)) = \mathbf{W}_i^T \beta + \gamma x_i + \sum_{l=1}^L u_{il} \phi_l + \sum_{k=1}^K v_{ik} \eta_k$. Similarly, model 2h sets

$\text{logit}(\Pr(Y_i=1 | X_i=0)) = \mathbf{W}_i^T \beta_1 + \sum_{l=1}^L u_{il} \phi_{1l} + \sum_{k=1}^K v_{ik} \eta_{1k}$ and sets

$\text{logit}(\Pr(Y_i=1 | X_i=1)) = \mathbf{W}_i^T \beta_2 + \sum_{l=1}^L u_{il} \phi_{2l} + \sum_{k=1}^K v_{ik} \eta_{2k}$.

In all these models the prior for α 's, β 's, ψ 's and η 's is taken to be normal with mean zero and variance 100, which is essentially non-informative. Note that, while prior studies have established association between recall or cancer incidence and one or more of our explanatory variables, it is difficult to convert this information to prior specification for coefficients in multistage models. We prefer to use vague specifications for these coefficients and let the data drive the inference. The various radiologist random effects, ψ 's and δ 's are all assumed *a priori* to be normally distributed with mean zero and corresponding variance component. For all variance components we assign proper but somewhat vague inverse gamma priors. That is, we assume that each prior has shape parameter equal to 2, yielding an infinite variance but finite mean. For the means of these priors we use crude data-based estimates of these variance components. ** The models are

** With our large data set and the fact that each radiologist in the study read at least 649 mammograms, the data carry considerable information about these variance components. Thus it is not surprising that we find little sensitivity to their prior specification.

fitted using tailored Markov Chain Monte Carlo algorithms (see Reference [17]) to accommodate the large sample size and the sparseness of cancers.

4. MODEL COMPARISON

From (2) the models proposed in Section 3 result in a log likelihood which is additive across the first and second stages. Hence any likelihood based model selection criterion can be used to provide a first stage selection, a second stage selection and, hence, an overall selection. In fact, all such criteria emerge as penalized likelihood forms. We compute AIC_1 and DIC_1 along with AIC_2 , and DIC_2 and can obtain for the full models, $AIC = AIC_1 + AIC_2$ and $DIC = DIC_1 + DIC_2$. In this notation AIC denotes the Akaike Information Criterion [18] and DIC, the recently introduced Deviance Information Criterion [19]. We do not take a position as to which criterion is preferable. We do note that, in the case of fixed effects, Spiegelhalter *et al.* show that DIC is approximately AIC. Moreover, in the case of mixed models, they argue that p_{D1} and p_{D2} measure the ‘effective number of parameters’ in the respective modelling stage. So perhaps DIC is most suitable for our setting. We did not employ BIC, the Bayesian Information Criterion [20] since, in theory, it is intended for fixed effects models and, in practice, it is known to provide too much support for simpler models.

The model comparison for the models presented in Section 3 is presented in Table III with Table IIIa giving the first stage results and Table IIIb the second stage results. Table IIIa shows that model 1d is preferred providing the smallest AIC_1 and DIC_1 . Table IIIb presents a less clear choice using AIC_2 and DIC_2 . That is, within the accuracy of the simulation-based fitting, the values are not significantly different. We select model 2d since it is nearly best under both criteria. So, in Section 5 we use the model 1d, 2d to carry out the data analysis.

5. ANALYSIS OF THE SELECTED MODEL

Following Section 4 we use a first stage model with patient characteristic effects, fixed facility effects and facility and radiologist random effects along with a second stage model which allows a different coefficient vector and different facility effects according to whether $X_j = 1$ or 0 but no radiologist effects. It is perhaps not surprising that the choice of radiologist significantly affects the chance of a positive assessment but does not significantly affect the cancer outcome given initial assessment.

Table IV summarizes the posterior distribution of the fixed effects in model 1d. We see that the presence of symptoms, history of breast biopsy and no previous mammogram for comparison all significantly increase the chance of positive assessment. Family history of breast cancer does not significantly affect the chance of positive assessment. Turning to breast density classification, relative to the reference classification 2, classification 3 provides the most increased risk of positive assessment. Classification 4 provides a smaller but still significant increase in risk. Classification 1 decreases the risk. With regard to menopausal status, relative to the ‘post-menopausal, no HRT’ classification, all other classifications increase the risk of positive assessment. Finally, for age, the youngest females in the study and the oldest females in the study have significantly reduced risk of positive assessment. Most of these results are consistent with findings of previous studies which investigated individual risk factors for positive assessment after a screening mammogram (see References [10, 21] and further references therein).

Table V summarizes the posterior distribution of the fixed effects in model 2d. The first three columns show the posterior summaries for the model coefficients given $X_j = 0$ (negative initial assessment), the next three columns show these posterior summaries given $X_j = 1$ (positive initial assessment). Given a negative initial assessment, only breast density

classification is significant. However, given a positive initial assessment, presence of symptoms increases chance of presence of cancer. Now, breast density classification becomes an insignificant risk factor. Now, the 'menopausal status/HRT' classification decrease the chance of presence of cancer. As expected younger women have a decreased chance of breast cancer, older women have an increased chance (See Reference [22] for comparable findings).

Figure 1 provides side by side box plots of the random radiologist effects in model 1d. The figure provides a clear indication of the magnitude of variability in radiologists. Several box plots lie entirely above zero. These radiologists introduce a significant positive random effect hence elevating the chance of positive initial assessment for the mammograms they interpret. Similarly, box plots falling below zero are associated with radiologists having a diminished chance of positive initial assessment. Note also that the box plots tend to grow wider as we move from left to right. This reflects the fact that, radiologists interpret fewer mammograms as we move from left to right. The facility effects (Table IV) are less dramatically different than those for the radiologists though the effects for facilities 1 and 5 are suggestively larger than those for facilities 3, 4 and 6 again elevating the chance of a positive initial assessment. Table V provides facility effects for model 2d. Here we see that, none of the facility effects differ significantly from each other. This is in accord with the fact that models 2b and 2d were not too different in the model comparison. It also agrees with our intuition. Variability in facilities would not be expected to strongly affect the incidence of cancer given initial assessment.

6. DEVELOPMENT OF RATES

As noted above, each of the rates of interest is a parametric function, hence within the Bayesian framework, has a posterior distribution. Due to our model formulation, prospective rates, i.e. those involving the probability of Y given X are immediately available from the second stage model. In particular, working with the model 2d, suppose a mammogram presents risk factors $\mathbf{R}_0$ and was interpreted in the k th facility. The predictive value of a positive test $P(Y=1 | X=1) \equiv \text{PPV}(\mathbf{W}_0, \boldsymbol{\beta}_2, \eta_{2k})$. The predictive value of a negative test, $P(Y=0 | X=0) \equiv \text{NPV}(\mathbf{W}_0, \boldsymbol{\beta}_1, \eta_{1k})$. Retrospective rates, e.g. sensitivity and specificity require both modelling stages and are computed under the joint model 1d/2d using Bayes' theorem. In particular, sensitivity

$$P(X=1 | Y=1) = \frac{P(Y=1 | X=1).P(X=1)}{(P(Y=1 | X=1).P(X=1) + P(Y=1 | X=0).P(X=0))} \equiv \text{se}(\mathbf{Z}_0, \mathbf{W}_0, \alpha, \delta_l, \psi_k, \beta_1, \beta_2, \eta_{1k}, \eta_{2k})$$

Similarly, specificity, $P(X=0 | Y=0)$, will be a function of all the model parameters.

The simulation based model fitting produces posterior samples for all the model parameters. Given any rate of interest each sampled parameter vector yields a posterior realization of the rate by inserting the sampled vector into the parametric function. Hence, we obtain a posterior sample for the rate to allow arbitrary inference about the rate.

In practice it would be of more interest to investigate a rate given $\mathbf{R}_0$ but not specifying a particular radiologist or facility. One possibility is to obtain the posterior associated with the radiologist/facility pair presenting the largest effects, another using the pair with the smallest effects. Hence, the extremes of the rates across radiologists/facility could be examined. Another option is to look at a rate 'averaged over radiologists and facility'. Within the Bayesian framework this requires integrating the rate with respect to the distribution of any effects involved in the rate. For example, to evaluate sensitivity as a function of risk factors and marginalizing over radiologist and facility effects, we seek

$$\tilde{se}(\mathbf{Z}_0, \mathbf{W}_0, \alpha, \beta_1, \beta_2) = \int se(\mathbf{Z}_0, \mathbf{W}_0, \alpha, \delta, \psi, \beta_1, \beta_2, \eta_1, \eta_2) f(\psi) f(\delta | \mu_\delta, \sigma_\delta^2) \times f(\eta_1) f(\eta_2) d\psi d\delta d\eta_1 d\eta_2$$

One can, as well envision marginalization only over radiologists but specifying facility, say k . Then the integration in (4) is only over δ with ψ, η_1, η_2 now subscripted by k .

The marginalization in (4) cannot be carried out explicitly. So we propose Monte Carlo integration, i.e. given a posterior draw $(\alpha^*, \beta_1^*, \mu_\delta^*, \sigma_\delta^{2*})$ we draw $\psi_b^*, \delta_b^*, \eta_{1,b}^*, \eta_{2,b}^*, b=1, \dots, B$, from the normal distributions $f(\psi)$, $f(\delta | \mu_\delta^*, \sigma_\delta^{2*})$, $f(\eta_1)$ and $f(\eta_2)$, respectively. Then, we take as a posterior realization of the sensitivity, $B^{-1} \sum_{b=1}^B se(\mathbf{Z}_0, \mathbf{W}_0, \alpha^*, \beta_1^*, \beta_2^*, \delta_b^*, \psi_b^*, \eta_{1,b}^*, \eta_{2,b}^*)$. B can be made arbitrarily large so this Monte Carlo integration can be implemented to arbitrary accuracy. Hence, each posterior sample of the parameters enables a Monte Carlo integration and gives a posterior realization of the sensitivity. All of these integrations are done after the MCMC model fitting using only the retained posterior samples.

In Figure 2(a) we present posterior point and interval estimates for the probability of positive initial assessment across radiologists when patient characteristics are set at highest risk levels under model 1d (from Table IV) and the highest risk facility (from Table IV) is selected. For comparison, the figure also provides these estimates at the lowest risk levels. Figure 2(b) provides the same comparison but now across facilities, choosing the highest and lowest risk radiologist (from Figure 1). Not surprisingly, the relative risks for ‘high’ vs ‘low’ are very large.

Figure 3 considers false negative rate (1—sensitivity) but with a different intention. First, as (3) reveals, required computation employs both models 1d and 2d. This makes ‘high’ and ‘low’ risk patient characteristics ill-defined. So, instead we choose ‘typical’ characteristics. Second, we look at these rates across age groups rather than fixing any particular group. Third, we introduce an ‘average’ facility (effect) and an ‘average’ radiologist (effect) and compare the resulting rates across age with those from marginalization over facility and radiologist following (4). As would be expected, point estimates are similar but variability is increased by marginalization. We would claim that this increased variability better reflects the uncertainty associated with these rates. Note that, from Table II, we can calculate an empirical overall false negative rate to be $79/(476 + 78) = 0.142$, in accord with the various rates in Figure 3.

7. CONCLUSION

In conclusion, the interpretation of mammography is an art and it is well-established that there is considerable variability among radiologists in interpreting mammograms (again, see Reference [10] and references therein.). We have presented here a flexible multistage hierarchical modelling approach which reflects the event order associated with screening mammography. Adjusting for patient characteristics, through the introduction of random effects we discover significant differences in radiologists with regard to initial assessment of the mammogram. We are able to use this hierarchical modelling approach to estimate any rates of interest, including, through suitable marginalization, rates without specifying radiologist or facility. The MCMC model fitting enables us to trivially attach uncertainty to these estimates.

Work in progress has focused on development of a survey of radiologists and a survey of facilities. Through the results of these surveys we shall endeavour to explain the differences in radiologists and in facilities. We will introduce additional levels in our modelling to

explain radiologist and facility effects with appropriate explanatory variables. We will report our findings in a subsequent manuscript.

Acknowledgments

Contract/grant sponsor: Agency for Health Research and Quality (AHRQ) and the National Cancer Institute (NCI); contract/grant number: HS10591 (J Elmore)

Contract/grant sponsor: National Cancer Institute; contract/grant number: KO5-CA104699

REFERENCES

1. NHANES. 2002. <http://www.cdc.gov/nchs/fastats/mammogram.htm>
2. U.S. Preventive Services Task Force. Screening for breast cancer: recommendations and rationale. *Annals of Internal Medicine*. 2002; 137:344–346. [PubMed: 12204019]
3. Gelfand AE, Ecker MD, Christiansen C, McLaughlin TJ, Soumerai SB. Conditional categorical response models with application to treatment of acute myocardial infarction. *Journal of the Royal Statistical Society, Series C*. 2000; 49:171–186.
4. Vineis P, Sinistrero G, Temporelli A, Azzoni L, Bigo A, Burke P, et al. Inter-observer variability in the interpretation of mammograms. *Tumori*. 1988; 74:275–279. [PubMed: 3400118]
5. Baines C, McFarlane D, Miller A. The role of the reference radiologist: estimates of interobserver agreement and potential delay in cancer detection in the national breast screening study. *Investigative Radiology*. 1990; 25:972–976.
6. Ciccone G, Vineis P, Frigerio A, Segnan N. Inter-observer and intra-observer variability of mammographic examination interpretation. *European Journal of Cancer*. 1992; 28A:1054–1058. [PubMed: 1627374]
7. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretation of mammograms. *New England Journal of Medicine*. 1994; 331:1493–1499. [PubMed: 7969300]
8. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by U.S. radiologists. *Archives of Internal Medicine*. 1996; 156:209–213. [PubMed: 8546556]
9. Kerlikowske K, Grady D, Barclay J, Frankel SD, Ominsky SH, Sickels EA, et al. Variability and accuracy in mammographic interpretation using the American college of radiology breast imaging reporting and data system. *Journal of the National Cancer Institute*. 1998; 90:1801–1809. [PubMed: 9839520]
10. Elmore JG, Miglioretti DL, Reish 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]
11. Ernster VL, Ballard-Barbash R, Barlow WE, et al. Detection of ductal carcinoma *in situ* in women undergoing screening mammography. *Journal of the National Cancer Institute*. 2002; 94:1546–1554. [PubMed: 12381707]
12. Thompson RS, Taplin SH, Carter AP, et al. Risk-based breast cancer screening program. *HMO Practice*. 1982; 2:177–191.
13. Taplin SH, Anderman CA. Risk-based breast cancer screening in an HMO: the first year's experience. *Group Health Institute Proceedings*. 1987:381–384.
14. American College of Radiology. Reston (VA): American College of Radiology; 1998. Breast Imaging Reporting and Data System (Bi-RADS™).
15. Kopans, DB.; D'Orsi, CJ.; Adler, DD., et al. Breast Imaging Reporting and Data System. 2nd edn. Reston VA: American College of Radiology; 1995.
16. Houn F, Elliott ML, McCrohan JL. The Mammography Quality Standards Act of 1992. *Radiology Clinical North America*. 1995; 33:1059–1065.
17. Gilks, WR.; Richardson, S.; Spiegelhalter, DJ., editors. Markov Chain Monte Carlo in Practice. New York: Chapman & Hall; 1995.

18. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1973; 19:716–723.
19. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*. 2002; 64:1–34.
20. Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978; 6:461–464.
21. 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; 92:1657–1666. [PubMed: 11036111]
22. Clemons M, Goss P. Estrogen and the risk of breast cancer. *New England Journal of Medicine*. 2001; 344(4):276–285. [PubMed: 11172156]
23. Taplin SH, Mandelson MT, Anderman CA, White E, Thompson RS, et al. Mammography diffusion and trends in late stage breast cancer: evaluation outcomes in a population. *Cancer Epidemiology, Biomarkers and Prevention*. 1997; 6:625–631.

Figure 1.
Random effects across the 28 radiologists under model 1d.

Figure 2.

(a) Estimated probability of positive initial assessment across radiologists for high risk (upper) and low risk (lower) women. See text for details; and (b) probability of positive initial assessment across facilities for high risk (upper) and low risk (lower) women. See text for details.

Figure 3. (1—sensitivity) (point and interval estimates) across age groups for a mammogram associated with ‘typical’ characteristics (See text for details). ■ is for an average radiologist and facility; • is for marginalization over radiologist and facility.

Table I

Summary of screening mammogram data.

Description of variable	Total	Per cent	Cancer	Positive
Total	77 912		555	8900
Age				
40–44	10 544	13.53	31	1355
45–49	11 965	15.36	53	1547
50–54	14 381	18.46	80	1752
55–59	9378	12.04	73	1086
60–64	7637	9.8	69	810
65–69	7244	9.3	70	774
70–74	7132	9.15	75	671
75–79	5484	7.04	56	517
80–84	2973	3.82	36	293
85–89	1174	1.51	12	95
Racial=ethnic group				
White	68 786	88.29	498	7832
Asian	4215	5.41	21	426
Black	2304	2.96	16	285
Native American=Alaskan native	374	0.48	5	56
Other (includesMixed)	1243	1.6	9	158
Unknown	990	1.27	6	143
Breast symptoms				
No symptom	73 770	94.68	495	8216
Lump=nipple discharge=other	4142	5.32	60	684
Family history of breast cancer				
No	64 449	82.72	434	7362
Yes	13 463	17.28	121	1538
History of breast biopsy=surgery				
No	62 862	80.68	407	6895
Yes	15 050	19.32	148	2005
Menopause=HRT * status				
1. Pre-menopausal	19 992	25.66	83	2702
2. Post-menopausal, no HRT	26 860	34.47	216	2528
3. Post-menopausal, HRT	25 901	33.24	204	3092
4. Post-menopausal, unknown HRT	5159	6.62	52	578
Breast parenchymal density				
1. Almost entirely fatty	7131	9.15	28	416
2. Scattered fibroglandular tissue	32 131	41.24	220	3291
3. Heterogenously dense	29 735	38.16	236	3956
4. Extremely dense	8915	11.44	71	1237
Previous mammogram				

Description of variable	Total	Per cent	Cancer	Positive
No	5469	7.02	36	914
Yes	72 443	92.98	519	7986

*
HRT = Hormone Replacement Therapy.

Table II

Three way table summary of initial assessment, final assessment and cancer outcome.

		Cancer Outcome		Total
		Positive	Negative	
Final assessment				
Initial assessment	Positive	467	6530	6997
	Negative	8	1876	1884
	Unknown	1	18	19
	Negative	79	68 933	69 012

Table III

Model selection.

(a) First stage model selection						
Model	$-2E \log(l(\theta_1; X) \text{Data})$	# of parameters	$-2 \log(l(E(\theta_1) \text{Data}))$	P_{D1}	AIC ₁	DIC ₁
	(i)	(ii)	(iii)	(i) – (iii)	(i) + 2(ii)	2(i) – (iii)
1a	54531	18	54513	18	54567	54549
1b	54105	23	54082	23	54151	54128
1c	53712	45	53668	44	53802	53756
1d	53643	50	53596	47	53743	53690
1e	53684	129	53594	90	53942	53774
(b) Second stage model selection						
Model	$-2E \log(l(\theta_2; X) \text{Data})$	# of parameters	$-2 \log(l(E(\theta_2) \text{Data}))$	P_{D2}	AIC ₂	DIC ₂
	(i)	(ii)	(iii)	(i) – (iii)	(i) + 2(ii)	2(i) – (iii)
2a	4848	19	4830	18	4886	4866
2b	4780	36	4754	26	4852	4806
2c	4789	24	4770	19	4837	4808
2d	4750	46	4724	26	4842	4776
2e	4780	46	4754	26	4872	4806
2f	4766	90	4730	36	4946	4802
2g	4772	51	4740	32	4874	4804
2h	4724	100	4680	44	4924	4768

Table IV

Summary of the posterior distribution for parameters in model 1d.

Description of variable	Median	(95 per cent L.C.L, 95 per cent U.C.L.)	
Breast symptoms			
No symptom	Referent		
Lump=nipple discharge=other	0.3970	(0.3090, 0.4840)	+
History of breast biopsy=surgery			
No	Referent		
Yes	0.2340	(0.1800, 0.2900)	+
Family history of BC			
No	Referent		
Yes	0.0290	(−0.0320, 0.0880)	0
Menopause=HRT status			
Pre-menopausal	0.2520	(0.1770, 0.3310)	+
Post-menopausal, no HRT	Referent		
Post-menopausal, HRT	0.2360	(0.1800, 0.2920)	+
Post-menopausal, unknown HRT	0.1370	(0.0420, 0.2360)	+
Parenchymal density			
Almost entirely fatty	−0.6140	(−0.7170, −0.5130)	−
Scattered Fibroglandular tissue	Referent		
Heterogenously dense	0.2060	(0.1550, 0.2550)	+
Extremely dense	0.0850	(0.0100, 0.1590)	+
Previous mammogram			
No	0.5380	(0.4560, 0.6190)	+
Yes	Referent		
Age			
40–44	−0.1670	(−0.2570, −0.0770)	−
45–49	−0.0170	(−0.0940, 0.0590)	0
50–54	Referent		
55–59	0.0150	(−0.0700, 0.0960)	0
60–64	−0.0460	(−0.1350, 0.0460)	0
65–69	−0.0210	(−0.1130, 0.0690)	0
70–74	−0.1500	(−0.2440, −0.0560)	−
75	−0.1890	(−0.2740, −0.1000)	−
Facility effect			
Facility 1	0.2735	(−0.4150, 0.9770)	0
Facility 2	0.0990	(−0.5820, 0.8110)	0
Facility 3	−0.3300	(−1.0080, 0.3770)	0
Facility 4	−0.2940	(−0.9620, 0.4170)	0
Facility 5	0.3280	(−0.3580, 1.0420)	0
Facility 6	−0.3590	(−1.0420, 0.3560)	0
Radiologist variance component	0.061	(0.032, 0.767)	

Table V

Summary of the posterior distribution for parameters in model 2d.

Description of variable	Negative initial assessment		Positive initial assessment	
	median	(LCL, UCL)	median	(LCL, UCL)
Breast symptoms				
No symptom	Referent		Referent	
Lump=nipple discharge=other	0.494	(-0.014, 0.954)	0	0.367 (0.091, 0.643) +
History of breast biopsy=surgery				
No	Referent		Referent	
Yes	0.202	(-0.192, 0.578)	0	0.128 (-0.080, 0.327) 0
Family history of breast cancer				
No	Referent		Referent	
Yes	0.083	(-0.329, 0.476)	0	0.244 (0.025, 0.462) 0
Menopause=HRT status				
1. Pre-menopausal	-0.168	(-0.617, 0.283)	0	-0.333 (-0.638, -0.037) -
2. Post-menopausal, no HRT	Referent		Referent	
3. Post-menopausal, HRT	0.062	(-0.324, 0.435)	0	-0.235 (-0.447, -0.028) -
4. Post-menopausal, unknown HRT	0.182	(-0.317, 0.666)	0	0.105 (-0.201, 0.398) 0
Breast parenchymal density				
1. Almost entirely fatty	-0.223	(-0.748, 0.281)	0	-0.012 (-0.379, 0.323) 0
2. Scattered fibroglandular tissue	Referent		Referent	
3. Heterogeneously dense	0.525	(0.153, 0.895)	+	-0.061 (-0.251, 0.132) 0
4. Extremely dense	0.677	(0.212, 1.122)	+	0.073 (-0.225, 0.356) 0
Previous mammogram				
No	0.095	(-0.424, 0.597)	0	-0.001 (-0.331, 0.316) 0
Yes	Referent	Referent	Referent	
Age				
40-44	-0.227	(-0.729, 0.272)	0	-0.524 (-0.885, -0.179) -
45-49	0.156	(-0.296, 0.602)	0	-0.405 (-0.720, -0.111) -
50-54	Referent		Referent	
55-59	0.223	(-0.237, 0.668)	0	0.172 (-0.117, 0.465) 0
60-64	0.015	(-0.476, 0.493)	0	0.459 (0.166, 0.743) +

Description of variable	Negative initial assessment		Positive initial assessment	
	median	(LCL, UCL)	median	(LCL, UCL)
65-69	0.229	(-0.252, 0.693)	0	0.451 (0.157950, 0.738025)
70-74	0.1215	(-0.380, 0.591)	0	0.685 (0.389, 0.965)
75	0.303	(-0.164, 0.741)	0	0.705 (0.439, 0.986)
Facility effect				
Facility 1	-6.171	(-7.021, -5.322)	—	-2.318 (-2.892, -1.750)
Facility 2	-5.873	(-6.748, -5.032)	—	-1.794 (-2.379, -1.212)
Facility 3	-5.825	(-6.741, -4.959)	—	-2.141 (-2.744, -1.534)
Facility 4	-5.5455	(-6.433, -4.665)	—	-2.049 (-2.649, -1.422)
Facility 5	-6.0525	(-7.253, -5.003)	—	-2.149 (-2.779, -1.489)
Facility 6	-6.403	(-7.792, -5.214)	—	-2.126 (-2.836, -1.428)