Statistics in Medicine

Research Article

Received 26 March 2014, Accepted 10 July 2015 Published online 16 August 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6619

Inference on cancer screening exam accuracy using population-level administrative data

H. Jiang,[a,b*†] P. E. Brown[a,b,c] and S. D.Walter[b]

This paper develops a model for cancer screening and cancer incidence data, accommodating the partially unobserved disease status, clustered data structures, general covariate effects, and dependence between exams. The true unobserved cancer and detection status of screening participants are treated as latent variables, and a Markov Chain Monte Carlo algorithm is used to estimate the Bayesian posterior distributions of the diagnostic error rates and disease prevalence. We show how the Bayesian approach can be used to draw inferences about screening exam properties and disease prevalence while allowing for the possibility of conditional dependence between two exams. The techniques are applied to the estimation of the diagnostic accuracy of mammography and clinical breast examination using data from the Ontario Breast Screening Program in Canada. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Bayesian inference; test accuracy; cancer screening; clustered analysis; random effect; latent-class model

1. Introduction

Population-level administrative data from screening programs for cancers and other diseases are a rich and extensive information source for evaluating screening effectiveness. The Ontario Breast Screening Program (OBSP) in Canada maintains one such database, containing information on hundreds of thousands of cancer screens, including individual-level characteristics such as age and family history [1–3]. In comparison with data from medical records or clinical trials, however, population-level data are more challenging to analyze because of two key difficulties not faced with clinical study data. The first is the observational nature of population-level data, with the test performance affected by heterogeneities in both the characteristics of the individuals being screened [4] and in the judgements and characteristics of the medical professionals carrying out the exams [5]. The second difficulty is the lack of a 'gold standard' reference test applied to all subjects, with the result that a number of cancers might remain unobserved or missed by the examiner and the health system during the follow-up periods.

The first of the difficulties faced with population-level data, which in statistical terms is that of correlated data with varying probabilities of incidence and detection, can be accounted for through mixed-effects models [e.g., 6]. A mixed-effects model includes both fixed and random effects and allows a wide variety of correlation patterns to be modeled explicitly. Individual factors such as age and family history, examiner-level factors such as years of experience, and institution-level variables such as annual number of patients screened can be explained by their inclusion as fixed effects. Variations in detection rates and false positive rates among examiners and institutions can be captured by random effects, as in the two-level hierarchical model used by Woodard *et al.* [7].

The solution to the second difficulty of some incident cases being unobserved is less straightforward, particularly, when more than one screening test is administered and the sensitivity of this second test is

[a] *Analytics Informatics, Cancer Care Ontario, Toronto, ON, Canada*
[b] *Department of Clinical Epidemiology & Biostatistics, McMaster University, Hamilton, ON, Canada*
[c] *Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada*
**Correspondence to: H. Jiang, Analytics Informatics, Cancer Care Ontario, Toronto, ON, Canada.*
[†] *E-mail: Hedy.Jiang@cancercare.on.ca*

variable. In the OBSP data considered here, each woman is examined twice, with a radiologist performing a mammography and a nurse giving a clinical breast exam (CBE). When mammography is performed first, a cancer missed by mammography might be either: (i) detected by CBE and diagnosed; (ii) missed by CBE and later became symptomatic and diagnosed by clinical practitioners; or (iii) missed by CBE and remaining undetected. The consequence of misclassifying the cancers in category (iii) as true negatives would be slight when considering a strata of the population for whom CBE is particularly effective and the number of such cancers is small. This misclassification would be consequential when comparing mammography sensitivity across different strata, such as age groups or health facilities, where the sensitivity of CBE is known to vary. Such an analysis should ideally acknowledge and account for the CBE test having created differences in misclassification rates among strata.

The goal of this paper is to make inferences on the performance of screening tests and to quantify the effect of specified explanatory variables using population-level administrative data. The particular objective considered is accommodating the heterogeneous nature of population-level data by explicitly modeling individual-level test result probabilities and creating and integrating out a ‘nuisance variable’ for unobserved cancers. The random effects model employed in this paper accommodates multilevel data structures, general covariate effects, partially unobserved disease status, and dependence between screening tests. A Markov Chain Monte Carlo (MCMC) algorithm is described for estimating the posterior distributions of sensitivity, specificity, and prevalence when the reference standard is imperfect. The MCMC approach extends naturally to the case of unobserved cancers by including cancer status in the model as an unknown (or partially observed) latent variable. The method is illustrated through the analysis of data from the OBSP.

1.1. Data and motivating problem

The data considered here is a cohort of 234,177 women screened, between January 1, 2002, and December 31, 2003, in the 73 OBSP screening centers, which offer both CBE an d mammography.

The Ontario Cancer Registry contains cancer incidence data from a number of data sources, including medical records and death certificates, and information about all cancers clinically diagnosed among the women screened is linked to the OBSP database. Cancers listed by the Registry as being diagnosed within 12 months of an individual’s screening date in the OBSP database are assumed to have been present and detectable during the screening exam. This 12-month threshold is fairly standard practice and used, for example, by Chiarelli *et al.* [2,3]; these references can be consulted for further information on the OBSP program and data they compile.

Table I. The number of women having negative screening tests and positive screening tests, and total number of screens administered for both mammography and CBE, subdivided according to whether cancer was diagnosed within 12 months of screening.

	CBE		
	Negative	Positive	Total
No cancer diagnosed			
Mammography negative	212190	5109	217,299
Mammography positive	13768	1288	15,056
Total	225958	6397	232,355
Cancer diagnosed			
Mammography negative	261	94	355
Mammography positive	1070	397	1467
Total	1331	491	1822
Total			
Mammography negative	212451	5203	217,654
Mammography positive	14838	1685	16,523
Total	227289	6888	234,177

CBE, clinical breast exam.

Table II. Summary of individual-level variables in OBSP dataset, with total number of women screened, number with positive screening tests, and number of cancers observed.

	Total screened		Positive tests		Cancer diagnosed	
Breast cancer in family						
No	205639	(87.8%)	19008	(87.5%)	1517	(83.3%)
Yes	28538	(12.2%)	2718	(12.5%)	305	(16.7%)
Breast density						
Low, 75%	220456	(94.1%)	19966	(91.9%)	1664	(91.3%)
High, ⩾75%	13721	(5.9%)	1760	(8.1%)	158	(8.7%)
Age (years)						
50–59	132000	(56.4%)	13316	(61.3%)	917	(50.3%)
60–69	102177	(43.6%)	8410	(38.7%)	905	(49.7%)

OBSP, Ontario Breast Screening Program.

Table I shows the number of positive and negative screening tests for each of CBE and mammography, subdivided according to whether cancers were subsequently diagnosed within 12 months of screening. There were 261 cancers diagnosed for women having negative screening test results for both mammography and CBE or 14% of all cancer cases. Of the individuals having a breast cancer diagnosis, the sensitivity of mammography is much higher than CBE, being $1467/1822 = 80.5\%$ versus $491/1822 = 26.9\%$ for CBE. Of the individuals for whom no cancer was diagnosed, the specificity of mammography is $217,299/232,355 = 93.5\%$, while the specificity of CBE is $225,958/23,255 = 97.2\%$. Table II summarizes the individual-level variables in OBSP dataset, with total number of women screened, number with positive screening tests, and number of cancers observed.

1.2. Literature review

Uncertainty in true disease status is known to affect estimates of sensitivity and specificity, with misclassification of a few breast cancer cases being able to cause a dramatic change in estimates of parameters in a logistic regression model [8]. Several authors [9–12] discuss the effects of misclassification on binary responses in different settings and claim that failure to account for measurement errors in covariates or responses causes biased and inconsistent parameter estimates. Approaches proposed for correcting the estimates of regression parameters include the simulation and extrapolation method [13] and a number of Bayesian methods [e.g., 12, 14].

To deal with imperfect reference tests, several authors have developed latent-class models in evaluating accuracy of diagnostic tests [15–20]. These methods generally require more than two independent tests in order to estimate parameters of interest. When there are fewer than two different tests, some parameters need to be constrained at fixed values based on model assumptions. An alternative is to use Bayesian inference, assigning a prior distribution to reflect the uncertainty in these parameter values. Assuming conditional independence between tests, Joseph *et al.* [21] and Black and Craig [22] use Bayesian inference to estimate disease prevalence, as well as model parameters, specifying Beta distributions as the prior for the prevalence, sensitivity, and specificity and obtaining the joint posterior distribution via the Gibbs sampling. Dendukuri and Joseph [23] extended their work and estimated the disease prevalence and test accuracy while adjusting for conditional dependence between two tests.

Heterogeneity in population-level data has been addressed with random effects models in a number of studies [20, 24–26]. Puggioni *et al.* [27] introduced a joint model for the four possible cell probabilities that determined the screening test result when two tests were applied. Through the model, the stochastic dependence between the estimates of sensitivity and specificity could be examined. The methodology presented here is based on a similar random effects structure and extends it to allow for multilevel correlation structures, covariates attached to observations, as well as examiners, and partially unobserved disease status.

2. A latent-variable model for cancer screening

2.1. Model description

Cancer screening outcomes are described with a random effects model, with the observed data including test results and cancer diagnoses and the unobserved latent variables including the individual's true disease status and the true and false positive rates of the screening tests being administered. The model structure is illustrated graphically in Figure 1 and explained in detail in the following paragraphs.

The observed data used in the cancer screening model have the following elements:

- T_{ij} is the result of screening test j performed on individual $i = 1 \dots I$ with $I = 234,117$, $j = 1$ indicating mammography and $j = 2$ denoting CBE. A positive or abnormal test is coded as $T_{ij} = 1$ with $T_{ij} = 0$ otherwise.
- Y_i is the observed cancer status of individual i, with $Y_i = 1$ if cancer was clinically diagnosed within 12 months of the screening exam and $Y_i = 0$ otherwise.
- X_i is a vector of covariates (or explanatory variables) associated with individual i.
- $s = 1 \dots S$ with $S = 73$ denotes a health facility (or screening site) with Q_s being a vector of covariates for facility s.
- Examiners are nested within health facilities, with E_{sj} different examiners at site s administering exams of type j, and R_{sje} (for $e = 1 \dots E_{sj}$) contains the covariates associated with the eth examiner of type j at site s.
- The triplet (s_i, j, e_{ij}) specifies that individual i was tested at screening site s_i and that test j was administered by the e_{ij}th examiner at site s_i.

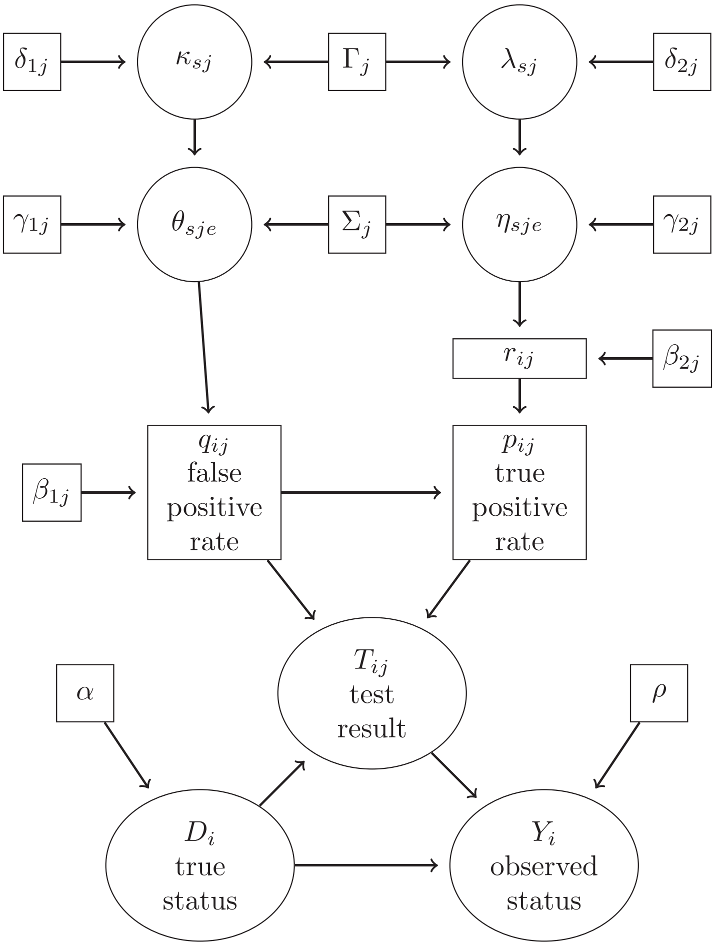


Figure 1. Graphical representation of the latent-variable screening model.

There are no examiner-level variables R_{sje} in the application of the model presented here, although this variable is retained in the model description for completeness.

The probability of observing any particular combination of test results T_{ij} and cancer incidence Y_i depends on the following, generally unobserved, quantities:

- D_i is the true disease status of subject i at the time of the screening tests. An observed cancer incidence with $Y_i = 1$ implies $D_i = 1$, although the converse is not always true, with unobserved cancers having $D_i = 1$ and $Y_i = 0$ being possible.
- p_{ij} and q_{ij} are the true positive and false positive rates for exam j as administered to individual i.
- ρ is the probability that a cancer that was not detected during screening exams will be subsequently diagnosed and recorded within 12 months.

The screening model is a four-level hierarchical model, with the first level specifying a joint distribution for T_{ij} and Y_i conditional on D_i:

$$\begin{aligned}
&\text{Test results:} && T_{ij}|D_i \begin{cases} \sim \text{Bernoulli}(q_{ij}) & \text{if } D_i = 0 \\ \sim \text{Bernoulli}(p_{ij}) & \text{if } D_i = 1 \end{cases} \\
&\text{Observed cancers:} && Y_i|D_i, T_{ij} \begin{cases} = D_i & \text{if } T_{i1} = 1 \text{ or } T_{i2} = 1 \\ = 0 & \text{if } D_i = 0, T_{i1} = T_{i2} = 0 \\ \sim \text{Bernoulli}(\rho) & \text{if } D_i = 1, T_{i1} = T_{i2} = 0. \end{cases}
\end{aligned} \tag{1}$$

Each test j is assumed to be conditionally independent given a true disease status and follow-up medical procedures resulting from a positive test are assumed to accurately identify disease status.

As an aside, note that while T_{i1} and T_{i2} are assumed to be conditionally independent given a true disease status, with $pr(T_{i1} = 1|D_i, T_{i2} = 1) = pr(T_{i1} = 1|D_i, T_{i2} = 0)$, the assumption does not hold when conditioning on the observed cases Y_i instead of the latent true disease status D_i. As a consequence, the relationship between the probability of an ‘observed’ true positive $pr(T_{i1} = 1|Y_i = 1)$ and an ‘actual’ true positive $pr(T_{i1} = 1|D_i = 1)$ is non-trivial and related to the properties of the second test administered, regardless of the ordering of tests being applied. The ratio of the odds corresponding to these two probabilities is derived in Appendix A.1 and is equal to

$$\begin{aligned}
\frac{pr(T_{i1} = 1|D_i = 1)}{pr(T_{i1} = 0|D_i = 1)} \Big/ \frac{pr(T_{i1} = 1|Y_i = 1)}{pr(T_{i1} = 0|Y_i = 1)} &= pr(Y_i = 1|T_{i1} = 0, D_i = 1) \\
&= \rho + p_{i2} - \rho p_{i2}.
\end{aligned} \tag{2}$$

When $\rho = 1$, with cancers missed by screening tests always being subsequently observed during the follow-up period, this ratio is 1.0 and conditioning on observing cancer is equivalent to conditioning on an individual’s true cancer status. When both $\rho < 1$ and $p_{i2} < 1$, however, this ratio will always be less than 1.0 and the ‘observed’ odds in the denominator will overestimate the ‘actual’ odds in the numerator by an amount depending on both ρ and p_{i2}.

When the true positive probability of the second test does not vary between individuals, with $p_{i2} = p_{02}$, the previous ratio will be constant for all i. In this scenario, the observed cases (having $Y_i = 1$) would represent instead of an equi-probable random thinning of all cancer cases (having $D_i = 1$), and a logistic regression for $pr(T_{i1} = 1|Y_i = 1)$ would yield inferences on log-odds ratios that would provide valid statements about $pr(T_{i1} = 1|D_i = 1)$. When the second test is more accurate for some individuals than others, and p_{i2} varies with i, making inference on true positive rates $pr(T_{i1} = 1|D_i = 1)$ based on observed cancer status Y_i is much less straightforward. A covariate such as an individual’s age or breast density could influence $pr(T_{i1} = 1|Y_i = 1)$ via both p_{i1} and p_{i2}, and some knowledge of the ρ parameter would be required to distinguish between these two mechanisms.

Returning to the model specification, the second level of the hierarchy assigns mixed-effects logistic relationships to the true and false positive probabilities p_{ij} and q_{ij} and a linear logistic expression for the cancer probabilities ψ_i. The terms in this level are

(1) α_0 and α, the intercept and vector of regression coefficients governing the probability ψ_i that individual i has cancer;

(2) pairs of intercepts and coefficients (μ_{1j}, β_{1j}) and (μ_{2j}, β_{2j}) giving the intercept and regression coefficients for the false positive and true positive test outcomes, respectively; and
(3) two random effects η_{sje} and θ_{sje} for each of the two examiners $(j = 1, 2)$ in question.

Two individuals seen by the same radiologist will not necessarily have had a CBE nurse in common, meaning that a notation where individuals i are nested within examiners e is not possible. The double-subscripted random effect $\eta_{s_ije_{ij}}$ is used when individual i's exam of type j is being specified, whereas η_{sje} is used when an examiner (but not a particular exam on a specific individual) is under consideration. The resulting model specification is

$$\begin{aligned}
\text{Cancer:} \qquad\qquad D_i &\sim \text{Bernoulli}\left(\psi_i\right) \\
\text{logit}\left(\psi_i\right) &= \alpha_0 + X_i\alpha \\
\text{False positives: } T_{ij} = 1|D_i = 0 &\sim \text{Bernoulli}\left(q_{ij}\right) \\
\text{logit}\left(q_{ij}\right) &= \mu_{1j} + X_i\beta_{1j} + \theta_{s_ije_{ij}} \\
\text{True positives: } T_{ij} = 1|D_i = 1 &\sim \text{Bernoulli}\left(p_{ij}\right) \\
p_{ij} &= 1 - \left(1 - r_{ij}\right)\left(1 - q_{ij}\right) \\
\text{logit}\left(r_{ij}\right) &= \mu_{2j} + X_i\beta_{2j} + \eta_{s_ije_{ij}}.
\end{aligned} \tag{3}$$

The relationship between p_{ij}, q_{ij}, and r_{ij} can be rewritten as an odds ratio $1 - r_{ij} = (1 - p_{ij})/(1 - q_{ij})$, and r_{ij} can be interpreted as the information conveyed to a test from an individual's true cancer status. A value of $r_{ij} = 0$ implies the two probabilities p_{ij} and q_{ij} are equal and the true cancer status D_i has no effect on test probabilities once X_i and $\theta_{s_ije_{ij}}$ are accounted for. An effective test should respond to cancer status by increasing the probability of a positive test when $D_i = 1$, and r_{ij} quantify this increase in a way that ensures $0 \leqslant q_{ij} \leqslant p_{ij} \leqslant 1$. The extreme case where $r_{ij} = 1$ implies a perfectly sensitive test with $p_{ij} = 1$ irrespective of q_{ij}. Further explanation and interpretation of this parametrization are given in Appendix A.2.

The third level of the model specifies the distributions of the examiner-level random effects η_{sje} and θ_{sje} from (3), with these distributions depending on site-level random effects κ_{sj} and λ_{sj}. It is assumed that these random effects are multivariate normal with following distributions:

$$\begin{aligned}
\begin{pmatrix}\eta_{sje}\\ \theta_{sje}\end{pmatrix} &\sim \text{N}\left(\begin{pmatrix}\kappa_{sj} + R_{sje}\gamma_{1j}\\ \lambda_{sj} + R_{sje}\gamma_{2j}\end{pmatrix}, \Sigma_j\right) \\
\begin{pmatrix}\kappa_{sj}\\ \lambda_{sj}\end{pmatrix} &\sim \text{N}\left(\begin{pmatrix}Q_s\delta_{1j}\\ Q_s\delta_{2j}\end{pmatrix}, \Gamma_j\right).
\end{aligned} \tag{4}$$

As mentioned previously, the R_{sje} and Q_s are vectors of covariates associated with examiners and health facilities (although no examiner-level covariates are used in this application). These covariates have regression coefficient parameters γ_{1j} and δ_{1j} affecting the true positive probabilities and γ_{2j} and δ_{2j} affecting the false positive probability of test j. Each test j has a variance matrix Σ_j for the examiner-level random effects and Γ_j for the site-level random effects. The fourth and final level to the model specifies the prior distributions of model parameters, as detailed in Section 2.2.

The previous formulation assumes conditional independence between the two tests, an assumption which can be relaxed through the inclusion of an additional random effect term. A random effect M_i can induce correlation between exams when added to the model for r_{ij} as

$$\begin{aligned}
\text{logit}(r_{ij}) &= \mu_{3j} + X_i\beta_{3j} + \eta_{js_ie_{ij}} + M_i \\
M_i &\sim \text{N}\left(0, \sigma^2\right).
\end{aligned} \tag{5}$$

2.2. *Prior distributions*

Informative priors are used for the intercept parameters μ_{1j} and μ_{2j} for the true detection probabilities for mammography and CBE, respectively. The American Cancer Society guidelines [28] list sensitivities for mammography from various clinical studies with a range from 0.64 to 0.84. Smith *et al.* [28] also cite the meta-analysis of Barton *et al.* [29] who estimate the sensitivity CBE with a 95% confidence interval between 0.48 and 0.60. As the intercept parameter μ_{2j} for the true positive probability p_{i1} in (3) relates to true positives in excess of the false positive rate, the priors for μ_{3j} target slightly lower probabilities than the aforementioned values.

- A prior of $\mu_{31} \sim \mathrm{N}(0.63, 0.24^2)$ is used for the intercept of mammography's true detection rate, giving a 95% prior interval for the baseline referral probability between 0.54 and 0.75 [28].
- The prior $\mu_{32} \sim \mathrm{N}(0.08, 0.16^2)$ is used for the intercept of CBE, giving a 95% interval for its true positive probability between 0.44 and 0.60 [29].
- The fixed-effects parameters β, γ, and δ refer to log-odds ratios for a 10-year change in age, one standard deviation change in the other continuous variables, or the presence/absence of a binary variable. For those parameters, prior distributions of $\mathrm{N}(0, 2^2)$ were chosen so that a change of 4 on the logit scale roughly corresponds to a change in probabilities from 0.5 to 0.99.
- A prior of $\mathrm{N}(-4, 1^2)$ was used for the intercept of the cancer model μ_1 to give a 95% prior interval between 2.5 and 12 cases per 1000 women [2].
- The missed cancer probability ρ is assumed to have an uninformative uniform prior of $\mathrm{Beta}(1.2, 1.2)$ giving a 95% prior interval of 0.04 and 0.96.

Priors of cancer personal-level risk factors were motivated by results from [30] and [31]. Family history of breast cancer has a prior of $\mathrm{N}(0.95, 0.21^2)$, giving a 95% interval for the odds ratio of (1.7, 3.9). For breast density, an informative prior is used with a distribution of $\mathrm{N}(0.75, 0.17^2)$ corresponding to a 95% interval for the odds ratio of (1.5, 3.0).

The inverse Wishart distribution was assumed for the variance matrices, with Γ^{-1} and Σ^{-1} distributed as $\mathrm{Wishart}(I/20, 6)$, giving 95% intervals for the standard deviations (square roots of diagonals of the variance matrices) of (0.06,0.25). While this is an informative prior concentrated at small values, the random effects operate on the log-odds scale, and the upper limit allows for a substantial amount of between-examiner variation. Consider a standard deviation of 0.25, giving an examiner in the 97.5 percentile an odds of $\exp(\kappa_{sj} + 1.96 \cdot 0.25)$ for producing a positive test and an examiner of the same type j at the same site s in the 2.5 percentile a corresponding odds of $\exp(\kappa_{sj} - 1.96 \cdot 0.25)$. The ratio of the 97.5th percentile odds to the 2.5th percentile odds is $\exp(2 \cdot 1.96 \cdot 0.25) \approx 2.7$, or in other words, these examiners would have odds for positive tests varying by factor of nearly 3. The lower end of the prior, where the standard deviation is 0.06, yields very little variation between examiners with the corresponding odds ratio between the 97.5th and 2.5th percentile being a fairly modest 1.25.

The families of distributions used for the priors were chosen because of their conjugacy, meaning that the conditional distributions of parameters have closed forms when possible. The Wishart distributions are conjugate with the Gaussian random effects, the beta prior on ρ is conjugate with the binary Y_i, and the Gaussian priors on the site-level covariates are conjugate with the Gaussian random effects. The choice of Gaussian distributions for the random effects was likewise a pragmatic decision.

2.3. *MCMC algorithm*

Inference on the model is performed with an MCMC algorithm for sampling from the posterior distributions of the parameters and the latent variables. The Bayesian inference via MCMC is used partly because of the difficulty in computing marginal probabilities with random effects and latent variables and partly in order to allow for the use of informative priors. A Gibbs sampling routine is described in detail in Appendix A.3, with each iteration n consisting of the steps given in the following.

- First, any unknown cancer status variables $D_i^{(n)}$ are sampled from the full conditional distribution given Y_i and T_{ij} and using parameter values $p_{ij}^{(n-1)}$, $q_{ij}^{(n-1)}$, and $\rho^{(n-1)}$ from the previous iteration.
- Second, the $\alpha^{(n)}$, $\beta^{(n)}$, $\mu^{(n)}$, and $\psi^{(n)}$ parameters and examiner-level random effects $\theta^{(n)}$ and $\eta^{(n)}$ receive the Metropolis–Hastings updates [32].
- Third, examiner-level covariances $\Sigma_j^{(n)}$ are sampled directly from the Wishart distributions using their conjugate Wishart priors, examiner-level random effects $\theta^{(n)}$ and $\eta^{(n)}$, and the previous iteration's examiner-level random effects $\kappa^{(n-1)}$ and $\lambda^{(n-1)}$.
- Fourth, site-level random effects $\kappa^{(n)}$ and $\lambda^{(n)}$ and coefficients δ are sampled directly from the Gaussian conditional distributions.
- Finally, site-level variances $\Gamma_j^{(n)}$ are drawn from the conditional Wishart distributions.

The algorithm is coded in R [33], and the code and synthetic data are available in http://pbrown.ca/screening. The package 'glmmBUGS' [34] was used for generating starting values, assuming no unobserved cancers and fitting four univariate models (true positive and true negative outcomes for each of the two tests). The starting values for α in the cancer model parameters are set to their maximum likelihood estimates from a logistic regression model. Five parallel MCMC chains were run using different values for the intercept α_0 of the cancer model between −4 and −3 (in increments of 0.2), values which exceed the maximum likelihood estimates of −4.8 and thereby inducing unobserved additional cancers. Each of the five chains was run for 3100 iterations, with the first 100 iterations discarded as burn-in and the results shown use every 15th iteration. Chain convergence was assessed through trace plots and the Gelman diagnostic plots and tests [35] from the 'coda' package [36] in R.

2.4. Sensitivities, specificities, and joint probabilities

Cancer researchers and planners are often more concerned with sensitivities and specificities associated with a particular screening regimen than with any of the parameters or random effects comprising the model in Section 2.1. Sensitivity and specificity for a single test j performed on an individual i are defined as $pr(T_{ij} = 1|D_i = 1) = \mathrm{E}(p_{ij})$ and $pr(T_{ij} = 0|D_i = 0) = 1 - \mathrm{E}(q_{ij})$, respectively. Assuming the 'either positive' rule for the combination of tests, a pair of tests performed on individual i has combined accuracy values of

$$\begin{aligned} &\text{sensitivity: } pr\left(T_{i1} = 1 \text{ or } T_{i2} = 1 |\, D_i = 1\right) = \mathrm{E}\left[1 - \left(1 - p_{i1}\right)\left(1 - p_{i2}\right)\right] \\ &\text{specificity: } pr\left(T_{i1} = 0 \text{ and } T_{i2} = 0 |\, D_i = 0\right) = \mathrm{E}\left[\left(1 - q_{i1}\right)\left(1 - q_{i2}\right)\right]. \end{aligned} \tag{6}$$

The earlier expectations are due to the p_{ij} and q_{ij} being random quantities, depending on site-level and examiner-level random effects η_{sje}, θ_{sje}, κ_{sj}, and λ_{sj}. The expectations in (6) are non-linear combinations of the model parameters, and the ease with which posterior samples of these quantities can be assembled from MCMC output is an additional argument for the Bayesian–MCMC inference methodology adopted here.

How the random effects are treated depends on whether the question of interest concerns a specific pair of examiners e_{01} and e_{02} at a given health facility s_0 in the dataset, or rather the outcomes expected from consulting a random or hypothetical pair of examiners and averaging out the variations among different examiners and screening sites. This latter scenario concerns the evaluation of a screening program as a whole, whereas the first question would be useful for evaluating individual examiners and understanding the variation among these examiners.

When considering a specific examiner, the expectations in (6) are computed using joint posterior samples of all model parameters and the examiner's random effect variables. For example, specificity estimated for test j by examiner e_{0j} at site s_0 on an individual with covariates X_0 is the sample mean of a set of $p_{0j}^{(n)}$ computed from $n = 1 \ldots N$ posterior MCMC samples $\mu_{1j}^{(n)}$, $\beta_{1j}^{(n)}$, and $\theta_{s_0 j e_{0j}}^{(n)}$ as

$$\operatorname{logit}\left(q_{0j}^{(n)}\right) = \mu_{1j}^{(n)} + \theta_{s_0 j e_{0j}}^{(n)} + X_0 \beta_{1j}^{(n)}. \tag{7}$$

When an average or hypothetical examiner is desired, the expectation is with respect to the posterior distribution of the model parameters and the unconditional distribution of the random effects given these parameters. Random effects $\theta_{sje}^{(n)}$ used to obtain $q_{0j}^{(n)}$ in (7), for example, would be drawn from the distributions

$$\begin{pmatrix} \kappa_{sj}^{(n)} \\ \lambda_{sj}^{(n)} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} Q_0 \delta_{1j}^{(n)} \\ Q_0 \delta_{2j}^{(n)} \end{pmatrix}, \Gamma_j^{(n)} \right) \text{ and } \begin{pmatrix} \eta_{sje}^{(n)} \\ \theta_{sje}^{(n)} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \kappa_{sj}^{(n)} + R_{j0} \gamma_{1j}^{(n)} \\ \lambda_{sj}^{(n)} + R_{j0} \gamma_{2j}^{(n)} \end{pmatrix}, \Sigma_j^{(n)} \right).$$

Notice that the calculation requires specifying covariates Q_0 (and possibly R_{j0}), although a distribution for sampling random $Q_0^{(n)}$ could be specified.

An individual i faced with a decision on whether to obtain a breast cancer screen would necessarily have an unknown cancer status D_i, and their decision would be better informed by joint probabilities of cancer status and test results than by probabilities conditional on cancer status. Of particular interest would be the probabilities

$$\begin{aligned} pr\left[\left(T_{i1} = 1 \text{ or } T_{i2} = 1\right) \text{ and } D_i = 1\right] &= \mathrm{E}\left\{\left[1 - \left(1 - p_{i1}\right)\left(1 - p_{i2}\right)\right]\psi_i\right\} \\ pr\left[\left(T_{i1} = 1 \text{ or } T_{i2} = 1\right) \text{ and } D_i = 0\right] &= \mathrm{E}\left\{\left[1 - \left(1 - q_{i1}\right)\left(1 - q_{i2}\right)\right]\left[1 - \psi_i\right]\right\}, \end{aligned}$$

corresponding to: (1) having a cancer and having it detected by screening; and (2) not having cancer and obtaining a false positive test result. A high probability of outcome 1 is evidence in favor of the individual i obtaining screening tests, because this outcome would likely lead to early cancer detection and a consequent improvement in life expectancy. Outcome 2 can cause emotional stress, inconvenience, cost to the healthcare system, and in extreme cases lead to serious adverse outcomes from subsequent medical treatment. A high probability of this second outcome would argue against this individual being screened, although individuals will differ in their risk aversion for this second probability. Unlike sensitivities and specificities, these joint probabilities depend on the cancer prevalence rate for type of individual concerned and hence the model parameters governing the prevalence. A posterior sample of the cancer probability $\psi_i^{(n)}$ is calculated for each MCMC sample and used in conjunction with the $q_{0j}^{(n)}$ and $p_{0j}^{(n)}$ to create posterior samples of the joint probabilities.

3. Results

Five parallel MCMC chains were used; each runs for 3100 iterations with the first 100 iterations discarded. Appendix B in the Supporting Information contains a full set of trace plots, Gelman diagnostic plots, and posterior densities.

3.1. Parameters and random effects

Table III(a) contains posterior distributions for each regression coefficient α, β, and δ. Four values are listed for each explanatory variable, with each type of test result (true positives $T_{ij} = 1|D_i = 1$ and false positives $T_{ij} = 1|D_i = 0$) having outcomes for each of the two screen tests (mammography when $j = 1$, and CBE for $j = 2$). The individual-level covariates X_i are breast density (binary high or low, with low as baseline), age (in decades), and family history as defined by a first-degree relative having had breast cancer (and with no family history as baseline). The site-level covariate Q_s is the number of screens given annually at the health facility, log transformed, and with the effect shown for the interquartile range. Note that all parameters have posterior distributions that are noticeably different from their non-informative prior distributions, indicating that the data have had a meaningful influence on determining the posteriors. Also, the posterior intervals are narrower for false positive outcomes, which is to be expected as there are many more false positives than true positive test results (cf. Table I).

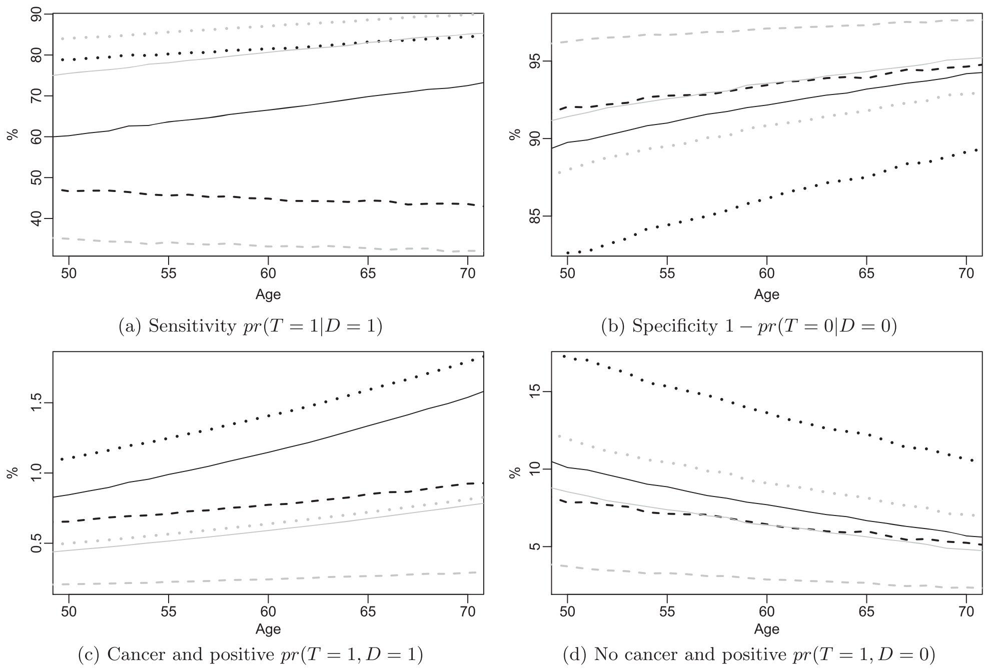
Breast density is the covariate with the strongest effect on test results, with high breast density having a pronounced effect on each of the four probabilities. High breast density increases the probability of a false positive for both screen tests, although the effect is much greater for CBE. The odds of a CBE false positive in the presence of high breast density are more than double the odds for a comparable woman with low breast density. The effect of high breast density on the true positive process is different for each screen test, with the odds of detecting a cancer with mammography on a woman with high breast density being roughly half of the comparable odds when breast density is low. Breast density increases the odds of cancer detection by CBE, and note that this true positive odds ratio reflects increased detection in excess of the corresponding increase in the false positive rate. Older women have lower odds of false positives and higher odds of a true positive than younger women, a result which applies to both screening modalities.

Table III(b) shows the prior and posterior distributions for the standard deviations and correlations of the random effects. Standard deviations for eight different random effects are shown: four at the examiner level sd[η_{sje}] and sd[θ_{sje}] for $j = 1, 2$; and four at the site level sd[κ_{sj}] and sd[λ_{sj}]. The amount of variation between sites and examiners is substantial, for the most part, exceeding the upper 97.5% prior quantile of 0.25. The posterior credible intervals of the standard deviation parameters nearly all contain 0.4, a value which would yield random effects in the range of -0.8 to 0.8 and odds ratios for pairs of sites or examiners as large as $e^{0.8}/e^{-0.8} \approx 5$. The heterogeneity of examiners and sites appears to be greater for nurse exams than for mammography.

There are four correlation parameters shown in Table III(b), showing the correlation between the odds of true positives and false positives cor[η_{sje}, θ_{sje}] (examiners) and cor[$\kappa_{sj}, \lambda_{sj}$] (sites) for each screening modality ($j = 1, 2$). The positive correlation between λ_{s2} and κ_{s2} (at the site level for nurse exams) is consistent with screening sites having different tolerances or thresholds for declaring CBE tests as abnormal. A screening site with a low 'abnormal' threshold could achieve an impressive true positive rate at the cost of a higher burden of false positives. There is no clear indication that this trade-off is manifested at the site level or in relation to mammography, although the uncertainty associated with the estimates of correlations is considerable.

Table III. Posterior means and 95% credible intervals for model parameters governing true positive and false positive test outcomes.

	Mean	2.5%	97.5%
(a) Odds ratios for breast density, a 1 standard deviation (or 5.5 year) change in age, having a family history of breast cancer, and for the interquartile range of the annual screening volume of screening sites.			
Nurse exam, false positive β_{22}			
Site volume δ_{22}	0.96	0.79	1.15
Family history	1.12	1.03	1.21
Age	0.94	0.91	0.96
Breast density	2.32	2.13	2.51
Mammography, false positive β_{21}			
Site volume δ_{21}	1.03	0.92	1.16
Family history	1.04	0.99	1.10
Age	0.93	0.91	0.95
Breast density	1.33	1.23	1.43
Nurse exam, true positive β_{12}			
Site volume δ_{12}	0.99	0.86	1.14
Family history	1.03	0.74	1.40
Age	1.13	1.00	1.27
Breast density	2.05	1.31	2.99
Mammography, true positive β_{11}			
Site volume δ_{11}	1.00	0.93	1.08
Family history	1.20	0.84	1.70
Age	1.24	1.07	1.41
Breast density	0.41	0.28	0.61
Prior distribution	7.46	0.02	50.90
(b) Random effect standard deviations and correlations at the site level and examiner level.			
Nurse exam, examiner level			
False positive sd(θ_{s2e})	0.49	0.41	0.57
True positive sd(η_{s2e})	0.49	0.40	0.59
Correlation(θ_{s2e}, η_{s2e})	0.07	−0.16	0.29
Mammography, examiner level			
False positive sd(θ_{s1e})	0.42	0.37	0.48
True positive sd(η_{s1e})	0.42	0.35	0.50
Correlation(θ_{s1e}, η_{s1e})	0.14	−0.10	0.35
Nurse exam, site level			
False positive sd(λ_{s2})	0.54	0.39	0.69
True positive sd(κ_{s2})	0.44	0.26	0.68
Correlation($\kappa_{s2}, \lambda_{s2}$)	0.78	0.40	0.97
Mammography, site level			
False positive sd(λ_{s1})	0.30	0.22	0.40
True positive sd(κ_{s1})	0.17	0.08	0.31
Correlation($\kappa_{s1}, \lambda_{s1}$)	0.29	−0.62	0.87
Prior distributions			
Standard deviations	0.12	0.06	0.25
Correlations	0.00	−0.78	0.74



(a) Sensitivity $pr(T = 1|D = 1)$

(b) Specificity $1 - pr(T = 0|D = 0)$

(c) Cancer and positive $pr(T = 1, D = 1)$

(d) No cancer and positive $pr(T = 1, D = 0)$

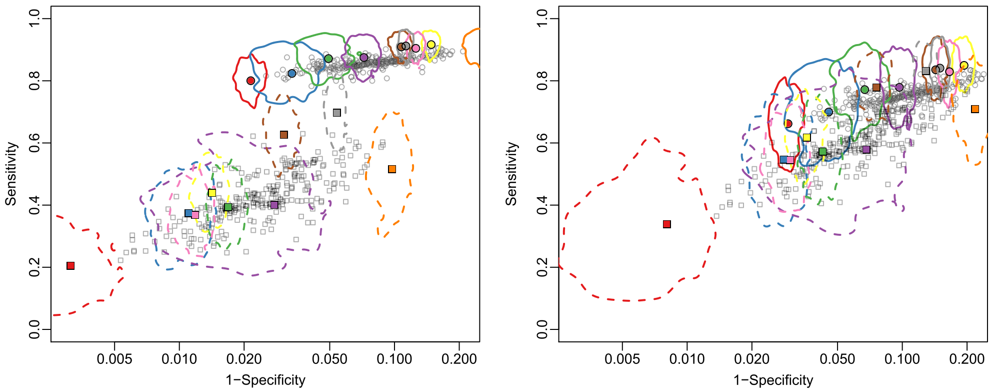
Figure 2. Sensitivity and specificity for mammography (—), clinical breast exam (- - -), and both (. . .) for high-risk women with high breast density and a family history of cancer (black lines) and low-risk women with low density without family history (gray lines).

3.2. *Sensitivities and specificities*

Screening outcomes are affected by individual characteristics such as breast density and family history, and results are shown in the following for women in a ‘high-risk’ group (with high breast density and having a first-degree relative with breast cancer) and a ‘low-risk’ group (having neither). Figure 2(a) and (b) shows predicted sensitivity and specificity of screening tests involving CBE and/or mammography as a function of an individual’s age for the two groups. Mammography (solid lines) has consistently higher sensitivity than CBE (dashed lines), although specificity is somewhat lower for the former. Unsurprisingly, undergoing both exams (dotted lines) increases sensitivity and decreases specificity slightly. Tests on women in the high-risk group (black lines) have lower specificity than tests on women in the low-risk group (gray lines) for both screening modalities, while sensitivity in the former group is lower for mammography and higher for CBE.

Figure 2(c) and (d) shows joint probabilities of an individual having cancer and obtaining a positive screening test (Figure 2(c)) and an individual not having cancer incurring a false positive test (Figure 2(d)). The joint cancer true positive probability in Figure 2(c) increases and the joint non-cancer false positive probability in Figure 2(d) decreases for both risk profiles. A pertinent practical question these figures can inform is whether an individual should opt for obtaining both CBE and mammography or forgo CBE in favor of mammography alone. Women in the high-risk group have a substantially higher joint cancer true positive probability with two exams than when undergoing mammography alone, as evidenced by the large gap between the dotted and solid black lines in Figure 2(c). The increase in this joint probability for low-risk women, shown by the dotted and solid gray lines, is slight. The rationale for undergoing both exams is therefore stronger for the high-risk women, although the combination of exams also increases the undesirable joint false positive probability in Figure 2(d).

The nature of variability among examiners can be inferred from the posteriors of covariance parameters in Table III(b), although a more intuitive illustration using sensitivities and specificities is presented in Figure 3. Each plotting symbol in Figure 3 shows the posterior mean of the sensitivity and specificity for an individual examiner seeing a 65-year-old individual, with Figure 3(a) relating to low-risk (low breast density, no family history) individuals with probabilities for high-risk individuals (high density, with a family history) shown in Figure 3(b). For an arbitrary selection of examiners, 75% posterior credible regions are shown as contour lines. Notice the clear separation between mammography and CBE, with mammography examiners having uniformly higher sensitivity than CBE and specificity, which is on



(a) Low breast density, without family history

(b) High breast density, with family history

Figure 3. Posterior means for sensitivity and specificity of individual examiners screening a 65-year-old patient with clinical breast exam (CBE) (□) and mammography (∘). Contour lines show 75% posterior credible regions and their corresponding posterior means for CBE (dashed lines) and mammography (solid lines). Note that the x-axis is on the log scale.

average lower. Also note that sensitivity is fairly stable among examiners (although more variable for CBE), whereas specificity is extremely heterogeneous. The differences in accuracy among examiners is therefore most evident in their false positive rates, with the cancer-detecting abilities of the most and least effective examiners being reasonably similar.

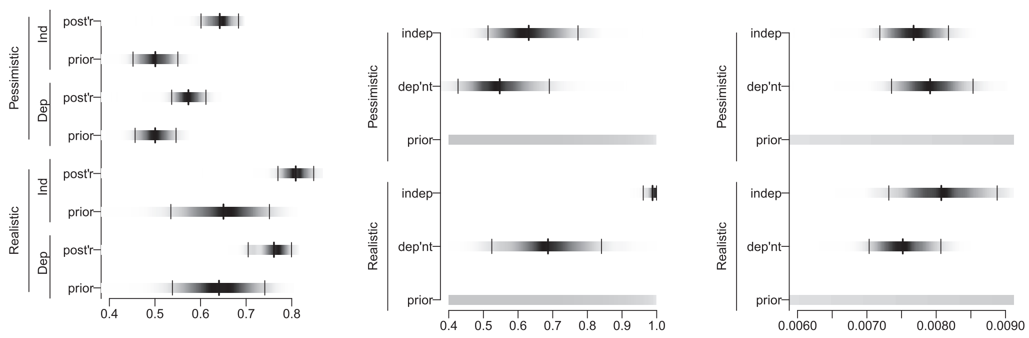
3.3. Prior sensitivity

To assess the influence and importance of prior distributions and the assumption of independence between screen types, posterior distributions from four variations on the screening model were computed. A ‘Pessimistic’ model specifies a prior distribution for the μ_1 parameter, which results in a true positive probability for mammography on a baseline individual tightly distributed around 50%, as opposed to the 54% to 73% range under the ‘Realistic’ prior assumptions in Section 2.2. The independence assumption was relaxed by fitting the ‘Dependence model’ from (5), with an individual-level random effect allowing for the possibility that a tumor that escapes detection from one screening modality is also more likely to escape detection by a second. This model and the standard ‘Independence model’ were fit using both the Realistic and Pessimistic sets of prior distributions. Figure 4(a) shows the distribution of the true positive probability $pr(T_{ij} = 1|D_i = 1)$ for mammography, on an individual with baseline covariates, from both prior and posterior samples. The priors for the dependence model differ slightly from the independence model as for the former the individual-level random effect is integrated out. Figure 4(b) shows the posterior probability for the proportion ρ of cancers missed by screening which are subsequently detected by other means. The baseline cancer rates on the natural scale are contained in Figure 4(c).

In each of the four models, mammography sensitivity is predicted to be at the upper end of the prior distribution. The Dependence model leads to estimates that are lower and closer to the priors than the independence model, a somewhat intuitive result as the additional σ parameter allows for greater flexibility. Also unsurprising is a lower posterior sensitivity in a model is accompanied by a higher number of unobserved cancers (a lower ρ) and a higher cancer rate. As the number of missed cancers is small, the baseline cancer rate needs only shift by a small amount in order to accommodate lower sensitivity.

One conclusion to draw from Figure 4 is that the choice of prior does affect the inferences made regarding sensitivity of a screening test when the number of unobserved cancers is unknown. The posterior distribution for ρ suggests an analysis of the OBSP data using a simpler model with an implicit assumption of all cancers being observed is not unreasonable if independence and the realistic priors are also assumed. However, one would not be able to assess the appropriateness of this simpler model without first having carried out an analysis where the number of unobserved cancers is allowed to vary.

With respect to the independence assumption, Figure 4(a) suggests that this assumption does not artificially inflate the estimate of sensitivity if one accepts the findings of clinical research on mammography sensitivity (the Realistic priors). Were the Pessimistic prior assumptions to be believed, the data would be inconsistent with the independence assumption and an individual-level random effect in the Dependence model would be required to induce a number of missing and unobserved cancers.



(a) Probability of true positives $pr(T_{ij} = 1|D_i = 1)$ for mammography on an individual in the baseline group.

(b) Probability ρ of of subsequent clinical observation of screen-undetected cancers.

(c) Probability of cancer incidence for a baseline individual $\exp(\alpha_0)/[1 + \exp(\alpha_0)]$.

Figure 4. Prior and posterior distributions from two sets of prior distributions (Realistic and Pessimistic) being used with each of the two cancer detection models. The dependent (Dep) model includes an individual-level random effect, inducing correlation between two screens carried out on the same individual, with the independent (Ind) model assuming the two screens are uncorrelated when the individual's true cancer status is conditioned upon.

4. Discussion

With electronic medical records and linked-health databases becoming increasingly common, large population-based observational datasets similar to the OBSP records will become increasingly available. The individuals in such databases are heterogeneous, and the outcome variables can be incomplete, although this paper has demonstrated with the OBSP data that explanatory variables, random effects, and latent variables for disease status can accommodate these complexities. Understanding the sources of variation in large screening databases can yield useful information for managing and improving screening programs, one example of which is provided by the estimates of examiner-level variance parameters. Examiner-level variations in false positive rates are substantially greater than the corresponding variations in true positive rates, implying that training and resources would be most effective when targeted at those examiners for whom false positive rates are the highest. Quantifying the influence of individual-level characteristics on screening outcomes can aid the development of individualized cancer screening policies and recommendations. Current guidelines from the American Cancer Society and the Canadian Cancer Society make breast cancer screening recommendations based on a woman's risk factor profile. The impact a woman's personal characteristics have on test sensitivity and specificity could also be reflected in screening recommendations with, for example, low-risk individuals encouraged to obtaining screening only if their predicted false positive probability is small.

The problem of estimating both cancer detection rates and the prevalence of unobserved cancers from the same dataset is addressed through the use of prior information on test accuracy from clinical studies. The model used involves a fundamental non-identifiability. Some of these non-identifiabilities are negated when more than one screening test is administered and an assumption of independence between screens is made, although Dendukuri and Joseph [23] discuss in some detail how substantial prior sensitivity remains when the number of screening tests is less than three. Informative prior distributions are therefore necessary for some of the model parameters. Inferences using this model are sensitive to these prior assumptions, with the 95% posterior credible interval (1%, 25%) for the proportion of screen-undetected cancers that are unobserved during follow-up changing to (25%, 50%) when a pessimistic prior positing a low true positive rate is substituted. Drawing conclusions from an analysis that relies on subjective prior information does provoke skepticism, although conclusions from an identifiable model with an implicit assumption that all screen-undetected cancers are observed should also be treated skeptically.

There can be subtleties involved with incorporating prior information into complex random effects models, with Menten *et al.* [37] demonstrating how different dependence models may result in similar fits to the data while resulting in different inferences. They concluded that the selection of appropriate latent-class models should be based on substantive subject matter knowledge. Our modeling framework can assist in this regard by explicitly separating true and observed disease status (D_i and Y_i) and specifying a model for test results conditional on true cancer status. To some degree, sensitivity to model specification is explored with the inclusion of a random effect term allowing for dependence between screening tests performed on the same individual. Although this dependence term results in a substantial increase in the number of unobserved cancers predicted (from 1–20% to 20–40%), the increase is not sufficiently large to cause more than a modest effect on the posterior for test sensitivity. There is certainly scope for a more detailed exploration of the possible dependence structures between screening tests with population-level data, such as the negative correlation between mammography and CBE noted by Walter *et al.* [38], although the fundamental non-identifiability expressed by Dendukuri and Joseph [23] is a substantial limitation for research in this regard.

A number of model extensions would be possible to more faithfully reflect the underlying biological processes of cancer incidence and screening outcomes. Variation in cancer risk throughout the population could be accommodated by including additional covariates or random effects terms in the model for ψ_i and testing for bimodality, or skewness in examiner-level random effects would result from using more flexible distributions or mixtures in place of the Gaussian. Such extensions are conceptually straightforward, and the size of the dataset would likely support the identification of these additional parameters. The computational challenges would be greatly increased, however, and a more sophisticated MCMC updating scheme would be required in place of the Metropolis-within-Gibbs algorithm used here.

Finally, no discussion of cancer screening methodology can avoid addressing the controversy regarding the degree to which cancer detection via screening ultimately provides a health benefit [e.g., 39]. Our model provides an evaluation of a screening program or test based on the simplest and most immediately obtainable success criterion, which is detection of lesions leading to a clinical diagnosis of cancer. However, the accommodation of heterogeneities among subjects and examiners in this model could form the basis of extended and expanded models for a more complex set of hypotheses and health outcomes. For example, the degree to which screen detection of cancers is overdiagnosis could be assessed by exploring the relationship between examiners' test sensitivity and the long-term health status of the individuals they screen. Although the variation in examiners' sensitivities is shown here to be modest (in the range of 10% to 20%), the large sample sizes in population-level datasets with random assignment of individuals to examiners within centers suggest that such an analysis would be insightful.

Appendix A: Model details

A.1. Odds ratio for actual versus observed true positives

First, note that $pr(D_i = 1|Y_i = 1) = 1$, so

$$pr\left(T_{i1} = 1|Y_i = 1\right) = pr\left(T_{i1} = 1|D_i = 1, Y_i = 1\right) pr\left(D_i = 1|Y_i = 1\right) = pr\left(T_{i1} = 1|D_i = 1, Y_i = 1\right).$$

Also, $pr(Y_i = 1|T_{i1} = 1, D_i = 1) = 1$ and

$$pr\left(T_{i1} = 1, Y_i = 1|D_i = 1\right) = pr\left(Y_i = 1|T_{i1} = 1, D_i = 1\right) pr\left(T_{i1} = 1|D_i = 1\right) = pr\left(T_{i1} = 1|D_i = 1\right).$$

Applying Bayes theorem

$$pr\left(T_{i1}|Y_i = 1, D_i = 1\right) = pr\left(T_{i1}, Y_i = 1|D_i = 1\right) / pr\left(Y_i = 1|D_i = 1\right)$$

yields an odd

$$\frac{pr\left(T_{i1} = 1|Y = 1\right)}{pr\left(T_{1i} = 0|Y = 1\right)} = \frac{pr\left(T_{i1} = 1, Y_i = 1|D_i = 1\right) / pr\left(Y_i = 1|D_i = 1\right)}{pr\left(T_{i1} = 0, Y_i = 1|D_i = 1\right) / pr\left(Y_i = 1|D_i = 1\right)} = \frac{pr\left(T_{i1} = 1|D_i = 1\right)}{pr\left(T_{i1} = 0, Y_i = 1|D_i = 1\right)}.$$

Inserting this odd into the odds ratio gives

$$\frac{pr\left(T_{i1} = 1|D_i = 1\right)}{pr\left(T_{i1} = 0|D_i = 1\right)} \bigg/ \frac{pr\left(T_{i1} = 1|Y_i = 1\right)}{pr\left(T_{i1} = 0|Y_i = 1\right)} = \frac{pr\left(T_{i1} = 0, Y_i = 1|D_i = 1\right)}{pr\left(T_{i1} = 0|D_i = 1\right)}$$
$$= pr\left(Y_i = 1|T_{i1} = 0, D_i = 1\right),$$

with the last step resulting from

$$pr\left(T_{i1}=0, Y_i=1|D_i=1\right)=pr\left(Y_i=1|T_{i1}=0, D_i=1\right)pr\left(T_{i1}=0|D_i=1\right).$$

Integrating out the possible values of T_{i2} gives

$$\begin{aligned}pr\left(Y_i=1|T_{i1}=0, D_i=1\right)=&pr\left(Y_i=1|T_{i2}=1, T_{i1}=0, D_i=1\right)pr\left(T_{i2}=1|T_{i1}=0, D_i=1\right)+\\&pr\left(Y_i=1|T_{i2}=0, T_{i1}=0, D_i=1\right)pr\left(T_{i2}=0|T_{i1}=0, D_i=1\right)\\=&pr\left(T_{i2}=1|T_{i1}=0, D_i=1\right)+\rho pr\left(T_{i2}=0|T_{i1}=0, D_i=1\right)\\=&pr\left(T_{i2}=1|D_i=1\right)+\rho pr\left(T_{i2}=0|D_i=1\right)\\=&p_{i2}+\rho(1-p_{i2})\end{aligned}$$

A.2. *Parametrization of true positive rates*

As an alternative to modeling the true positive and false positive, consider the following:

- W_{ij} represents an examiner correctly identifying a cancerous lesion on an individual, modeled $W_{ij}|D_i \sim \text{Bernoulli}(r_{ij})$ when $D_i=1$ and $W_{ij}=0$ whenever $D_i=0$; and
- $V_{ij} \sim \text{Bernoulli}(q_{ij})$ represents an examiner declaring a test to be positive for any other reason, likely from concern over a feature that would not be confirmed as cancerous from further medical procedures.

An individual without cancer can only test positive with $V_{ij}=1$. An individual with cancer can obtain a positive test in one of two ways: the examiner detects the cancer ($W_{ij}=1$); or the examiner does not detect the cancer ($W_{ij}=0$) but assigns a positive result because of some other issues ($V_{ij}=1$). Assuming V_{ij} is independent of W_{ij}, the probability of a positive test of an individual with cancer is

$$pr\left(W_{ij}=1 \text{ or } V_{ij}=1\right)=1-pr\left(W_{ij}=0 \text{ and } V_{ij}=0\right)=1-\left(1-r_{ij}\right)\left(1-q_{ij}\right).$$

The earlier equation is identical to the expression of p_{ij} in (3).

As a further motivation for this parametrization, consider an examiner with tendency to give positive tests to a large proportion of cancer-free individuals with $q_{ij}=0.9$. This examiner should have a true positive rate of at least 0.9, and a true positive rate of 0.91 would not indicate exceptional cancer-detecting abilities. In contrast, an examiner having $q_{ij}=0.01$ and having positive tests for 91% of patients with cancer should be regarded as exceptional. Making inference on the covariates influencing r_{ij} as opposed to p_{ij} leads to coefficients being interpretable as ‘cancer-detecting ability’ independently of the covariate’s effect on the false positive rate.

A.3. *The algorithm*

At each iteration n, the data augmentation step samples true cancer status $D_i^{(n)}$ and possible ‘accidental’ positives $V_{ij}^{(n)}$ (Appendix A.2) conditional on the parameters and random effects. Unobserved cancers are sampled from the distribution

$$pr\left(D_i=1|T_{ij}=0, Y_i=0\right)=\frac{(1-\rho)(1-p_{i.})\psi_i}{(1-\rho)(1-p_{i.})\psi_i+(1-\psi_i)}$$

with $p_{i.}=pr(T_{i1}=1 \text{ or } T_{i2}=1|D_i=1)=1-(1-p_{i1})(1-p_{i2})$. Accidental positives for screen-detected cancers sampled with

$$pr\left(V_{ij}=1|T_{ij}=1, D_i=1\right)=\frac{q_{ij}-\psi p_{ij}q_{ij}}{\psi(1-(1-p_{ij})(1-q_{ij})}.$$

The observation-level fixed-effects parameters $\beta^{(n)}$ are block-updated, conditional on the other parameters and latent variables using random walk Metropolis, with five blocks (β_{1j} and β_{2j} for $j=1,2$ and the cancer model α). New values are proposed with a multivariate Normal distribution with standard deviations ranging between 0.005 and 0.05, chosen by trial and error to give acceptance rates in the 0.5 to 0.8 range.

At each iteration, 10 updates of this step are performed, as the computational time for this step is relatively undemanding.

Examiner-level random effects $[\theta^{(m)}_{sje}, \eta^{(m)}_{sje}]$ undergo random walk Metropolis updating, again repeating 10 times. Proposal standard deviations are 0.8 for the true positive variables and 0.2 for the false positives. Screening site random effects $(\kappa^{(n)}_{sj}, \lambda^{(n)}_{sj})$ and fixed-effects parameters δ are sampled directly from Gaussian conditional distribution given the examiner-level random effects and variance matrices. Variance parameters $\sigma^{(n)}$, $\Sigma^{(n)}_{\cdot}$, and $\Gamma^{(n)}_{\cdot}$ are directly sampled from Wishart conditional distributions given the random effects. The distribution of ρ is Beta-distributed, conditional on the other variables in the model, and it is sampled directly.

Acknowledgements

We would like to thank Anna Chiarelli for the provision of the data. The research project was supported by grants from the Canadian Breast Cancer Foundation Ontario Region and the Natural Sciences and Engineering Research Council. The authors also thank the Ontario Breast Screening Program, a program of Cancer Care Ontario, for use of its data for this study. The Associate Editor and two anonymous reviewers deserve particular recognition for their attention and suggestions for improvements to this paper.

References

1. Chiarelli A, Mai V, Moravan V, Halapy E, Majpruz V, Tatla R. False-positive result and reattendance in the Ontario Breast Screening Program. *Journal of Medical Screening* 2003; **10**(3):129–133.
2. Chiarelli A, Majpruz V, Brown P, Thériault M, Shumak R, Mai V. The contribution of clinical breast examination to the accuracy of breast screening. *Journal of the National Cancer Institute* 2009; **101**(18):1236–1243.
3. Chiarelli A, Majpruz V, Brown P, Theriault M, Edwards S, Shumak R, Mai V. Influence of nurses on compliance with breast screening recommendations in an organized breast screening program. *Cancer Epidemiology Biomarkers and Prevention* 2010; **19**(3):697–706.
4. McPherson K, Steel C, Dixon J. Breast cancer-epidemiology, risk factors, and genetics. *British Medical Journal* 2000; **321**(7261):624–628.
5. Beam C, Layde P, Sullivan D. Variability in the interpretation of screening mammograms by us radiologists: findings from a national sample. *Archives of Internal Medicine* 1996; **156**(2):209–213.
6. Diggle P, Heagerty P, Liang KY, Zeger S. *Analysis of Longitudinal Data*, 2nd edn. Oxford University Press: New York and Oxford, 2013.
7. Woodard D, Gelfand A, Barlow W, Elmore J. Performance assessment for radiologists interpreting screening mammography. *Statistics in Medicine* 2007; **26**(7):1532–1551.
8. Carroll R. *Measurement Error in Nonlinear Models: A Modern Perspective*, Vol. 105. CRC Press: Boca Raton (FL), 2006.
9. Neuhaus J. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* 1999; **86**(4):843–855.
10. Rosychuk R, Thompson M. A semi-Markov model for binary longitudinal responses subject to misclassification. *Canadian Journal of Statistics* 2001; **29**(3):395–404.
11. Roy S, Banerjee T, Maiti T. Measurement error model for misclassified binary responses. *Statistics in Medicine* 2005; **24**(2):269–283.
12. Rosychuk R, Islam S. Parameter estimation in a model for misclassified Markov data—a Bayesian approach. *Computational Statistics and Data Analysis* 2009; **53**(11):3805–3816.
13. Kuchenhoff H, Mwalili S, Lesaffre E. A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics* 2006; **62**(1):85–96.
14. McGlothlin A, Stamey J, Seaman Jr J. Binary regression with misclassified response and covariate subject to measurement error: a Bayesian approach. *Biometrical Journal* 2008; **50**(1):123–134.
15. Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Statistics in Medicine* 1986; **5**(1): 21–27.
16. Hui S, Walter S. Estimating the error rates of diagnostic tests. *Biometrics* 1980; **36**(1):167–171.
17. Walter S, Irwig L. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology* 1988; **41**(9):923–937.
18. Espeland M. Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics* 1989; **45**(2):587–599.
19. Uebersax J, Grove W. Latent class analysis of diagnostic agreement. *Statistics in Medicine* 1990; **9**(5):559–572.
20. Qu Y, Tan M, Kutner M. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* 1996; **52**(3):797–810.
21. Joseph L, Gyorkos T, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* 1995; **141**(3):263–272.
22. Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine* 2002; **21**(18):2653–2669.
23. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 2001; **57**(1):158–167.

24. Rutter C, Gatsonis C. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001; **20**(19):2865–2884.
25. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology* 2004; **57**(9):925–932.
26. Arends L, Hamza T, Van Houwelingen J, Heijenbrok-Kal M, Hunink M, Stijnen T. Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making* 2008; **28**(5):621–632.
27. Puggioni G, Gelfand A, Elmore J. Joint modeling of sensitivity and specificity. *Statistics in Medicine* 2008; **27**(10): 1745–1761.
28. Smith R, Saslow D, Andrews Sawyer K, Burke W, Costanza M, Evans III W, Foster Jr R, Hendrick E, Eyre H, Sener S. American Cancer Society guidelines for breast cancer screening: update 2003. CA: *a Cancer Journal for Clinicians* 2003; **53**(3):141–152.
29. Barton M, Harris R, Fletcher S. Does this patient have breast cancer?: the screening clinical breast examination: should it be done? How? *The Journal of the American Medical Association* 1999; **282**(13):1270–1280.
30. Boyd N, Guo H, Martin L, et al. Mammographic density and risk of breast cancer. *New England Journal of Medicine* 2007; **356**:227–236.
31. Madigan M, Ziegler R, Benichou J, Byrne C, Hoover R. Proportion of breast cancer cases in the United States explained by well-established risk factors. *Journal of the National Cancer Institute* 1995; **87**(22):1681–1705.
32. Chib S, Greenberg E. Understanding the Metropolis–Hastings algorithm. *The American Statistician* 1995; **49**(4):327–335.
33. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2014.
34. Brown P, Zhou L. MCMC for generalized linear mixed models with glmmBUGS. *R Journal* 2010; **2**:13–17.
35. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 1998; **7**(4):434–455.
36. Plummer M, Best N, Cowles K, Vines K. Coda: convergence diagnosis and output analysis for MCMC. *R News* 2006; **6**(1):7–11.
37. Menten J, Boelaert M, Lesaffre E. Bayesian latent class models with conditionally dependent diagnostic tests: a case study. *Statistics in Medicine* 2008; **27**(22):4469–4488.
38. Walter S, Macaskill P, Lord S, Irwig L. Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. *Statistics in Medicine* 2012; **31**(11-12):1129–1138.
39. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. *New England Journal of Medicine* 2012; **367**(21):1998–2005.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.