10-3-2007

MULTIPLE MODEL EVALUATION ABSENT THE GOLD STANDARD VIA MODEL COMBINATION

Edwin J. Iversen, Jr.
Department of Statistics, Duke University, iversen@nonsense.isds.duke.edu

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Sining Chen
Department of Environmental Health Science, Johns Hopkins Bloomberg School of Public Health

Multiple Model Evaluation Absent the Gold Standard via Model Combination

Edwin S. Iversen, Jr., Giovanni Parmigiani and Sining Chen*

Abstract

We describe a method for evaluating an ensemble of predictive models given a sample of observations comprising the model predictions and the outcome event measured with error. Our formulation allows us to simultaneously estimate measurement error parameters, true outcome — aka the gold standard — and a relative weighting of the predictive scores. We describe conditions necessary to estimate the gold standard and for these estimates to be calibrated and detail how our approach is related to, but distinct from, standard model combination techniques. We apply our approach to

*This work was funded in part by the Cancer Genetics Network at Duke, U24 CA78157, and at The Johns Hopkins University, U24 CA78148. In addition, G.P. and S.C. were supported by the Johns Hopkins SPORE in Breast Cancer (P50CA88843) and NCI grants R01CA105090-01A1 and P50CA62924-05; E.I. also received support from NCI through R01CA105090-01A1. The authors wish to thank this manuscript's editors and referees for their comments and suggestions and the following individuals for their invaluable contributions to the success of the CGN BRCA1/2 Models Validation Study: Tara Friebel, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania; Dianne Finkelstein, The Massachusetts General Hospital; Hoda Anton-Culver, Department of Medicine, University of California, Irvine; Argyrios Ziogas, Department of Medicine, University of California, Irvine; Barbara L. Weber, Abramson Cancer Center, University of Pennsylvania; Andrea Eisen, Hamilton Regional Cancer Centre; Kathleen E. Malone, Program in Epidemiology, Fred Hutchinson Cancer Research Center; Li Hsu, Public Health Sciences Division, Fred Hutchinson Cancer Research Center; Leif E. Peterson, Departments of Medicine and Molecular and Human Genetics Baylor College of Medicine; Joellen M. Schildkraut, Department of Community and Family Medicine, Duke University; Claudine Isaacs, Georgetown University Lombardi Cancer Center; Beth N. Peshkin, Georgetown University Lombardi Cancer Center; Camille Corio, Georgetown University Lombardi Cancer Center; Leoni Leondaridis, Georgetown University Lombardi Cancer Center; Gail Tomlinson, Department of Pediatrics, University of Texas Southwestern; Christopher I. Amos, Department of Epidemiology, University of Texas M. D. Anderson Cancer Center; Louise C. Strong, Department of Pediatrics, University of Texas M. D. Anderson Cancer Center; Donald A. Berry, Department of Biostatistics, University of Texas M. D. Anderson Cancer Center; Jeffrey Weitzel, City of Hope National Medical Center; Sharon Sand, City of Hope National Medical Center; Debra Dutson, Huntsman Cancer Center, University of Utah; Rich Kerber, Huntsman Cancer Center, University of Utah; David M. Euhus, Department of Surgery, University of Texas Southwestern.

data from a study to evaluate a collection of BRCA1/BRCA2 gene mutation prediction scores. In this example, genotype is measured with error by one or more genetic assays. We estimate true genotype for each individual in the dataset, operating characteristics of the commonly used genotyping procedures and a relative weighting of the scores. Finally, we compare the scores against the gold standard genotype and find that Mendelian scores are, on average, the more refined and better calibrated of those considered and that the comparison is sensitive to measurement error in the gold standard.

Keywords: Model Evaluation; Model Combination; Measurement Error; Breast Cancer Susceptibility Genes; Bayesian Analysis.

1 Introduction

Numerous medical conditions can be diagnosed only to a level of uncertainty via one or more tests, be they diagnostic, biomarker–based or probabilistic. In addition to the obvious clinical implications, this complicates evaluation of competing diagnostic tools because the true condition of patients in the evaluation sample, sometimes referred to as the 'gold standard' for comparison, is not known (Hui and Walter 1980). In what follows, we describe and apply a method for evaluating an ensemble of prognostic models given a sample on which those models have been evaluated and for which the outcome event is measured with error.

This work is inspired by a multi–center validation study of BRCA1/2 mutation carrier probability models carried out within the Cancer Genetics Network (CGN). These models are routinely used in high risk cancer clinics as a counseling tool prior to formal genetic testing, but their predictions may be inconsistent. For each study participant, the CGN validation data set contains genetic test results at BRCA1 and BRCA2, BRCA1/2 carrier scores, denoted $S_1, \cdots, S_M$, calculated for $M = 9$ models and a small number of family history summaries. Section 3 provides a complete listing of fields collected in context of the

CGN validation study. Each carrier probability score S_m is a function of the individual's family history and, with a few exceptions, each is an estimate of $\Pr(G \mid \mathbf{F})$ where $\mathbf{F}$ denotes the individual's family history of breast and ovarian cancer and G is an indicator of whether or not the individual carries a disease associated mutation at BRCA1 or BRCA2. A subset of these scores was built by modeling test result T as a function of $\mathbf{F}$. In principle, these scores could be compared to those built explicitly to predict G by multiplying the latter class by test sensitivity thereby putting them on the same — but clinically less relevant — footing, given test specificity is 1. That the test used to train the former class of scores is usually not the same used in clinic further complicates this approach.

It is often the case that measurement error in the gold standard is ignored in evaluations of competing prognostic models. Indeed, in an accompanying report (Parmigiani et al. 2007) we compare the various BRCA1/2 carrier scores using assay result as the gold standard. This is also the approach taken in Barcenas et al. 2006. Here we demonstrate that such a comparison can be sensitive to this practice and provide the foundation for a less biased evaluation.

2 Modeling Approach

2.1 The Sampling Model

Most CGN study subjects were from centers that collected, or ascertained, them on the basis of the extent of disease present in their family histories. We account for this in our model by conditioning on family history, $\mathbf{F}$, through the carrier probability scores, each of which is a function of $\mathbf{F}$ (in fact, it is not uncommon for recruitment of individuals to studies of women at high genetic risk of breast cancer to explicitly depend on one or more of the scores under evaluation). In particular, we use a retrospective model (Kraft and Thomas 2000) for genetic test results at BRCA1 and BRCA2, denoted T_1 and T_2 respectively, given carrier

probability scores $S_1, \cdots S_M$, covariates $\mathbf{X}$ and model parameters θ. The parameter vector θ is comprised of test sensitivity and mutation prevalence parameters and parameters relating carrier probabilities to genotype. The model is also conditional on P_1 and P_2, the BRCA1 and BRCA2 test protocols employed.

The conditional sampling model we utilize relates carrier probability model scores to test results via latent variables for genotype. In particular, we write

$$\Pr(T_1, T_2 \,|\, P_1, P_2, S_1 \cdots S_M, \mathbf{X}, \theta) = \sum_{\substack{(G_1,G_2)\in \\ \{(0,0),(0,1),(1,0)\}}} \Pr(T_1, T_2 \,|\, G_1, G_2, P_1, P_2, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}) \Pr(G_1 \,|\, G, \mathbf{X}, \boldsymbol{\gamma_1}) \Pr(G \,|\, S_1 \cdots S_M, \mathbf{X}, \boldsymbol{\gamma_2}). \quad (1)$$

We code test result $T_g = 1$ if the individual tests positive for a disease associated mutation at BRCAg, $g \in \{1, 2\}$; $T_g = 0$, otherwise. Similarly, we code genotype $G_g = 1$ if the individual truly carries a disease associated mutation at BRCAg; $G_g = 0$, otherwise. We define $G = G_1 \vee G_2$. The operating characteristics of the genetic tests are imperfect. While it is well established that they are specific (Iversen Jr. et al. 1999; Myriad Genetics 2003), their sensitivities vary and may be significantly less than one. We introduce gene– and test–modality–specific sensitivity parameters, $\boldsymbol{\beta_g} = (\beta_{g,1} \cdots \beta_{g,N_g})$, for test protocols $P_g \in \{1 \cdots N_g\}$. We ignore the extremely rare possibility that a tested individual may harbor disease-associated mutations at both genes ($(G_1, G_2) = (1, 1)$). Further assumptions made in Equation 1 are given below.

Test Results Given Bivariate Genotype. We assume that test results are independent of $\mathbf{X}$ and the scores given genotype and that the two tests are conditionally independent given genotype and test protocols, yielding

$$\Pr(T_1, T_2 \mid G_1, G_2, P_1, P_2, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}) = \Pr(T_1 \mid G_1, P_1, \boldsymbol{\beta_1}) \Pr(T_2 \mid G_2, P_2, \boldsymbol{\beta_2}), \quad (2)$$

Each factor $\Pr(T_g \mid G_g, P_g, \boldsymbol{\beta_g})$ in this expression is Bernoulli with modality–specific success probability β_{g,P_g} given $G_g = 1$ and a point mass at $\{T_g = 0\}$ otherwise. This expression is

easily adapted to tests with specificity less than 1.

Bivariate Given Univariate Genotype. The second factor in Equation 1 relates joint BRCA1, BRCA2 genotype (G_1, G_2) to combined genotype, G, and covariates. Here, we assume that the likelihood of BRCA1 genotype G_1 is Bernoulli and conditionally independent of the carrier probability scores given $G = 1$ and $\mathbf{X}$, where $\mathbf{X}$ includes variables that modify prevalence of BRCA1 mutations or distinguish the BRCA1/2 familial phenotypes. These are related to G_1 through a logistic regression on the subset of the sample with $G = 1$. When $G = 0$, this component places a point mass at $(G_1, G_2) = (0, 0)$.

Univariate Genotype Given Scores. While some of the scores make joint BRCA1 and BRCA2 predictions, the CGN study only collected combined predictions. Thus we relate the carrier probability scores and covariates in $\mathbf{X}$ to combined genotype, G, through a regression model of the form $\Pr(G \mid S_1, \cdots, S_M, \mathbf{X}, \boldsymbol{\gamma_2})$ where $\boldsymbol{\gamma_2}$ is the vector of regression coefficients. This component is central to the analysis: it relates the carrier scores to the gold standard variable G. One way of thinking about it is as another score of the form $\Pr(G \mid \mathbf{F})$, one that assumes that G is conditionally independent of $\mathbf{F}$ given the various scores $S_1(\mathbf{F}), \cdots, S_M(\mathbf{F})$. It serves as, and could function in the role of, a composite carrier probability score.

We consider two formulations for this regression model: a multiplicative model

$$\text{logit}(\pi_j) \;=\; \gamma_{2,0} \;+ \sum_{m=1}^{M} \gamma_{2,m}\, f(S_{j,m}), \tag{3}$$

where $f(\cdot)$ defines a transformation of the scores; and an additive model

$$\pi_j \;=\; \gamma_{2,0} + \sum_{m=1}^{M} \gamma_{2,m}\, S_{j,m} \quad \text{subject to } \sum_{m=0}^{M} \gamma_{2,m} \;=\; 1 \;\text{ and }\; \gamma_{2,m} > 0 \;\forall\; m. \tag{4}$$

Identifiability. In both formulations, identifiability of γ_2**,** β_1 and β_2 is guaranteed given all but exceptional arrangements of the data and a collection of non–trivial scores. Gold

standard genotypes G are estimable when γ_2, β_1 and β_2 are identified. To see this, note that the likelihood in Equation 1 marginalized over G_1 and G_2 is a product of individual–specific multinomial terms with success probabilities $(\beta_1\rho_j\pi_j,\ \beta_2(1-\rho_j)\pi_j,\ 1-\beta_1\rho_j\pi_j-\beta_2(1-\rho_j)\pi_j)$ when there is one test modality employed and where ρ_j denotes $\Pr(G_{1,j} \mid G_j = 1, \mathbf{X_j}, \gamma_\mathbf{1})$ and j indexes the individual. Note that β_1 or β_2 always multiplies π_j. Hence multiplying β_1 and β_2 and dividing π_j by the same constant $0 < c < 1$ yields the same value of the likelihood. In the additive formulation, this corresponds to dividing $\gamma_\mathbf{2}$ by that constant. However, the constraint that it sums to 1 prevents this and thus its parameters are identifiable. In the multiplicative formulation, identifiability is guaranteed if the design matrix $\mathbf{S}$ is of full rank and no components of $\gamma_\mathbf{2}^*$ are uniquely determined by the system of equations $\mathbf{S}^{\mathrm{T}}{\gamma_\mathbf{2}}^* = \mathrm{logit}(c \cdot \mathrm{expit}(\mathbf{S}^{\mathrm{T}}\gamma_\mathbf{2}))$ for $0 < c < 1$ and arbitrary $\gamma_\mathbf{2}$. The same applies in the case where there are multiple assays.

Calibration. The multiplicative model is a prospective model for G given the S's. It corresponds to a joint formulation $\Pr(G, S_1, \cdots, S_M \mid \text{parameters})$ in which the S_m's are conditionally independent beta distributed random variables given G and in which G is Bernoulli(π). In the prospective version, elements of $\gamma_\mathbf{2}$ are simple functions of π and the beta distribution parameters. While interpretation of the joint model's parameters is complicated in cases where the implied conditional independence assumption is not tenable, as is the case in the analysis described below, inference for the latent genotypes is not necessarily compromised. For our purposes, all that is necessary is for the prospective model to be calibrated (a score S is calibrated for G if $\Pr(G = 1 \mid S = s) = s$). This would be achieved if, for example, any one score entering the retrospective model, say S_m, is calibrated for G and $\gamma_{2,m} = 1$ and $\gamma_{2,m'} = 0$ for all $m' \neq m$. In contrast, the additive model has a corresponding joint formulation under certain restrictions on the coefficients that are outlined in Genest and Schervish 1985. It is calibrated if, among other scenarios, the intercept and coefficients associated with uncalibrated scores are zero. In our analysis we set the intercept to zero.

An alternate approach to Equation 1 is to model the joint distribution of test results and scores, i.e. $Pr(S_1, \cdots, S_M, T_1, T_2 \mid \theta)$. This model is related to that described by Hui and Walter 1980 who write a joint likelihood for an arbitrary number of binary tests with unknown error rates for a sample that can be divided into subpopulations with unknown, but different, disease prevalences (see also Walter and Irwig 1988). One shortcoming of taking this approach in the current context is that it ignores the differential modes of ascertainment employed across and sometimes within the various centers and could result in biased inferences. The retrospective likelihood is 'ascertainment assumption free' to the extent that ascertainment of subjects in the data set depends on family history through a function of the scores. Indeed, in many settings this is exactly the case.

2.2 Relation to Model Combination

The purpose of the models described in Equations 3 and 4 is to generate a "third party" consensus for the latent genotypes that does not *a priori* favor any score or subset of scores; a byproduct is to identify a composite carrier probability model. We view this modeling as a device for combining expert (statistical) opinion across M experts *vis-a-vis* the unknown genotype of test–negative probands. When we choose the logit transformation for f in Equation 3 we define a logarithmic opinion pool (Genest and Zidek 1986) of the form described by Bordley 1982. This approach represents a quasi–Bayesian combination of evidence – 'quasi' in the sense that prior odds, specified as a function of $\gamma_{2,0}$, are a mathematical construct not elicited prior to assembling the expert evidence (French 1985; Genest and Zidek 1986; Dawid et al. 1995). The model of Equation 4 defines a linear opinion pool.

Our procedure departs from the typical problem of combining probabilities in two important respects. First, typically there is a single defined event (or event space) of interest. In the current setting, we have a series of events, each corresponding to the genotype of one of the probands. A fixed set of M automated experts proffer their opinions, in the form

of scores, on each individual's genotype. The parameters $\gamma_{2,1} \cdots \gamma_{2,M}$ define the relative weights accorded the scores. These weights are taken to be the same from event to event (*i.e.* across probands). Second, we observe the events of interest, albeit measured with error. This allows us to estimate the weights.

This approach, a regression based combination of model forecasts, is one that has been used to improve meteorological (Kharin and Zwiers 2002; Gel et al. 2004; Gneiting et al. 2005) and econometric (Makridakis and Winkler 1983) predictions. In the meteorological literature, forecasts derived from collections of predictive models, termed multi-model ensembles, are combined using either deterministic weights (Thompson 1977) or empirically derived weights (Fraedrich and Smith 1989; Krishnamurti et al. 1999; Kharin and Zwiers 2002), for example from a regression of forecasts given a set of observations of the event. This procedure corresponds to a linear opinion pool with empirically determined weights.

While the models being combined may be similar, e.g. reflecting like–minded analyses of nonidentical data sets in the sense of (Dawid et al. 1995), there is evidence that there is strength in combining predictions based on very dissimilar models (Makridakis and Winkler 1983; French 1995) that derives from accounting for model uncertainty. There is a growing literature on formal Bayesian methods for incorporating model uncertainty through model averaging (Clyde and George 2004). Bayesian model averaging combines models in weighted averages where the weights are determined by the marginal likelihood of the data associated with each model. This approach arises in the so called M–closed setting, where the true model is assumed to be included in the average (Key et al. 1999). In the M–open setting, where this is not the case, model forecasts may be optimally combined via a weighted multivariate regression of observed forecasted events on out–of–sample forecasts, resulting in an empirically weighted linear opinion pool.

3 CGN Analysis

3.1 CGN Data Set

The Cancer Genetics Network's (CGN) Validation of BRCA1&2 Carrier Probability Models study assembled a multi–center database of carrier probability scores, family history summaries and other data on BRCA1/2 tested individuals for use in evaluating performance of those models. Data were provided by eleven centers. Center by center summaries are provided in Table 1. With the exception of three, all provided data on high risk families recruited in counseling clinics. The cases from Seattle were selected from a population–based series of breast cancer cases diagnosed before age 45 either on the basis of early age at diagnosis (< 35) or incidence of breast cancer in a first–degree family member. The cases from Baylor comprise a cohort from a community–based study who underwent testing for the 185delAG BRCA1 Ashkenazi Jewish founder mutation and provided a complete family history of cancer. Cases from UC Irvine are the first 803 participants in a study of all breast cancer cases diagnosed in Orange County, California during the period 3/1/94 to 2/28/95. Centers carried out the various model calculations themselves and assembled their datasets using CaGene (Euhus et al. 2002), a software package used in genetic counseling; in most cases, we did not have access to the raw family data.

A variety of genetic testing strategies were employed by the centers participating in this study. At some centers, multiple assays were implemented serially to improve sensitivity of the overall testing procedure. The various assays and their combinations are specific and their false positive rates can be assumed to be negligible, but their sensitivities vary. Table 1 summarizes the assays employed by the participating centers.

The project evaluated nine BRCA1/2 carrier probability models. Four are empirical models built by modeling test data as a function of various family history summaries. They

Center	*Population*	*Number Tested for BRCA1&2*	*Number Tested for BRCA1 Only*	*Number Tested for BRCA2 Only*	*BRCA1 Testing Method*	*BRCA2 Testing Method*
Baylor	AJ[a] volunteers	0	282	0	AJ Sequencing	NA[d]
Duke	High risk	275	0	0	SSCP, CSGE	SSCP, CSGE
Georgetown	High risk	230	11	1	Seq[c], 0ncormed, AJ Seq	ASO
Johns Hopkins	High risk	102	3	0	Seq, AJ Sequencing	Seq, AJ Seq
Penn	High risk	472	159	45	CSGE, Seq, AJ Seq, SSCP	Seq, CSGE, AJ Seq
Seattle	High risk[b]	384	199	0	SSCP	SSCP
MD Anderson	High risk	115	2	0	Sequencing	Sequencing
UTSW	High risk	115	6	0	Sequencing	Sequencing
UC Irvine	Popn–Based	803	0	0	ASO, AJ Seq	AJ ASO, Unknown
Utah	High risk	61	0	0	Sequencing	Sequencing
City of Hope	High risk	76	1	0	Sequencing, AJ Seq	Seq, AJ Seq

Table 1: Centers participating in the CGN validation study, their source population, sample contribution and testing methods employed. The high risk population refers to individuals presenting to clinics or involved in research studies for reason of a family history of breast and ovarian cancer. Population based samples are tied via a known sampling scheme to a population cancer registry. Volunteers to the Baylor study responded to an advertisement. Assays used in testing for disease associated variants are listed by gene in order of their frequency of use.

[a]Ashkenazi Jewish
[d]Not Applicable
[c]Sequencing
[b]High risk subsample of population–based series

are the Myriad (`www.myriad.com`) model (Frank et al. 1998; Frank et al. 2002) for predicting results of testing at BRCA1 and BRCA2; the NCI model (Hartge et al. 1999) for testing positive for one of the three Ashkenazi Jewish BRCA1/2 founder mutations; the University of Pennsylvania model (Couch et al. 1997) for predicting whether one or more individuals in a family will test positive for a mutation at BRCA1, modified to predict for the individual in the family seeking testing (the "counseland"); and the Finnish model (Vahteristo et al. 2001) for predicting results of a test for a BRCA1 or BRCA2. The three variants of BRCAPRO (Berry et al. 1997; Parmigiani et al. 1998) and the Yale model (Claus et al. 1990; Claus et al. 1991) are Mendelian models, i.e. models for family history data that formally account for the underlying Mendelian genetic model using a formal likelihood based on that described in Elston and Stewart 1971. BRCAPRO predicts the joint mutation status at BRCA1 and BRCA2, while the Yale model predicts that of a single autosomal dominant disease gene predisposing to breast and ovarian cancer (effectively combining BRCA1 and BRCA2).

The empirical and Mendelian models predict different quantities. The former predict the likelihood of testing positive to a genetic test with sensitivity and specificity equal to that

used the model building data set while the latter predict an individual's actual genotype. The remaining model, Family History Assessment Tool (FHAT) (Gilpin et al. 2000), is a quantitative 'expert–based' score developed to assist clinicians in identifying appropriate candidates for counseling. It is derived from responses to a family history questionnaire and is measured on a 45 point scale; we rescaled it to range between 0 and 1. More detail on the CGN study can be found in Parmigiani et al. 2007.

3.2 Model Specification

The model in Equation 1 is factored into three components. In what follows, we discuss specification of each, including our choice of priors.

Test Results Given Bivariate Genotype. Equation 2 specifies the structure of this component whose parameters are the sensitivities of the genetic testing procedures listed in Table 1. We derived informative priors by conducting a meta–analysis of published data (Geisler et al. 2001; Eng et al. 2001; Andrulis et al. 2002) in which we treated these studies as independent trials and pooled their results. This analysis suggested a beta(73.5, 43.5) prior on sensitivity of SSCP and SSCP followed by ASO and a beta(27.5, 18.5) on sensitivity of CSGE. The beta(3.27,0.39) prior on sensitivity for testing for BRCA1 185delAG but *not* BRCA1 5382insC was based on the relative proportion of these mutations in the Ashkenazi population as observed in Roa et al. 1996; Hartge et al. 1999. The remaining testing modalities were accorded uniform priors on their sensitivities. In addition, we applied to these priors a set of ordering constraints. For BRCA1, these reflected the beliefs that full gene sequencing is more sensitive than any other assay, including selected search for an AJ founder mutation by sequencing; that both sequencing strategies are more sensitive than SSCP or CSGE; that sensitivity of sequencing or ASO for one AJ mutation is less sensitive than sequencing or ASO for both; and that any assay is less sensitive alone than when used in combination with any other. Similar constraints were applied to the assays employed for

BRCA2. Individuals tested for mutations at only one of the two genes were treated as if they had been tested using an assay with sensitivity zero for the untested gene.

We believe these studies to be of high quality and that incorporating data from them improves the model. We explore the robustness of our analysis to these prior data in Section 3.6, where we report on sensitivity of the latent genotypes, and of the comparison of scores given those genotypes, to this prior.

Bivariate Given Univariate Genotype. The event that a mutation carrier's mutation is located at BRCA1 is modeled by a logistic regression on $\mathbf{X}$. In the current analysis, $\mathbf{X}$ includes only ethnicity. If the data had been available, we would also have included summaries of family history that help to distinguish the BRCA1 and BRCA2 phenotypes. However, without data on tumor markers, whose significance is only now becoming clear, BRCA1 and BRCA2 carriers are difficult to distinguish from one another even given family history data, and these data probably would not contribute much. The parameters most sensitive to misspecification of this component will be the sensitivity parameters as misspecification will affect the model's ability to discriminate whether likely false negatives are truly BRCA1 or BRCA2 carriers. These are ancillary inferences, however; our interest is in constructing an accurate estimate of the marginal genotype G — the relevant quantity for clinical management and genetic testing decisions — and in understanding its relationship to the scores. We placed diffuse independent 10 degree of freedom, mean zero Student–t priors on the coefficients; the prior standard deviation was 100 for the intercept and 5 for ethnicity.

Univariate Genotype Given Scores. This component is either a multiplicative (Equation 3) or additive (Equation 4) regression of latent genotype on the carrier probability scores. We included dummy variables to indicate the subsets of observations to which the NCI, Penn and Finnish scores apply to allow for the fact that not all models are universally applicable. We include a term for each carrier score under evaluation. Three variants of the

BRCAPRO model are included: its standard formulation, one calculated assuming 50% of the standard formulation's penetrance and one calculated assuming penetrances generated by applying age–independent relative risks (AIRR) to the phenocopy rates of breast and ovarian cancer. This is to allow for sensitivity of the Mendelian scores to assumed penetrance. In addition, we include indicator variables for a family history of prostate cancer, colon cancer and endometrial cancer as tools for discovering new factors that might be used to improve existing carrier probability models. Finally, we adjust for the counseland's phenotype — the counseland's age, whether she had breast cancer, bilateral breast cancer, or ovarian cancer — and mode of data ascertainment (population–based or high–risk). Because coefficients in the additive formulation are restricted to be positive, we include in the additive model indicators and scores for the complementary events. The model is constrained so that a variable and its complement will never co–appear. Terms appearing in the genotype model are listed in Table 2.

3.3 Out–of–Sample Evaluation of Genotype Model

In this and the next two subsections we focus on model choice and evaluation. Here we evaluate the multiplicative (Equation 3) and additive (Equation 4) formulations. In the next subsection we focus on specification of the genotype model's linear predictor and estimation of the gold standard. Finally, we utilize these estimates to compare the carrier scores.

Integrity of the evaluation and comparison of the scores depends on accuracy of the imputed genotypes. In light of this, we evaluated out–of–sample predictive accuracy of the additive and multiplicative formulations using the subset (n=640) of probands with family history data who received full gene sequencing. Sensitivity of sequencing is high, on the order of 0.9; specificity is effectively 1. For the purpose of this analysis, we ignored assay error and equated test result with binary genotype, G. We randomly divided the dataset into 100 subsets. For each subset, we fit both models to the remaining subsets and, under

each model, calculated predictions for the left–out subset.

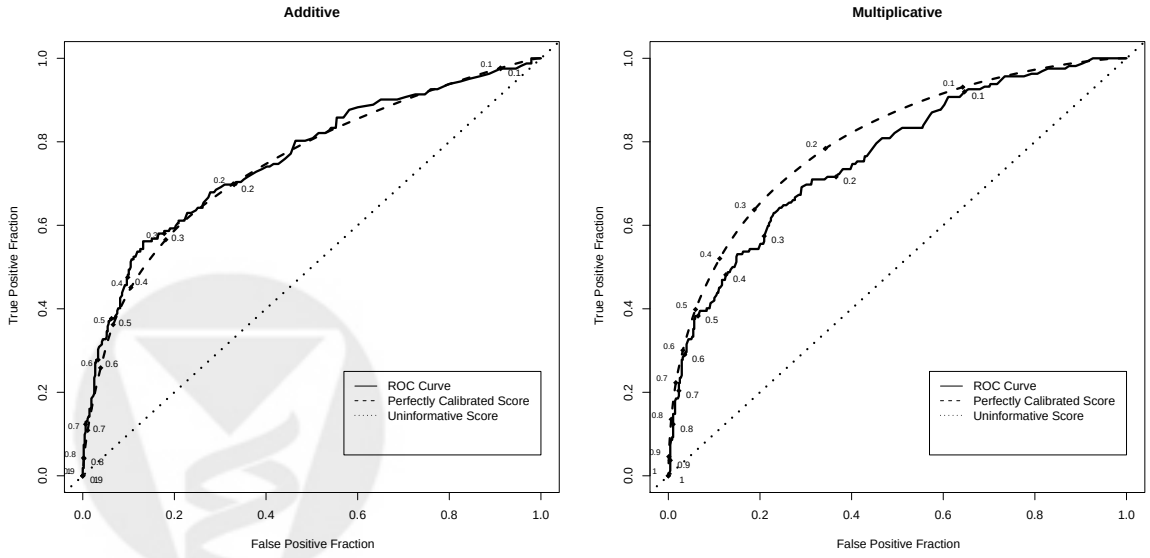
We judged the models on the basis of the overall accuracy, calibration and refinement of their out–of–sample predictions. We found the additive formulation to be superior on the basis of these criteria and in terms of interpretability. Using root mean squared error (RMSE) as the measure of overall accuracy, we found the additive model to be more accurate (RMSE=0.387) than the multiplicative model (RMSE=0.391). In both cases, we found the bias to be negligible (-0.0031 for the additive predictions; -0.0005 for the multiplicative predictions). Model calibration, a measure of reliability in repeated forecasts, quantifies agreement between the forecast value and the frequency of the forecasted event as a function of the forecast value (DeGroot and Fienberg 1983). We addressed calibration by comparing the accuracy of the predictions within deciles of the predicted values. One–way analyses of variance of the deviations given the factor 'predictive decile' suggest that the additive predictions are better calibrated on the whole and suffer less from systematic deviations from calibration: the model mean square for the additive predictions was 0.115 and the residual mean square was 0.151; the associated quantities for the multiplicative model were 0.213 and 0.152. In these analyses, decile represents an *ad hoc* correction to calibration that explains nearly twice as much deviation from calibration under the multiplicative than under the additive model.

Model refinement is related to forecaster accuracy. In the case of a binary event, both the constant forecast equal to the true prevalence of the event and the binary forecast that is always correct are well calibrated; the latter is refined, the former is not (DeGroot and Fienberg 1983). We report two measures of refinement: the receiver operating characteristic (ROC) curve and the area under the curve (AUC), also termed the concordance index (Harrell Jr et al. 1982). Figure 1 depicts ROC curves for the additive (left panel) and multiplicative (right panel) predictions. To form a basis of comparison, each plot also includes curves for two hypothetical scores, one of which is perfectly calibrated and has the

same marginal distribution as the predictions (heavy dashed line) while the other is completely uninformative (dashed diagonal). Note that the additive predictions substantially overlap the associated hypothetical calibrated score while this is not true of the multiplicative predictions. The area under the ROC curve is 0.757 for the additive and 0.762 for the multiplicative predictions.

To calculate the perfectly calibrated reference curve, note that a score S is perfectly calibrated for G if $s = \Pr(G = 1 \,|\, S = s)$. Let $\bar{S}$ denote the sample mean of S and note that $\Pr(S \,|\, G = 1) \simeq S/(n\bar{S})$ for S in $\{S_i : i = 1, \cdots, n\}$ when $\Pr(S)$ is estimated by its empirical distribution in the sample of interest; similarly, $\Pr(S \,|\, G = 0) \simeq (1 - S)/(n(1 - \bar{S}))$. The associated ROC curve plots $(\sum_{\{i:S_i>s\,\&\,G_i=0\}} \Pr(S_i \,|\, G_i), \sum_{\{i:S_i>s\,\&\,G_i=1\}} \Pr(S_i \,|\, G_i))$ as s varies between 0 and 1.

Figure 1: ROC plots for the out–of–sample evaluation of the additive (left panel) and multiplicative (right panel) models. The ROC curve plots the positive predictive value (PPV) against the negative predictive value (NPV) of a continuous score as a function of a threshold that varies through the range of scores (here 0 to 1). Each plot depicts 3 ROC curves: the curve associated with the predictions (solid line), the curve associated with a hypothetical perfectly calibrated curve with the same marginal (heavy dashed line) and a hypothetical uninformative score (dash line on the diagonal).



In the course of exploratory analysis using the multiplicative and additive formulations we examined various additional explanatory variables, including transformations of the scores

and a variety of summaries of extent of family history, for their ability to improve the accuracy of the model for genotype given scores. None were found to add to the explanatory ability of the models as originally formulated. We conclude that the additive model is more accurate and better calibrated than the multiplicative model. For these reasons and the fact that it is easier to interpret, we focus on the additive model in its original formulation. An evaluation of scores in the subset of sequenced individuals using assay result as the standard of comparison can be found in Parmigiani et al. 2007.

3.4 Model Fit & Estimation of Gold Standard

We fit the additive variant (Equation 4) of the model described by Equation 1 to the full CGN dataset using a Markov chain Monte Carlo (MCMC) algorithm. We placed a Dirichlet$(1, \cdots, 1)$ prior on $\boldsymbol{\gamma_2}$. Latent genotypes G_{1i} and G_{2i} were drawn from individual–specific multinomial distributions; β_1 and β_2 were drawn from their full conditional distributions and γ_1 was updated using an independence sampler (Tierney 1994) that achieved an acceptance rate of 93%. Finally, we updated γ_2 and the specification of the linear predictor in Equation 4 using a reversible jump step. Three move types were employed: (1) update γ_2 associated with the current model, (2) add an additional variable and (3) remove a variable. Moves were constrained so that a variable would never enter the model with its complement. We tested the code by removing the likelihood from the acceptance calculation and observed that the MCMC estimate of the distribution over the number of regressors coincided with its theoretical counterpart.

The chain was run for a lengthy burn–in period and one million subsequent iterations. The chain was thinned to every 5th realization, leaving two hundred thousand draws for inference. We employed a battery of convergence diagnostics implemented in the R package CODA to evaluate the convergence of non–dimension–varying parameters. All passed the Raftery and Lewis diagnostic (Raftery and Lewis 1996), the Heidelberger and Welch sta-

Table 2: Estimates of the coefficients in the additive regression model associating latent binary BRCA1/2 genotype and the various carrier scores, a set of family history summaries, and subset and population indicators. Coefficients were constrained to be positive and sum to 1. Each variable was allowed to enter the model either as itself (x; columns 2-4) or as its complement ($1 - x$; columns 5-7). Age was scaled to range between 0 and 1. Model specification was considered a random variable and models were visited via a random walk. Estimates in the table are of posterior means (columns 2, 3, 5 and 6) and standard deviations (columns 4 and 7) derived from the MCMC analysis. The second column provides estimates of the probability that the associated variable is in the model. Columns 3 and 4 summarize the posterior on the coefficient averaging over models that include the variable.

	Variable "x"			Complement of Variable "1-x"		
Predictor	Pr(Included)	E($\gamma_{2,p}$|In)	SD($\gamma_{2,p}$|In)	Pr(Included)	E($\gamma_{2,p}$|In)	SD($\gamma_{2,p}$|In)
NCI Indicator	0.5425	0.0239	0.0131	0.0077	0.0030	0.0026
Penn Indicator	0.0374	0.0046	0.0038	0.0857	0.0111	0.0099
Finnish Indicator	0.0319	0.0044	0.0038	0.0858	0.0097	0.0085
High Risk Indicator	1.0000	0.1070	0.0158	0.0000	– – –	– – –
BRCAPRO Model	0.9947	0.2518	0.0768	0.0002	0.0052	0.0045
BRCAPRO, 50% Penetrance	0.9999	0.3520	0.0793	0.0000	– – –	– – –
BRCAPRO, AIRR	0.4394	0.0739	0.0578	0.0196	0.0042	0.0037
Yale Model	0.0850	0.0131	0.0123	0.0425	0.0050	0.0047
Myriad/Frank Model	0.4061	0.0614	0.0483	0.0198	0.0042	0.0035
NCI Model	0.1082	0.0164	0.0151	0.0345	0.0044	0.0040
Penn Model	0.4097	0.0644	0.0500	0.0186	0.0041	0.0040
FHAT Model	0.1559	0.0259	0.0242	0.0300	0.0049	0.0044
Finnish Model	0.9666	0.1691	0.0619	0.0012	0.0036	0.0034
Counseland First Breast Cancer	0.0337	0.0047	0.0043	0.0668	0.0074	0.0064
Counseland Breast Cancer Recur	0.0519	0.0081	0.0077	0.0528	0.0058	0.0050
Counseland Ovarian Cancer	0.1864	0.0212	0.0167	0.0298	0.0041	0.0040
Age, Scaled to [0,1]	0.0413	0.0060	0.0055	0.0756	0.0097	0.0088
(Scaled Age)^2	0.0514	0.0077	0.0072	0.0623	0.0072	0.0066
Family History Prostate Cancer	0.0338	0.0053	0.0050	0.0671	0.0072	0.0057
Family History Colon Cancer	0.4726	0.0215	0.0121	0.0085	0.0030	0.0030
Family History Endometrial Cancer	0.1403	0.0168	0.0138	0.0250	0.0038	0.0037

tionarity and half–width tests (Heidelberger and Welch 1983) and the Geweke diagnostic (Geweke 1992) with each diagnostic's parameters set to their CODA default values.

Table 2 tabulates estimated posterior means (columns 2 and 3) and standard deviations (column 4) of the coefficients in the genotype regression (i.e. components of γ_2). The second column provides estimates of the probability that the associated variable is in the model. Columns 3 and 4 summarize the posterior on the coefficient averaging over visited models that included the variable. Columns 5–7 provide the same data for each variable's complement. Of the indicator variables for subpopulations, only the NCI indicator (indicating Ashkenazi Jewish ancestry) and ascertainment mode show evidence of a significant modifying effect.

BRCAPRO, its half–penetrance version and the Finnish score are included in virtually all models. Together their coefficients sum to about 0.77 on average. The BRCAPRO age–independent relative risk (AIRR), Myriad/Frank and Penn scores are each included in about 40% of models. The average coefficients accorded each range between 0.06 and 0.075. The Yale, NCI and FHAT scores are each included in fewer than 20% of models and, when

included, none has a coefficient that exceeds 0.03 on average. No score is included in more than 5% of models as its complement and when one is, its average coefficient is estimated to be smaller than 0.01. Interpretation of the coefficients associated with the various scores is complicated by the fact that they reflect multivariate corrections for correlations between them (Genest and McConway 1990); that a score's coefficient is small or not significant does not imply that it is a poor performer. We address the relative merits of the various scores in the next section.

With exception of ovarian cancer, the counseland phenotype variables, including age, rarely enter the genotype model. The ovarian cancer variable appears in roughly 19% of models and its average coefficient is estimated to be about 0.02. This suggests that the ensemble of carrier probability scores adequately accounts for the counseland's breast and ovarian cancer phenotype and age. While there is evidence that prostate, colon and endometrial cancers are associated with BRCA1 or BRCA2 (Aretini et al. 2003; Edwards et al. 2003; Risch et al. 2001), the carrier probability scores under consideration do not explicitly account for a family history of these cancers with the exception of FHAT which depends on prostate cancer. Our results suggest that accounting for these cancer sites may improve accuracy of the scores. In particular, a family history of colon cancer is included in about half of all models and, when included, its coefficient averages about 0.022. Further, using the realizations of the BRCA1 and BRCA2 genotype vectors as covariates in binary regressions of the 3 family history indicators, we see evidence of an association between counseland's BRCA1 genotype and a family history of colon cancer (averaging over genotype realizations, odds ratio = 1.51, p-value = 0.0007) and weak evidence of an association between BRCA1 and family history of endometerial cancer (average $OR = 1.36$, $p = 0.11$). This evidence is indirect because the family history indicators do not take into account the number of affected relatives, their relation to the counseland and their ages at diagnosis.

Table 3 tabulates *a posteriori* estimates of sensitivity for each genetic testing modality.

Table 3: Estimated posterior means and standard deviations (SD) of test sensitivity for the various BRCA1 and BRCA2 testing procedures employed in the CGN validation dataset. Targeted mutation screening and assays for the Ashkenazi founder mutations BRCA1 185delAG, BRCA1 5382insC and BRCA2 6174delT are gene–specific; the remaining are general procedures.

	BRCA1		BRCA2	
Testing Modality	Mean	SD.	Mean	SD
SSCP	0.52	0.043	0.55	0.05
Sequencing	0.93	0.038	0.92	0.04
ASO	0.44	0.125	0.77	0.14
Targeted Mutation Screening	0.87	0.075	—	—
Sequencing for 185delAG & 5382insC	0.84	0.073	—	—
Sequencing for 185delAG only	0.72	0.127	—	—
Sequencing for 6174delT	—	—	0.84	0.07
CSGE	0.68	0.052	0.60	0.06
Sequencing, then Follow–up Test	0.95	0.044	0.96	0.03
Targeted Mutation Screening + Sequencing	0.96	0.030	0.96	0.03
SSCP + ASO	0.67	0.040	0.80	0.10
Other	0.93	0.056	0.92	0.07

Sensitivity of BRCA1 and BRCA2 testing modalities vary widely. In the case of BRCA1, they range from 0.44 to 0.96; for BRCA2 this range is 0.55 to 0.96. Modalities that involve gene sequencing are the most accurate. When used in isolation, SSCP, CSGE and ASO are the least accurate. The targeted strategies (sequencing for a panel of known mutations or for founder or family mutations) are intermediate to these extremes, but largely because they rely on prior information such as ethnicity or knowledge that a specific variant was found in a family member. Interestingly, with exception of ASO, methods applied to both BRCA1 and BRCA2 are estimated to have comparable sensitivities across genes. While it is possible that imperfect sensitivity of these tests may result from an undetected "BRCA3," there is no evidence that an additional major breast cancer gene exists, though some minor candidates could contribute to the negatives (Lacroix et al. 2005).

3.5 Evaluation of Scores Against the Gold Standard

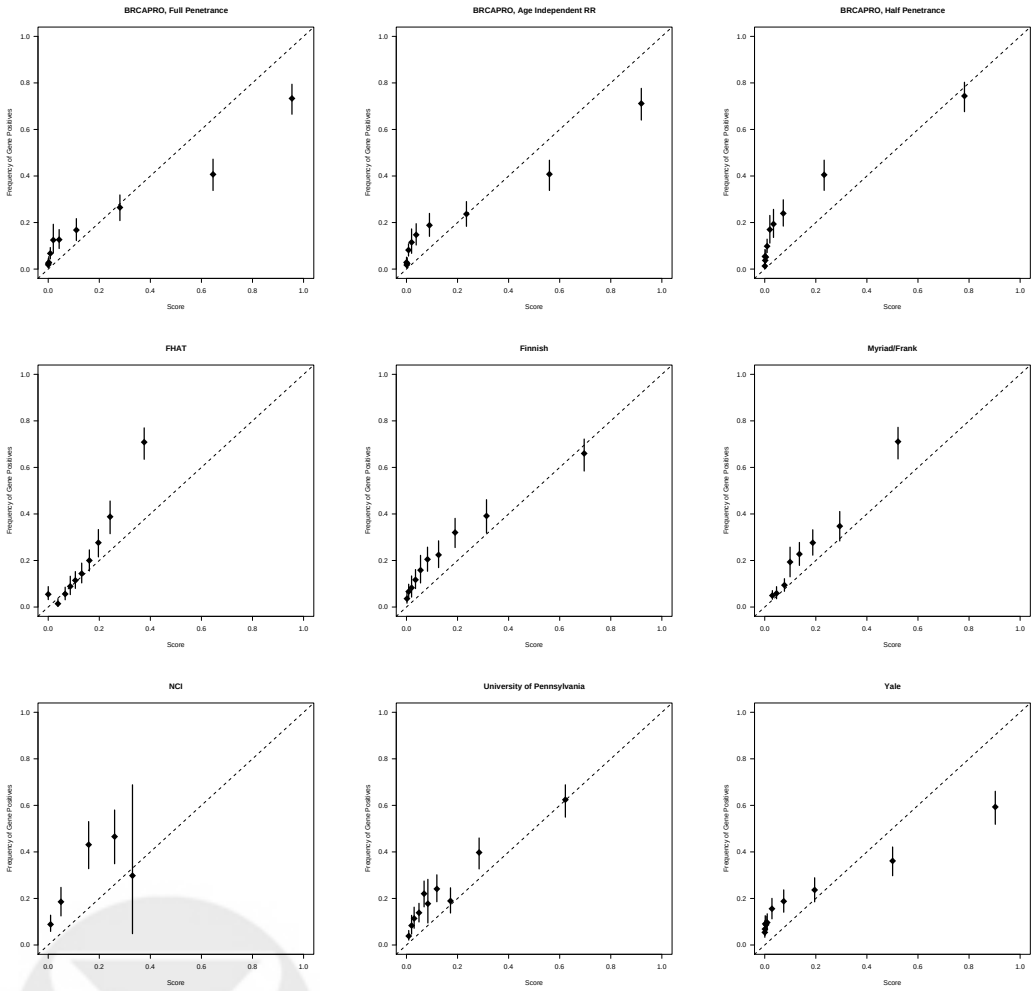
In what follows, we evaluate the carrier scores against the average over MCMC realizations of the latent genotype vector. The measures we focus on address model calibration, refinement and overall predictive performance.

Calibration. Figure 2 presents calibration plots for the 9 carrier probability models. This figure depicts point estimates (diamonds) and 95% equal-tailed interval estimates (vertical segments) of carrier frequencies at mean scores within deciles of score. Variation reflected in the segments results from sampling variability and uncertainty in the carrier status of test negative subjects. Note that there are fewer than ten categories in summaries of the Myriad/Frank and NCI scores because of their discrete nature. In all cases, calibration is for predicting BRCA1 or BRCA2 genotype among all individuals with the exception of the Penn model, which is for BRCA1 carriers among individuals not testing positive for a BRCA2 mutation, and the NCI model which is restricted to AJ subjects. In these plots, a well–calibrated score would have bin means that follow the diagonal; significant departures from this indicate systematic biases. Note that the FHAT score was not designed to be calibrated and hence its plot should not be expected to be linear.

To provide a quantitative measure of departure from calibration, we calculated root mean squared error (RMSE) for each score's ability to predict genotype, G, and observed test result, $T = T_1 \vee T_2$. For each score, we calculated RMSE for the unadjusted score, for the score corrected for an estimate of bias, RMSE(1), and for the score corrected for decile–specific estimates of bias, RMSE(D). Summaries for each carrier score are calculated using only data from the population to which the score pertains (for example, Ashkenazi Jewish individuals for the NCI score). These data are presented in Table 4. RMSE(D) serves as an *ad hoc* measure of calibration: the closer its value is to RMSE, the better the calibration of the score. For example, the unadjusted NCI score has an RMSE for *test result* of 0.372 while its adjusted (for D) RMSE is 0.356. Of this difference, most is due to a global downward bias: RMSE(1) is 0.360. In contrast, correcting BRCAPRO for D reduces its RMSE for *genotype* from 0.348 to 0.323 with most of this difference due to a correction for mis–calibration.

Overall Accuracy. Root mean–squared error (RMSE) and bias provide measures of

Figure 2: Calibration plots for the 9 carrier probability models. For each model, subjects are divided into deciles bins according to their carrier score. The proportion of carriers in each bin is indicated by a circle (mean) and a vertical segment (95% interval). Variation reflected in the segments results from sampling variability and uncertainty in the carrier status of test negative subjects. In all cases, calibration is for joint BRCA1 and BRCA2 genotype among all individuals with exception of the Penn model, which is for BRCA1 carriers among individuals not testing positive for a BRCA2 mutation, and the NCI model which is for a BRCA1 or BRCA2 founder mutation among AJ subjects.



overall predictive performance. Estimates of these quantities are presented in Table 4 both for predicting genotype and test result. Across subjects, BRCAPRO and its variants have the lowest RMSE and bias for predicting genotype. In addition, it has the highest concordance index (AUC) against both genotype and test result. The score labeled "Prevalence" corresponds to using the empirical prevalence of mutation (columns 2–6) or test result (columns 7–11) as the score assigned to all individuals. This score is constant, hence unrefined. It

Table 4: Accuracy, calibration and refinement for each score's ability to predict genotype, G, (columns 1 – 5) and test result, $T = T_1 \vee T_2$, (columns 6 – 10) among all study subjects. The columns headed RMSE(1) and RMSE(D) are described in the text. AUC is area under the ROC curve. Note that summaries for a carrier score are calculated using only data from the population to which the score pertains (for example, Ashkenazi Jewish individuals for the NCI score). Note: (†) "Prevalence" corresponds to using the empirical prevalence of mutation (columns 2–6) or test result (columns 7–11) as the score assigned to all individuals in the highlighted subset; (‡) estimated prevalence can be found in the 'Bias' column; bias is zero when using the fixed prevalence of mutation as a score.

	Predicting Genotype					Predicting Test Result				
Carrier Score	Bias	RMSE	RMSE(1)	RMSE(D)	AUC	Bias	RMSE	RMSE(1)	RMSE(D)	AUC
BRCAPRO	0.014	0.348	0.348	0.323	0.838	0.061	0.352	0.346	0.310	0.830
BRCAPRO, 50%	-0.078	0.342	0.333	0.326	0.825	-0.031	0.320	0.319	0.307	0.817
BRCAPRO, AIRR	-0.007	0.348	0.348	0.333	0.830	0.040	0.344	0.342	0.312	0.823
Yale	-0.021	0.384	0.384	0.361	0.736	0.026	0.376	0.375	0.335	0.716
Myriad/Frank	-0.054	0.354	0.350	0.346	0.764	-0.007	0.317	0.317	0.316	0.763
NCI	-0.138	0.412	0.388	0.381	0.626	-0.097	0.372	0.360	0.356	0.621
Penn Model	-0.072	0.377	0.370	0.367	0.736	-0.019	0.338	0.337	0.332	0.744
FHAT	-0.057	0.360	0.355	0.341	0.796	-0.010	0.322	0.322	0.317	0.785
Finnish	-0.072	0.379	0.372	0.369	0.767	-0.017	0.345	0.344	0.339	0.768
Prevalence†	0.191 ‡	0.393	—	—	—	0.144 ‡	0.351	—	—	—

is also unbiased; in the bias columns for this "score" we report the empirical prevalence of mutation carriers (column 2) and of test positive subjects (column 7).

Were the comparisons to be made using test result as the gold standard, the Myriad/Frank model, the half–penetrance variant of BRCAPRO and the FHAT model would have appeared superior on basis of RMSE. This highlights that fact that comparison of the scores is sensitive to the choice of gold standard. While multiplying the Mendelian predictions by test sensitivity would likely improve their bias and RMSE for predicting test result, we chose not to make this adjustment as test result is not the clinically relevant quantity.

Carrier probability scores are often employed for a specific purpose (e.g. genetic counseling) in a particular setting or population (e.g. high–risk clinic), hence it is of interest to evaluate the scores accordingly. The two primary population distinctions we choose to focus on are high–risk versus population–based ascertainment and subjects with versus without Ashkenazi Jewish heritage. A comparison of the scores against genotype in these populations is presented in panels (a) through (d) of Table 5. Note that using an empirical estimate of prevalence as the score among population–based cases is better calibrated for mutation status than five of the scores under investigation.

The subject's cancer history is the single most important input to each of the carrier

Table 5: Summaries related to the over all accuracy, calibration and refinement of the various carrier probability models as displayed in Table 4. Each performance summary was calculated and is displayed for (a) high risk subjects only, (b) population–based subjects only, (c) Ashkenazi Jewish Subjects only and (d) non–Ashkenazi Jewish subjects only. (e) cancer–free subjects, (f) subjects affected with breast cancer only, (g) subjects affected with ovarian cancer only and (h) subjects affected with both breast and ovarian cancer. Note: (†) the score named "Prevalence" corresponds to using the empirical prevalence of mutation as the score assigned to all individuals in the highlighted subset; (‡) estimated prevalence can be found in the 'Bias' column. Bias is zero when using the fixed prevalence of mutation as a score.

Carrier Score	Bias	RMSE	RMSE(1)	RMSE(D)	AUC	Bias	RMSE	RMSE(1)	RMSE(D)	AUC
	(a) High Risk Subjects					(b) Population–Based Subjects				
BRCAPRO	0.008	0.431	0.430	0.406	0.770	0.020	0.239	0.238	0.215	0.842
BRCAPRO, 50%	-0.136	0.432	0.410	0.402	0.761	-0.019	0.216	0.215	0.214	0.799
BRCAPRO, AIRR	-0.022	0.433	0.433	0.412	0.756	0.009	0.233	0.233	0.217	0.838
Yale	-0.060	0.483	0.479	0.441	0.651	0.018	0.249	0.248	0.224	0.751
Myriad/Frank	-0.135	0.449	0.428	0.426	0.692	0.027	0.220	0.218	0.218	0.788
NCI	-0.229	0.511	0.457	0.452	0.538	-0.030	0.245	0.243	0.241	0.553
Penn Model	-0.140	0.454	0.432	0.428	0.689	0.011	0.252	0.252	0.246	0.740
FHAT	-0.142	0.457	0.434	0.422	0.718	0.028	0.223	0.221	0.217	0.827
Finnish	-0.133	0.450	0.430	0.423	0.728	0.010	0.256	0.256	0.251	0.776
Prevalence†	0.321 ‡	0.467	—	—	—	0.061 ‡	0.238	—	—	—
	(c) Ashkenazi Jewish Subjects					(d) Non–Ashkenazi Jewish Subjects				
BRCAPRO	0.081	0.396	0.387	0.356	0.809	-0.007	0.333	0.333	0.319	0.846
BRCAPRO, 50%	-0.141	0.394	0.368	0.353	0.778	-0.059	0.325	0.319	0.314	0.846
BRCAPRO, AIRR	0.054	0.390	0.387	0.360	0.796	-0.025	0.334	0.333	0.321	0.839
Yale	-0.105	0.405	0.391	0.379	0.673	0.004	0.378	0.378	0.349	0.762
Myriad/Frank	-0.056	0.372	0.367	0.362	0.719	-0.053	0.348	0.344	0.339	0.758
NCI	-0.138	0.412	0.388	0.381	0.626	—	—	—	—	—
Penn Model	-0.044	0.392	0.390	0.387	0.737	-0.079	0.373	0.365	0.361	0.728
FHAT	-0.109	0.394	0.378	0.361	0.776	-0.042	0.348	0.346	0.331	0.814
Finnish	-0.164	0.429	0.396	0.389	0.729	-0.048	0.365	0.362	0.358	0.787
Prevalence†	0.214 ‡	0.410	—	—	—	0.184 ‡	0.387	—	—	—
	(e) Unaffected Subjects					(f) Breast Cancer Affected Subjects				
BRCAPRO	-0.040	0.288	0.286	0.284	0.730	0.041	0.371	0.369	0.340	0.834
BRCAPRO, 50%	-0.079	0.299	0.289	0.281	0.682	-0.069	0.353	0.346	0.338	0.823
BRCAPRO, AIRR	-0.042	0.292	0.289	0.287	0.725	0.009	0.369	0.369	0.347	0.822
Yale	-0.069	0.304	0.296	0.293	0.583	0.031	0.400	0.399	0.364	0.766
Myriad/Frank	-0.033	0.295	0.293	0.287	0.693	-0.062	0.374	0.369	0.365	0.721
NCI	-0.087	0.321	0.309	0.306	0.401	-0.158	0.463	0.435	0.431	0.554
Penn Model	-0.057	0.323	0.318	0.316	0.754	-0.077	0.385	0.377	0.374	0.696
FHAT	-0.029	0.288	0.286	0.283	0.706	-0.056	0.375	0.371	0.359	0.771
Finnish	-0.081	0.331	0.321	0.316	0.785	-0.068	0.386	0.380	0.376	0.730
Prevalence†	0.105 ‡	0.306	—	—	—	0.205 ‡†	0.404	—	—	—
	(g) Ovarian Cancer Affected Subjects					(h) Breast and Ovarian Cancer Affected Subjects				
BRCAPRO	-0.035	0.354	0.352	0.336	0.880	0.049	0.357	0.353	0.325	0.875
BRCAPRO, 50%	-0.135	0.385	0.360	0.338	0.844	-0.125	0.401	0.380	0.348	0.874
BRCAPRO, AIRR	-0.036	0.358	0.356	0.340	0.873	0.066	0.357	0.350	0.334	0.865
Yale	-0.232	0.472	0.411	0.371	0.572	-0.260	0.508	0.435	0.375	0.837
Myriad/Frank	-0.025	0.358	0.357	0.327	0.839	-0.155	0.424	0.394	0.370	0.791
NCI	-0.337	0.600	0.495	0.470	0.430	-0.629	0.719	0.343	0.170	0.873
Penn Model	-0.076	0.463	0.456	0.430	0.640	-0.043	0.410	0.407	0.373	0.784
FHAT	-0.083	0.397	0.388	0.348	0.824	-0.345	0.546	0.423	0.367	0.828
Finnish	-0.057	0.412	0.407	0.365	0.822	-0.134	0.436	0.414	0.361	0.827
Prevalence†	0.260 ‡	0.437	—	—	—	0.651 ‡	0.474	—	—	—

scores under consideration, hence it is of interest to evaluate sensitivity of the scores to this history. Panels (e) through (h) of Table 5 compare the scores when the sample is stratified by prior history of breast and ovarian cancer. We estimate that, on average, each of the carrier scores underestimates the probability of a mutation among subjects with no history of breast or ovarian cancer and among subjects with only an ovarian cancer history. In all

strata, a BRCAPRO variant is the most calibrated score.

Refinement. Figure 3 plots ROC curves for the various carrier probability scores. In all cases these curves are for predicting genotype. The three BRCAPRO variants dominate this plot and are followed by the FHAT and Myriad/Frank scores.

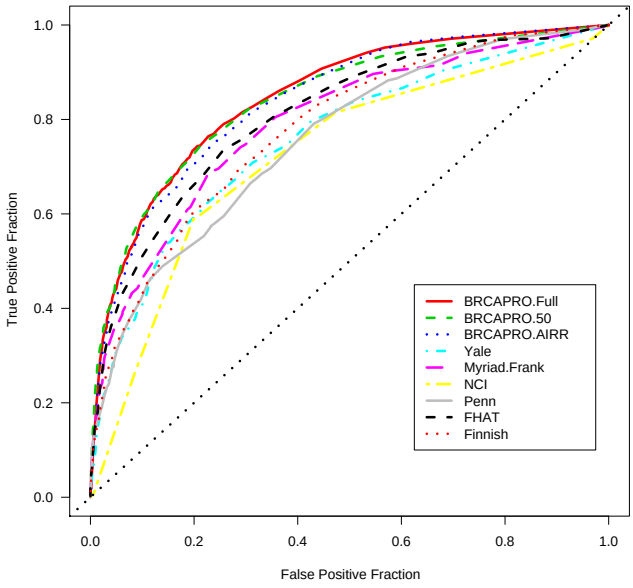


Figure 3: Receiver operating characteristic (ROC) curves for the various carrier probability models predicting binary genotype (presence or absence of a disease associated mutation at BRCA1 or BRCA2). By this measure, BRCAPRO and its variants are the most refined scores. Estimates of area under the curve (AUC) are presented in Table 4.

3.6 Sensitivity to the Prior

We carried out a parallel analysis with unconstrained, independent uniform priors on the test sensitivity parameters as a check to the robustness of the model. The remaining priors in the model, which were chosen to be minimally informative, were left unchanged. We compared the average value of the latent BRCA1 and BRCA2 genotypes in the subset of test–negative individuals under the two models and found them to be highly correlated: we calculated this correlation to be 0.87 for the gold standard BRCA1 genotype and 0.79 for

the gold standard BRCA2 genotype. Further, 87% of the averaged latent BRCA1 genotypes under the two priors were within 0.05 of one another; for BRCA2 this figure was 89%. Given this level of concordance, it is not surprising that the comparison of scores was highly robust to the choice of prior: correlations between the estimates in columns 1–5 of Table 4 and their counterparts estimated under the model with the uninformative prior on assay sensitivities, presented in Table 6 all exceeded 0.99.

Table 6: Main results under alternative prior on assay sensitivity parameters: accuracy, calibration and refinement for each score's ability to predict genotype, G, among all study subjects calculated with independent uniform priors on the assay sensitivity parameters. Compare with columns 1–5 of Table 4. Note that summaries for a carrier score are calculated using only data from the population to which the score pertains (for example, Ashkenazi Jewish individuals for the NCI score).

Carrier Score	Bias	RMSE	RMSE(1)	RMSE(D)	AUC
BRCAPRO	-0.001	0.351	0.350	0.335	0.839
BRCAPRO, 50%	0.091	0.353	0.341	0.334	0.830
BRCAPRO, AIRR	0.020	0.353	0.352	0.340	0.830
Yale	0.034	0.390	0.388	0.368	0.754
Myriad/Frank	0.067	0.365	0.358	0.354	0.760
NCI	0.156	0.428	0.398	0.389	0.627
Penn Model	0.086	0.388	0.378	0.375	0.735
FHAT	0.070	0.371	0.364	0.349	0.792
Finnish	0.086	0.389	0.379	0.376	0.766

4 Discussion

In this paper, we have presented and applied a method for evaluating an ensemble of predictive models for genotype given a sample on which those models have been evaluated and for which the outcome event is measured with error. Its formulation allows for simultaneous estimation of the measurement error parameters, latent genotype and parameters that represent a relative multivariate weighting of the carrier scores. We have found that the relative performance of the scores is affected by whether genotype or an imperfect test of genotype is used in the evaluation. The Mendelian models are more accurate for predicting genotype, the quantity of interest. When assay error is ignored, performance evaluations are biased in favor of the empirical models. Our results are broadly consistent with those of (Barcenas et al. 2006) who compare a set of carrier models overlapping those studied here on a set of 472 sequenced individuals. They found that the Myriad model was comparable

to the Mendelian models BRCAPRO and BOADICEA on basis of AUC and that all performed better than the Penn model. We found that the Myriad model does less well than the Mendelian models for predicting both genotype and test result using this metric.

Because the entire analysis is conditional on the scores which are summaries of family history, it is robust to biases resulting from ascertainment to the extent these scores capture the features of family history upon which decisions to sample were made. Further, while it is possible that a score's performance may be overstated when the score is evaluated on basis of an estimated genotype that depends in part on the score, the parameterization of the regression of latent genotype on scores is designed not to favor any score or subset of scores *a priori*.

While the application we describe is specific to the setting of evaluating quantitative carrier scores for BRCA1 and BRCA2, the approach we take is more generally applicable. First, there are numerous other familial cancer syndromes (Lindor et al. 1998). Family history based scores are being developed for some of these, including the hereditary nonpolyposis colon cancer (HNPCC) syndrome with its associated disease genes MLH1 and MSH2. An evaluation of MLH1/MSH2 carrier scores would involve a similar analysis to that presented here. This isn't to suggest that the approach is limited to evaluating carrier scores for disease genes. Indeed, it is applicable in any setting where an event of interest is observed with error and where, in practice, that event is predicted using one or more quantitative scores.

References

Andrulis, I., Anton-Culver, H., Beck, J., Bove, B., Boyd, J., Buys, S., Godwin, A., Hopper, J., Li, F., Neuhausen, S., Ozcelik, H., Peel, D., Santella, R., Southey, M., van Orsouw, N., Venter, D., Vijg, J., and Whittemore, A. (2002). Comparison of DNA- and RNA-based methods for detection of truncating BRCA1 mutations. *Human Mu-*

tations, 20(1):65–73.

Aretini, P., D'Andrea, E., Pasini, B., Viel, A., Costantini, R., Cortesi, L., Ricevuto, E., Agata, S., Bisegna, R., Boiocchi, M., Caligo, M., Chieco-Bianchi, L., Cipollini, G., Crucianelli, R., D'Amico, C., Federico, M., Ghimenti, C., DeGiacomi, C., DeNicolo, A., Puppa, L., Ferrari, S., Ficorella, C., Iandolo, D., Manoukian, S., Marchetti, P., Marroni, F., Menin, C., Montagna, M., Ottini, L., Pensotti, V., Pierotti, M., Radice, P., Santarosa, M., Silingardi, V., Turchetti, D., Bevilacqua, G., and Presciuttini, S. (2003). Different expressivity of BRCA1 and BRCA2: Analysis of 179 Italian pedigrees with identified mutation. *Breast Cancer Research and Treatment*, 81:71 – 79.

Barcenas, C., Hosain, G. M. M., Arun, B., Zong, J., Zhou, X., Chen, J., Cortada, J. M., Mills, G. B., Tomlinson, G. E., Miller, A. R., Strong, L. C., and Amos, C. I. (2006). Assessing brca carrier probabilities in extended families. *Journal of Clinical Oncology*, 24:354–360.

Berry, D. A., Parmigiani, G., Sanchez, J., Schildkraut, J., and Winer, E. (1997). Probability of carrying a mutation of breast-ovarian cancer gene BRCA1 based on family history. *J Natl Cancer Inst*, 89:227–238.

Bordley, R. F. (1982). A multiplicative formula for aggregating probability assessments. *Management Science*, 28:1137–1148.

Claus, E. B., Risch, N., and Thompson, W. D. (1990). Age of onset as an indicator of familial risk of breast cancer. *American Journal of Epidemiology*, 131:961–972.

Claus, E. B., Risch, N., and Thompson, W. D. (1991). Genetic analysis of breast cancer in the cancer and steroid hormone study. *American Journal of Human Genetics*, 48:232–242.

Clyde, M. and George, E. (2004). Model uncertainty. *Statistical Science*, 19:81—94.

Couch, F. J., DeShano, M. L., Blackwood, M. A., Calzone, K., Stopfer, J., Campeau, L., Ganguly, A., Rebbeck, T., Weber, B. L., Jablon, L., Cobleigh, M. A., Hoskins, K., and Garber, J. E. (1997). BRCA1 mutations in women attending clinics that evaluate the risk of breast cancer. *N Engl J Med*, 336:1409–15.

Dawid, A., DeGroot, M., and Mortera, J. (1995). Coherent combinations of experts' opinions (with discussion). *Test*, 4(2):263 – 313.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *The Statistician*, 32:12–22.

Edwards, S., Kote-Jarai, Z., Meitz, J., Hamoudi, R., Hope, Q., Osin, P., Jackson, R., Southgate, C., Singh, R., Falconer, A., Dearnaley, D., Ardern-Jones, A., Murkin, A., Dowe, A., Kelly, J., Williams, S., Oram, R., Stevens, M., Teare, D., Ponder, B., Gayther, S., Easton, D., and Eeles, R. (2003). Two percent of men with early-onset prostate cancer harbor germline mutations in the BRCA2 gene. *American Journal of Human Genetics*, 72:1 – 12.

Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity*, 21:523–542.

Eng, C., Brody, L., Wagner, T., Devilee, P., Vijg, J., Szabo, C., Tavtigian, S., Nathanson, K., Ostrander, E., and Frank, T. (Dec, 2001). Interpreting epidemiological research: blinded comparison of methods used to estimate the prevalence of inherited mutations in BRCA1. *Journal of Medical Genetics*, 38(12):824–833.

Euhus, D., Smith, K., Robinson, L., and *et al.* (2002). Pretest prediction of BRCA1 or BRCA2 mutation by risk counselors and the computer model BRCAPRO. *Journal of the National Cancer Institute*, 94(11):844 – 851.

Fraedrich, K. and Smith, N. (1989). Combining predictive schemes in long-range forecasting. *Journal of Climate*, 2:291 – 294.

Frank, T., Deffenbaugh, A., Reid, J., and *et al.* (2002). Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: analysis of 10,000 individuals. *Journal of Clinical Oncology*, 20(6):1480 – 1490.

Frank, T., Manley, S., Olopade, O., Cummings, S., Garber, J., Bernhardt, B., Antman, K., and et al. (1998). Sequence analysis of brca1 and brca2: Correlation of mutations with family history and ovarian cancer risk. *J Clin Oncol*, 16:2417–2425.

French, S. (1985). Group consensus probability distributions: a critical survey. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics*, volume II, pages 183–201. Amsterdam: North Holland.

French, S. (1995). Discussion of coherent combinations of experts' opinions. *Test*, 4(2):294–296.

Geisler, J., Hatterman-Zogg, M., Rathe, J., Lallas, T., Kirby, P., and Buller, R. (Oct, 2001). Ovarian cancer BRCA1 mutation detection: Protein truncation test (PTT) outperforms single strand conformation polymorphism analysis (SSCP). *Human Mutations*, 18(4):337–344.

Gel, Y., Raftery, A., and Gneiting, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (gop) method (with discussion). *Journal of the American Statistical Association*, 99:575–590.

Genest, C. and McConway, K. (1990). Allocating the weights in the linear opinion pool. *Journal of Forecasting*, 9:53 – 73.

Genest, C. and Schervish, M. J. (1985). Modeling expert judgements for Bayesian updating. *The Annals of Statistics*, 13:1198–1212.

Genest, C. and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1:114–148.

Geweke, J. (1992). Evaluating the accuracy of sampling–based approaches to calculating posterior moments. In Bernado, J., erger, J., AP, D., and Smith, A., editors, *Bayesian Statistics 4*. Clarendon Press, Oxford, UK.

Gilpin, C., Carson, N., and Hunter, A. (2000). A preliminary validation of a family history assessment form to select women at risk for breast or ovarian cancer for referral to a genetics center. *Clinical Genetics*, 58(4):299 – 308.

Gneiting, T., Raftery, A., Westveld, A., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133:1098 – 1118.

Harrell Jr, F., Califf, R., Pryor, D., Lee, L., and Rosati, R. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247:2543 – 2546.

Hartge, P., Struewing, J. P., Wacholder, S., Brody, L. C., and Tucker, M. A. (1999). The prevalence of common BRCA1 and BRCA2 mutations among Ashkenazi Jews. *Am. J. Hum. Genet*, 64:963–970.

Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31:1109–1144.

Hui, S. L. and Walter, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36:167–171.

Iversen Jr., E. S., Parmigiani, G., and Berry, D. (1999). Validating Bayesian prediction models: a case study in genetic susceptibility to breast cancer. In *Case Studies in Bayesian Statistics Volume IV*, pages 321–338.

Key, J., Pericchi, L., and Smith, A. (1999). Bayesian model choice: What and why? In *Bayesian Statistics 6*, pages 343–370.

Kharin, V. and Zwiers, F. (2002). Climate predictions with multimodel ensembles. *Journal*

of Climate, 15:793 – 799.

Kraft, P. and Thomas, D. C. (2000). Bias and efficiency in family–based gene–characterization studies : Conditional, prospective, retrospective, and joint likelihoods. *American Journal of Human Genetics*, 66:1119–1131.

Krishnamurti, T., Kishtawal, C., LaRow, T., Bachiochni, D., Zhang, Z., Willifor, C., Gadgil, S., and Surendran, S. (1999). Improved weather forecasts and seasonal climate forecasts from multimodel superensemble. *Science*, 285:1548 – 1550.

Lacroix, M., Leclercq, G., and BreastMed Consortium (2005). The portrait of hereditary breast cancer. *Breast Cancer Research and Treatment*, 89:297–304.

Lindor, N., Greene, M., and the Mayo Familial Cancer Program (1998). The concise handbook of family cancer syndromes. *Journal of the National Cancer Institute*, 90(14):1039–1071.

Makridakis, S. and Winkler, R. (1983). Averages of forecasts: some empirical results. *Management Science*, 29(9):987–996.

Myriad Genetics (2003). BRACAnalysis technical specification. Technical report, Myriad Genetics Laboratories.

Parmigiani, G., Berry, D. A., and Aguilar, O. (1998). Determining carrier probabilities for breast cancer susceptibility genes BRCA1 and BRCA2. *American Journal of Human Genetics*, 62:145–158.

Parmigiani, G., Friebel, T., Iversen Jr., E., Chen, S., Finkelstein, D., Anton-Culver, H., Ziogas, A., Weber, B., Eisen, A., Malone, K., Hsu, L., Peterson, L., Schildkraut, J., Isaacs, C., Corio, C., Leondaridis, L., Tomlinson, G., Amos, C., Strong, L., Berry, D., Weitzel, J., Sand, S., Dutson, D., Kerber, R., Peshkin, B., and Euhus, D. (2007). Validity of models for prediction of BRCA1 and BRCA2 mutations. *Annals of Internal Medicine*, 147(5):To Appear.

Raftery, A. E. and Lewis, S. M. (1996). Implementing MCMC. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, pages 115–127, London. Chapman and Hall.

Risch, H. A., McLaughlin, J. R., Cole, D. E. C., Rosen, B., Bradley, L., Kwan, E., Jack, E., Vesprini, D. J., Kuperstein, G., Abrahamson, J. L. A., Fan, I., Wong, B., and Narod, S. A. (2001). Prevalence and penetrance of germline BRCA1 and BRCA2 mutations in a population series of 649 women with ovarian cancer. *American Journal of Human Genetics*, 68:700–710.

Roa, B. B., Boyd, A. A., Volcik, K., and Richards, C. S. (1996). Ashkenazi Jewish population frequencies for common mutations in BRCA1 and BRCA2. *Nature Genetics*, 14:185–187.

Thompson, P. (1977). How to improve accuracy by combining independent forecasts. *Monthly Weather Review*, 105:228 – 229.

Tierney, L. (1994). Markov chains for exploring posterior distributions (disc: P1728-1762). *The Annals of Statistics*, 22:1701–1728.

Vahteristo, P., Eerola, H., Tamminen, A., Blomqvist, C., and Nevanlinna, H. (2001). A probability model for redicting BRCA1 and BRCA2 mutations in breast and breast-ovarian cancer families. *British Journal of Cancer*, 84(5):704 – 708.

Walter, S. D. and Irwig, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology*, 41:923–937.