DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Statistical and Machine Learning Methods for Clinical Risk Prediction

Citation

Guan, Zoe. 2020. Statistical and Machine Learning Methods for Clinical Risk Prediction. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Permanent link

https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37365776

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

©2020 – Zoe Guan
all rights reserved.

Statistical and Machine Learning Methods for Clinical Risk Prediction

Abstract

In many areas of healthcare, clinical prediction models are used to assess disease risk and guide decisions about prevention and treatment. Accurate risk stratification is key to reducing morbidity and mortality through the effective delivery of precision medicine. This dissertation proposes and compares methods for improving the accuracy of risk prediction models through the integration of different models and/or datasets and the adaptation of machine learning algorithms that have achieved high accuracy in other prediction problems. Chapters 1 and 2 focus on cancer risk prediction, while Chapter 3 addresses general settings where multiple studies are available for training and validation.

In Chapter 1, we propose to combine existing breast cancer risk prediction models that embed complementary information. Numerous models have been developed, but they often give predictions with conflicting clinical implications. Integrating information from different models can potentially improve the accuracy of risk predictions. BRCAPRO and BCRAT are two widely used models that are based on different risk factors and methodologies. BRCAPRO is a Mendelian model that uses detailed family history information to estimate the probability of carrying a BRCA1/2 mutation, as well as future risk of breast and ovarian cancer, based on mutation prevalence and penetrance (age-specific probability of developing cancer given genotype). BCRAT uses a relative hazard model based on first-degree family history and non-genetic risk factors. We consider two approaches for combining BRCAPRO and BCRAT: 1) modifying the penetrance functions in BRCAPRO using relative hazard estimates from BCRAT, and 2) training an ensemble model that takes as input BRCAPRO and BCRAT predictions. We assess the performance of the combination models in simulations and data from the Cancer Genetics Network, and show that they achieve performance gains over BRCAPRO and BCRAT among individuals with a strong family history of cancer.

In Chapter 2, we propose to adapt neural networks for family history-based breast cancer risk prediction. The prevailing models for assessing familial risk of breast cancer are Mendelian models, but these models rely on many assumptions about cancer susceptibility genes. Training more flexible models, such as neural networks, on large datasets can potentially lead to accuracy gains. While there is an extensive literature on neural networks and their state-of-the-art performance in many tasks, there is little work applying them to family history data. The neural network models we propose eliminate the need to explicitly specify the effects of cancer susceptibility genes, overcoming one of the main limitations of Mendelian models. In data simulated under Mendelian inheritance, we demonstrate that neural networks are able to achieve nearly optimal prediction performance. Moreover, when the data generated from a Mendelian model are subject to misreporting of cancer diagnoses, neural networks are able to outperform the Mendelian model. Using a large dataset of over 200,000 family histories from the Risk Service, we train neural networks to predict future risk of breast cancer. We validate them using data from the Cancer Genetics Network and show that they achieve competitive performance with BRCAPRO.

In Chapter 3, we compare methods for training prediction models using multiple studies. In precision medicine and other settings, systematic data sharing and data curation initiatives are opening opportunities for developing and validating models on multiple studies, which can lead to improved generalizability. Two general approaches for integrating information across studies are: 1) merging all of the studies and training a single model and 2) multi-study ensembling, which involves training a separate model on each study and combining the resulting predictions. We provide theoretical and empirical analyses comparing the performance of these approaches in the presence of potential heterogeneity in predictor-outcome relationships across studies. In a linear regression setting, we show analytically and confirm via simulations that merging yields lower prediction error than ensembling when the effects of the predictors are relatively homogeneous across studies. However, as heterogeneity increases, there exists a transition point beyond which ensembling outperforms merging. We provide analytic expressions for the transition point in various scenarios, study asymptotic properties, and illustrate how transition point theory can be used for deciding when to merge versus when to ensemble using metagenomic data.

Contents

Author List

The following authors contributed to Chapter 1: Theodore Huang, Anne Marie McCarthy, Kevin S. Hughes, Alan Semine, Hajime Uno, Lorenzo Trippa, Giovanni Parmigiani, Danielle Braun.

The following authors contributed to Chapter 2: Giovanni Parmigiani, Danielle Braun, Lorenzo Trippa.

The following authors contributed to Chapter 3: Giovanni Parmigiani, Prasad Patil.

List of Figures

List of Tables

Acknowledgments

I am extremely grateful to my advisor, Giovanni Parmigiani, and my de facto co-advisor, Danielle Braun, for their keen insights, thoughtful guidance, and invaluable support throughout my graduate studies.

Many thanks to my other committee members, Lorenzo Trippa and Hajime Uno, and to Prasad Patil, who generously gave their time and expertise to advise me on various aspects of my research.

I would like to thank the BayesMendel and Multi-Study Machine Learning groups for their feedback and support. In particular, Theodore Huang provided helpful code and suggestions for Chapter 1, Matthew Ploenzke provided helpful suggestions for Chapter 2, and Boyu Ren provided helpful suggestions for Chapter 3.

I would also like to thank my collaborators, Anne Marie McCarthy and Kevin Hughes, for providing valuable data and feedback.

Finally, thank you to my family and friends for their constant support.

1

Combining Breast Cancer Risk Prediction Models

1.1 Introduction

Breast cancer is the second most common cancer and the second leading cause of cancer death in women in the U.S (Siegel et al., 2020; American Cancer Society, 2020). Identifying individuals at high risk is critical for guiding decisions about risk management and prevention, including screening, genetic counseling and testing, and preventative procedures. In clinical practice, at least 24 breast cancer risk prediction models have been developed to help identify higher-risk individuals (Cintolo-Gonzalez et al., 2017). These models estimate an individual's risk of carrying a pathogenic mutation in a breast cancer susceptibility gene and/or an individual's future risk of breast cancer, and they are based on a wide range of risk factors, methodologies, and study populations. Some models, such as the Breast Cancer Risk Assessment Tool (BCRAT) (Gail et al., 1989, 2007; Matsuno et al., 2011; Banegas et al., 2016), are regression-based models that use hormonal/reproductive risk factors (such as age at first live birth) and summaries of family history. Others, such as BRCAPRO (Parmigiani et al., 1998), BOADICEA (Antoniou et al., 2004, 2008; Lee et al., 2019), and IBIS, (Tyrer et al., 2004), use detailed family history information and principles of genetic inheritance. IBIS and BOADICEA (Lee et al., 2019) also take into account non-genetic risk factors. Different models can output discordant risk predictions with conflicting treatment implications (Jacobi et al., 2009; Ozanne et al., 2013). One solution is to select a single model upon which to base intervention decisions (Collins et al., 2016; Phillips et al., 2019). However, identifying the most accurate model for a given patient can be difficult, and, even if such a model is identified, other models could still contribute additional relevant information. Thus, it is important to systematically integrate information from different models to improve risk stratification and reduce cancer morbidity.

We investigate methods for combining BRCAPRO (Parmigiani et al., 1998) and BCRAT (Gail

et al., 1989, 2007; Matsuno et al., 2011; Banegas et al., 2016), two widely used and validated (Spiegelman et al., 1994; Rockhill et al., 2001; Berry et al., 2002; Amir et al., 2003; Terry et al., 2019; McCarthy et al., 2019) breast cancer risk prediction models based on different approaches and risk factors. BRCAPRO is a family history-based model that provides carrier probabilities for breast cancer susceptibility genes BRCA1 and BRCA2 as well as future risk estimates for invasive breast cancer and for ovarian cancer. It translates family history data into risk estimates using Mendelian laws of inheritance, Bayes' rule, and literature-based estimates of mutation prevalence and penetrance (age-specific probability of developing cancer given genotype). BCRAT estimates an individual's future risk of invasive breast cancer based on a relative hazard model that includes age, hormonal and reproductive risk factors, breast biopsy history, and first-degree family history of breast cancer. The model was originally developed using case-control data from white women participating in a U.S. mammography screening program and was later updated for African-American (Gail et al., 2007), Asian-American (Matsuno et al., 2011), and Hispanic (Banegas et al., 2016) women.

Although there is some overlap in the inputs to BRCAPRO and BCRAT, the two models are largely complementary (Figure 1.1). A validation study in a U.S. screening population found that 6-year risk predictions from BRCAPRO and BCRAT had a moderate correlation of 0.53 (McCarthy et al., 2019). BRCAPRO uses extensive family history information while BCRAT considers only first-degree relatives. BCRAT considers several non-genetic risk factors that are not considered by BRCAPRO, including age at menarche, age at first live birth, and breast biopsies. Therefore, combining the two models could potentially improve predictions compared to BRCAPRO or BCRAT alone. We consider two combination approaches: 1) penetrance modification and 2) training an ensemble model.

The first approach involves modifying penetrance functions (age-specific probabilities of developing cancer given genotype) to account for additional risk factors. We develop a penetrance modification model, BRCAPRO+BCRAT (M), based on an approach that incorporates BCRAT relative hazards into the BRCAPRO penetrance functions using a relative hazard model (Liu et al., 2013). This approach has similarities to the one used in IBIS, a hybrid of a family history-based model and a proportional hazards model for other risk factors.

Ensemble learning involves training multiple base models and combining their predictions. A wide variety of ensemble methods have been developed, including stacking (Wolpert, 1992), bagging (Breiman, 1996, 2001), boosting (Freund et al., 1996), and Bayesian model averaging (Hoeting et al., 1998). The extensive literature (mostly empirical, though some theoretical work has been done) shows that ensembles can outperform their base models (Opitz and Maclin, 1999; Dietterich, 2000b; Kleinberg, 1990; Perrone and Cooper, 1992; Schapire et al., 1998; Kuncheva, 2002; Van der Laan, Polley, and Hubbard, Van der Laan et al.), especially when the base models produce dissimilar predictions (Krogh and Vedelsby, 1995; Cunningham and Carney, 2000). Ensembling often leads to accuracy gains because averaging reduces variance and can expand the sets of functions that can be represented by the base models (Dietterich, 2000a). In the setting of breast cancer risk prediction, Ming et al. (2019) compared various machine learning models to BCRAT and BOADICEA and showed that boosting and random forest, which is a form of bagging, had higher discriminatory accuracy than the existing models. In this paper, we consider an ensemble model, BRCAPRO + BCRAT (E), based on stacking, which involves training a meta-model to optimally combine predictions from the base models. We use logistic regression for the meta-model.

We assess the added predictive value of the combination models compared to BRCAPRO and BCRAT in simulations and a real data application, where we train the ensemble model on data from the Newton-Wellesley Hospital (NWH) and validate all models on data from the Cancer Genetics Network (CGN). In the data application, we also use IBIS as a reference for comparison to evaluate the relative performance of combining existing models versus developing a hybrid model from the ground up.

1.2 Methods

1.2.1 General Notation

Given a female proband (individual who presents for risk assessment) without a previous diagnosis of breast cancer, the goal is to predict her risk of developing invasive breast cancer within τ years based on family history H (described in Section 1.2.2) and other risk factors X (described in Section

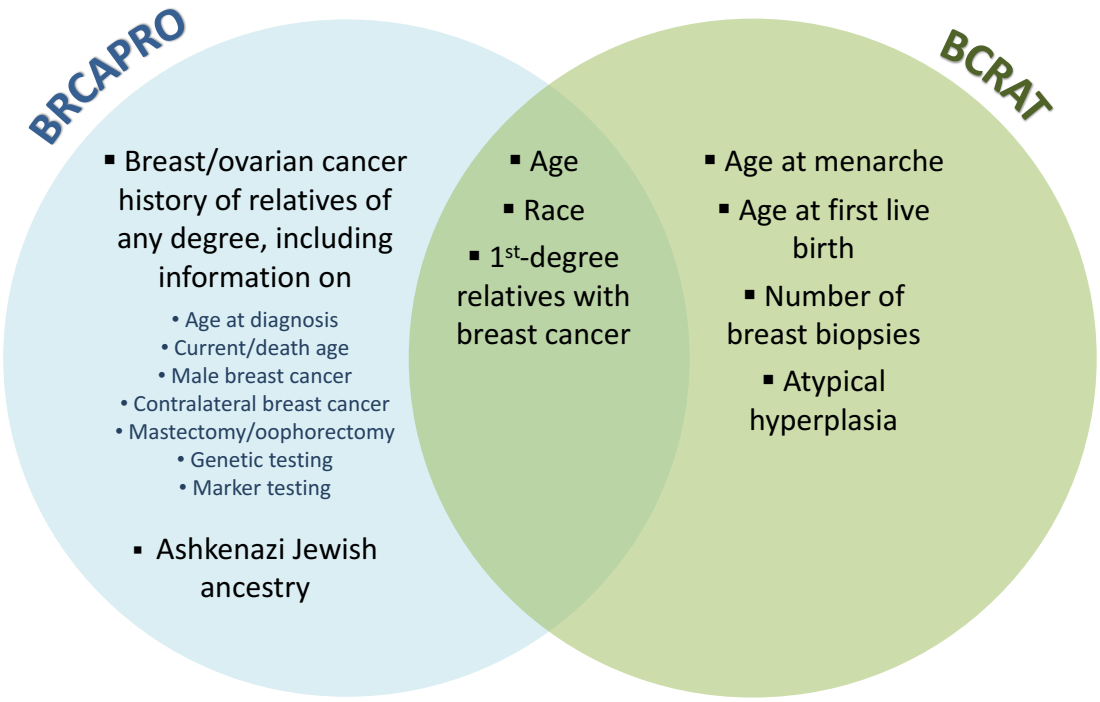

Figure 1.1: Inputs to BRCAPRO and BCRAT.

1.2.2) while accounting for death from other causes as a competing risk. τ is a pre-specified positive integer.

Let $\tilde{a}$ be the proband's current age, $\tilde{T}_B$ the age at onset of breast cancer, $\tilde{T}_D$ the age at death from other causes, and $\tilde{T} = \min(\tilde{T}_B, \tilde{T}_D)$ the time to the first event (either breast cancer or death), with $\tilde{a}$, $\tilde{T}_B$, $\tilde{T}_D$, and $\tilde{T}$ taking on continuous values in the interval $[0, \infty)$. Let $a = \lfloor \tilde{a} \rfloor$, $T_B = \lfloor \tilde{T}_B \rfloor$, $T_D = \lfloor \tilde{T}_D \rfloor$, and $T = \lfloor \tilde{T} \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function. a, T_B, T_D, and T are discrete versions of $\tilde{a}$, $\tilde{T}_B$, $\tilde{T}_D$, and $\tilde{T}$ observed at yearly intervals. Let J denote the type of the first event, with $J = B$ if $\tilde{T}_B \leq \tilde{T}_D$ and $J = D$ otherwise.

BRCAPRO uses a discrete model for T while BCRAT and IBIS use continuous models for $\tilde{T}$. We will consider discrete T in the combination models.

1.2.2 Existing Models

BRCAPRO

BRCAPRO (Parmigiani et al., 1998) estimates the probability of carrying a deleterious germline mutation in BRCA1 and BRCA2 using Bayes' rule, laws of Mendelian inheritance, mutation prevalence and penetrance, and family history. It also estimates future risk of breast and ovarian cancer based on the carrier probabilities and penetrances.

Family history can be represented as a pedigree, or a graph where each node is a family member and edges flow from parents to offspring. Let R be the number of relatives in the pedigree besides the proband and let $r = 0, 1, \ldots, R$ index the family members, where $r = 0$ corresponds to the proband. For each family member r, let H_r be a vector containing the following information on demographics and breast/ovarian cancer history: current age or age at death, gender, race/ethnicity, relation to the proband, breast cancer status, age at breast cancer diagnosis if affected, ovarian cancer status, age at ovarian cancer diagnosis if affected, genetic testing results if available, whether the individual has had a prophylactic mastectomy, mastectomy age if applicable, whether the individual has had a prophylactic oophorectomy, and oophorectomy age if applicable. Let $H = (H_0, \ldots, H_R)$.

Additionally, let G_r be family member r's BRCA1/BRCA2 genotype, with $G_r = 0$ if r is a non-carrier, $G_r = 1$ if r is a carrier of a mutation in BRCA1 only, $G_r = 2$ if r is a carrier of a mutation in BRCA2 only, and $G_r = 3$ if r is a carrier of mutations in both BRCA1 and BRCA2.

Using Bayes' rule and the assumption of conditional independence of phenotypes given genotypes (across individuals as well as cancer types), the proband's probability of having genotype G_0 is

$$P(G_0|H) = \frac{P(G_0) \sum_{G_1,\ldots,G_R} \prod_{r=0}^{R} P(H_r|G_r) P(G_1, \ldots, G_R|G_0)}{\sum_{G_0} P(G_0) \sum_{G_1,\ldots,G_R} \prod_{r=0}^{R} P(H_r|G_r) P(G_1, \ldots, G_R|G_0)}. \tag{1.1}$$

The summation over genotypes is calculated using the Elston-Stewart peeling algorithm (Elston and Stewart, 1971) and $P(G_1, \ldots, G_R|G_0)$ is calculated based on Mendelian laws of inheritance. The prevalences $P(G_0)$ are obtained from the literature and are ethnicity-specific (in particular, different prevalences are used for Ashkenazi Jewish and non-Ashkenazi Jewish individuals). $P(H_r|G_r)$ is calculated using literature-based penetrances for breast and ovarian cancer. The penetrances are

functions of age and are cancer- and sex-specific. The penetrance functions for non-carriers are based on rates from the Surveillance, Epidemiology, and End Results (SEER) program and are race-specific, while the penetrance functions for carriers are from a meta-analysis of published studies (Chen et al., 2020).

After estimating the carrier probabilities, BRCAPRO calculates future risk of breast and ovarian cancer through a weighted average of the genotype-specific risks. To simplify the notation, from here on we will omit the subscript 0 from G_0 (throughout the rest of the paper, we only refer to the proband's genotype and not the genotypes of the other family members). The proband's risk of developing breast cancer between ages a and $a+\tau$, conditional on having genotype g and not having breast cancer by age a, is

$$\begin{aligned} P(T \leq a+\tau, J = B | T > a, G = g) &= \sum_{t=a+1}^{a+\tau} \frac{P(T = t, J = B | T \geq t, G = g) P(T \geq t | G = g)}{P(T > a | G = g)} \\ &= \sum_{t=a+1}^{a+\tau} \lambda_B^g(t) \prod_{u=a+1}^{t-1} \left(1 - \lambda_B^g(u) - \lambda_D(u)\right) \end{aligned} \tag{1.2}$$

where $\lambda_B^g(t) = P(T = t, J = B | T \geq t, G = g)$ is the cause-specific hazard of breast cancer conditional on genotype g and $\lambda_D(t) = P(T = t, J = D | T \geq t)$ is the cause-specific hazard of death from causes other than breast cancer. $\lambda_B^g(t)$ is calculated from the female breast cancer penetrance for genotype g, $P(T = t, J = B | G = g)$, using the recursive formula

$$\lambda_B^g(t) = \frac{P(T = t, J = B | G = g)}{\prod_{u=1}^{t-1}(1 - \lambda_B^g(u) - \lambda_D(u))}, \tag{1.3}$$

while $\lambda_D(t)$ is estimated based on SEER mortality rates for all causes except breast cancer.

The final risk estimate is

$$P(T \leq a+\tau, J = B | T > a, H) = \sum_{g=0}^{3} P(T \leq a+\tau, J = B | T > a, G = g) P(G = g | H). \tag{1.4}$$

Software for running BRCAPRO is available through the BayesMendel R package (Chen et al., 2004b). We used v2.1-6.1 (selecting the crude risk option).

BCRAT

BCRAT (Gail et al., 1989, 2007; Banegas et al., 2016; Matsuno et al., 2011) estimates the relative hazard of developing breast cancer based on age (dichotomized into < 50 and ≥ 50) and the following risk factors: $X_1 =$ age at menarche, $X_2 =$ number of benign breast biopsies, $X_3 =$ age at first live birth (if nulliparous, set $X_3 = 25$), $X_4 =$ number of female first-degree relatives with breast cancer, and $X_5 =$ presence of atypical hyperplasia (0, 1, or unknown). Let $X = (X_1, \ldots, X_5)$.

The relative hazard for an individual of age t with risk factors X compared to an individual of age t with no BCRAT risk factors (other than age) is

$$\begin{aligned} r(t, X) = \exp(&\beta_1 I[X_1 \in [12, 13]] + \beta_2 I[X_1 < 12] + \\ &\beta_3 I[X_2 = 1] + \beta_4 I[X_2 \geq 2] + \beta_5 I[t \geq 50] I[X_2 = 1] + \beta_6 I[t \geq 50] I[X_2 \geq 1] + \\ &\beta_7 I[X_3 \in [20, 24]] + \beta_8 I[X_3 \in [25, 29]] + \beta_9 I[X_3 > 29] + \\ &\beta_{10} I[X_4 = 1] + \beta_{11} I[X_4 = 2] + \\ &\beta_{12} I[X_3 \in [20, 24]] I[X_4 = 1] + \beta_{13} I[X_3 \in [25, 29]] I[X_4 = 1] + \beta_{14} I[X_3 > 29] I[X_4 = 1] + \\ &\beta_{15} I[X_3 \in [20, 24]] I[X_4 \geq 2] + \beta_{16} I[X_3 \in [25, 29]] I[X_4 \geq 2] + \beta_{17} I[X_3 > 29] I[X_4 \geq 2] + \\ &\beta_{18} I[X_2 > 0] I[X_5 = 0] + \beta_{19} I[X_2 > 0] I[X_5 = 1]), \end{aligned} \tag{1.5}$$

where $I[\cdot]$ denotes the indicator function (equal to 1 if the bracketed expression is true and 0 otherwise). The relative hazard model includes interactions between age and number of biopsies, as well as age at first live birth and number of affected relatives. The regression coefficients were estimated from U.S. case-control studies. Separate models were fit to data from white, African-American, Asian, and Hispanic women to obtain race-specific estimates.

The risk of developing breast cancer between ages $\tilde{a}$ and $\tilde{a} + \tau$, conditional on not having breast cancer at age $\tilde{a}$, is

$$P(\tilde{T} \leq \tilde{a} + \tau, J = B | \tilde{T} > \tilde{a}, X) = \int_{\tilde{a}}^{\tilde{a}+\tau} \tilde{\lambda}_{B,0}(t) r(t, X) \exp\left\{ - \int_{\tilde{a}}^{t} (\tilde{\lambda}_{B,0}(u) r(u, X) + \tilde{\lambda}_D(u)) du \right\} dt, \tag{1.6}$$

where $\tilde{\lambda}_{B,0}(t) = \lim_{dt \to 0} P(t \leq \tilde{T} < t + dt, J = B | \tilde{T} \geq t, X = 0)/dt$ is the cause-specific hazard of breast

cancer for those with no BCRAT risk factors and $\tilde{\lambda}_D(u)) = \lim_{dt\to 0} P(t \le \tilde{T} < t+dt, J = D|\tilde{T} \ge t)/dt$ is the cause-specific hazard of death from causes other than breast cancer. $\tilde{\lambda}_{B,0}(t)$ is calculated from $\tilde{\lambda}_B(t) = \lim_{dt\to 0} P(t \le \tilde{T} < t+dt, J = B|\tilde{T} \ge t)/dt$, the cause-specific hazard of breast cancer in the general population, using the formula (see Gail et al. (1989))

$$\tilde{\lambda}_{B,0}(t) = \tilde{\lambda}_B(t)(1 - AR(t)), \tag{1.7}$$

where, letting $P(X|t)$ be distribution of X for age t,

$$AR(t) = 1 - \frac{1}{\sum_X r(t,X)P(X|t)}, \tag{1.8}$$

is the population attributable risk due to X for those of age t. $P(X|t)$ and $r(t,X)$ are both assumed to be constant for $t < 50$ and for $t \ge 50$, so $AR(t)$ is as well. Race-specific estimates of $\tilde{\lambda}_B(t)$ and $AR(t)$ are obtained from SEER data and $\tilde{\lambda}_D(t)$ is estimated based on SEER mortality rates for all causes except breast cancer. In the implementation of the model, the age scale is divided into 13 intervals and $\tilde{\lambda}_B(t)$ and $\tilde{\lambda}_D(t)$ are assumed to be constant on each interval (see Gail et al. (1989) for more details).

Software for running BCRAT is available through the BRCA R package (`https://cran.r-project.org/web/packages/BCRA/index.html`). We used version 2.1 of the R package.

IBIS

In our data application, we also use the IBIS model (Tyrer et al., 2004; Brentnall and Cuzick, 2019) as a reference for comparison since it combines detailed family history information with non-genetic risk factors. It first calculates carrier probabilities and risk of breast cancer based on family history, then incorporates additional risk factors Y (age at menarche, age at menopause, height, body mass index, age at first live birth, menopausal hormone therapy, atypical hyperplasia, lobular carcinoma in situ (LCIS), breast density, and SNPs) via a relative hazard model. The carrier probabilities are calculated using a similar approach as in BRCAPRO, but in addition to BRCA1 and BRCA2, IBIS considers a hypothetical low-penetrance susceptibility gene that acts as a surrogate for all other breast cancer susceptibility genes. The prevalence and penetrance of

BRCA1 and BRCA2 are obtained from the literature and the prevalence and penetrance of the hypothetical gene are estimated using data from a Swedish population-based study. The penetrance function for non-carriers is based on rates from the Thames Cancer Registry.

IBIS calculates a weighted average of the cumulative penetrances for each genotype:

$$P(\tilde{T} \leq t, J = B|H) = \sum_{g,\gamma} P(\tilde{T} \leq t, J = B|G, \Gamma)P(G = g, \Gamma = \gamma|H), \tag{1.9}$$

where Γ denotes the proband's carrier status with respect to the hypothetical gene ($\Gamma = 0$ for non-carriers and $\Gamma = 1$ for carriers). In IBIS, joint carriers of BRCA1 and BRCA2 are modelled as BRCA1 carriers.

The risk of developing breast cancer between ages a and $a + \tau$, conditional on not having breast cancer at age a (we use a instead of $\tilde{a}$ here because while IBIS is based on the continuous-time framework, integer-valued ages are used in the implementation), is

$$P(\tilde{T} \leq a + \tau, J = B|\tilde{T} > a, H, Y) = \int_a^{a+\tau} \tilde{\lambda}_{B,H}(t)s(Y) \exp\left\{-\int_a^t (\tilde{\lambda}_{B,H}(u)s(Y) + \tilde{\lambda}_D(u))du\right\} dt, \tag{1.10}$$

where $\tilde{\lambda}_{B,H}(t) = \lim_{dt\to 0} P(t \leq \tilde{T} < t + dt, J = B|\tilde{T} \geq t, H)/dt$ and $s(Y)$ is a normalized version of the relative hazard of breast cancer associated with risk factors Y where the normalization factor is the average relative hazard in the general population:

$$s(Y) = \phi(Y)(1 - AR) = \frac{\phi(y)}{\int \phi(Y)f(Y)dy} \tag{1.11}$$

where AR denotes the population attributable risk due to Y, $\phi(Y)$ is the relative hazard associated with Y (relative to the no-risk population), and $f(Y)$ is the prevalence of Y in the population. $s(Y)$ is approximated by

$$s(Y) \approx \prod_{j=1} \frac{\phi(Y_j)}{\int \phi(Y_j)f(Y_j)dy} \tag{1.12}$$

using the assumption that the risk factors are independent (j indexes the risk factors in Y).

Software for running IBIS is available at `http://www.ems-trials.org/riskevaluator/`. We used the command line program for version 8 and the competing mortality option.

1.2.3 Model Combination Approaches

Penetrance Modification: BRCAPRO+BCRAT (M)

Liu et al. (2013) proposed to combine BRCAPRO and BCRAT by incorporating the relative hazards for BCRAT covariates into the genotype-specific hazard functions in BRCAPRO. Since BCRAT is not recommended for known carriers of BRCA1/2 mutations (one of the patient eligibility criteria for using the risk calculator at `https://bcrisktool.cancer.gov/calculator.html` is the absence of a positive test result for BRCA1/2), we consider applying the relative hazards for BCRAT covariates to only the non-carrier hazard function in BRCAPRO.

We extend the BCRAT relative hazard model,

$$\tilde{\lambda}_B(t|X) = \tilde{\lambda}_{B,0}(t) r(t, X), \tag{1.13}$$

where $\tilde{\lambda}_B(t|X) = \lim_{dt \to 0} P(t \leq \tilde{T} < t + dt, J = B | \tilde{T} \geq t, X)/dt$, to obtain a model for non-carriers:

$$\tilde{\lambda}_B(t|X, G = 0) = \tilde{\lambda}^0_B(t) r^0(t, X), \tag{1.14}$$

where $\tilde{\lambda}_B(t|X, G = 0) = \lim_{dt \to 0} P(t \leq \tilde{T} < t + dt, J = B | \tilde{T} \geq t, X, G = 0)/dt$, $\tilde{\lambda}^0_B(t) = \lim_{dt \to 0} P(t \leq \tilde{T} < t + dt, J = B | \tilde{T} \geq t, G = 0)/dt$, and $r^0(t, X)$ is the relative hazard of breast cancer compared to the average hazard among non-carriers (discussed in more detail below). Models 1.13 and 1.14 are continuous-time models. To incorporate the hazard modification into the discrete-time framework used by BRCAPRO, we consider the discrete-time analogue induced by Equation 1.14 under the setting where we observe only integer-valued t (see Chapter 2.4.2 of Kalbfleisch and Prentice (2011)):

$$\lambda_B(t|X, G = 0) = 1 - \left(1 - \lambda^0_B(t)\right)^{r^0(t,X)}, \tag{1.15}$$

where $\lambda_B(t|X, G = 0) = P(T = t, J = B | T \geq t, X, G = 0)$. Note that, while BCRAT uses the continuous-time framework and BRCAPRO the discete-time framework, for white women, the estimated general population hazard in BCRAT, $\tilde{\lambda}_B(t) = \tilde{\lambda}_{B,0}(t)/(1 - AR(t))$, is similar to the estimated non-carrier hazard in BRCAPRO, $\lambda_B(t|X, G = 0)$ (Figure S.1.1 in Supplementary

Section S.1.1).

We then modify the calculation of the non-carrier risk in BRCAPRO by replacing $\lambda_B^0(t)$ in Equation 1.2 with $\lambda_B(t|X, G = 0)$ to get

$$P(T{\leq}a{+}\tau, J{=}B|T{>}a, G{=}0, X){=}\sum_{t=a+1}^{a+\tau}\Big(1{-}\big(1{-}\lambda_B^0(t)\big)^{r^0(t,X)}\Big)\prod_{u=a+1}^{t-1}\Big(\big(1{-}\lambda_B^0(u)\big)^{r^0(u,X)}{-}\lambda_D(u)\Big). \tag{1.16}$$

As in BRCAPRO, the final risk is a weighted average of the genotype-specific risks:

$$P(T{\leq}a{+}\tau, J{=}B|T{>}a, H, X){=}\sum_{g=0}^{3}P(T{\leq}a{+}\tau, J{=}B|T{>}a, G{=}g, X)P(G{=}g|H), \tag{1.17}$$

This combination approach assumes that the BCRAT risk factors modify the hazard of breast cancer in the same way for all non-carriers (conditional on age). It is similar to replacing the non-carrier future risk from BRCAPRO with the future risk from BCRAT (with some adjustments for the different baseline hazards used in BRCAPRO and BCRAT). The modification of the hazard function induces a modification of the corresponding penetrance function (see Equation 1.3), so we refer to the combination model as BRCAPRO+BCRAT (M), where "M" stands for (penetrance) modification.

The relative hazard approach for incorporating the BCRAT risk factors has similarities to the one used in IBIS, but IBIS averages the genotype-specific risks before incorporating non-genetic risk factors, while BRCAPRO+BCRAT (M) incorporates the BCRAT risk factors before averaging the genotype-specific risks. The advantage of the latter is that it allows for the effects of the BCRAT risk factors to differ by genotype. Differing effects by genotype have been found for some BCRAT risk factors, such as age at menarche (see Milne and Antoniou (2016) for a review). However, in general, the effects of the BCRAT risk factors on carriers are not well-studied (only a limited number of prospective studies have been done and they had small sample sizes (Milne and Antoniou, 2016)), so the current version of BRCAPRO+BCRAT (M) only modifies the non-carrier hazards.

Similar to $s(Y)$ from IBIS (Equation 1.11), $r^0(t, X)$ is a normalized version of $r(t, X)$ where the

normalization factor is the average relative hazard among non-carriers:

$$r^0(t, X) = r(t, X)(1 - AR^0(t)) = \frac{r(t, X)}{\sum_X r(t, X)P(X|t, G = 0)dy} \tag{1.18}$$

where $AR^0(t)$ is the population attributable risk fraction among non-carriers. The normalization is necessary because $r(t, X)$ modifies $\tilde{\lambda}_{B,0}(t)$ and is with respect to the no-risk population; in order to modify $\lambda_B^0(t)$, we need the relative hazard with respect to the non-carrier population.

Due to low mutation prevalence, we approximate $AR^0(t)$ with $AR(t)$, i.e. we assume $P(X|t, G = 0) \approx P(X|t)$. Therefore, BRCAPRO+BCRAT (M) takes parameters from existing models and does not need to be trained on new data (however, the parameters should be updated as new data becomes available). Though race-specific estimates of $AR(t)$ are available from BCRAT, they are based on data from the 1980s to early 2000s, so we re-estimated $AR(t)$ based on the distribution of BCRAT covariates in more recent data from the 2015 National Health Interview Survey (NHIS), which uses a cross-sectional sample of U.S. adults designed to be representative of the U.S. general population. As in IBIS (Equation 1.12), we assumed that the risk factors are independent, except we used the joint distribution of age at first live birth and number of affected first-degree relatives because Equation 1.5 includes an interaction between these variables. The race-specific estimates from the NHIS are given in Supplementary Section S.1.1.

Since $P(G = g|H)$ already accounts for family history, it may seem redundant to include the BCRAT family history variable (X_4, the number of affected first-degree relatives) among the penetrance-modifying risk factors. However, its inclusion could be useful because 1) there is a strong interaction between the family history variable and age at first live birth in BCRAT, and 2) the BCRAT family history variable could potentially account for residual familial risk due to non-BRCA-related factors, such as low-penetrance genes and shared environmental factors, which are not currently considered by BRCAPRO.

Logistic Regression Ensemble: BRCAPRO+BCRAT (E)

The second model combination approach involves training a logistic regression ensemble model that uses BRCAPRO and BCRAT as the base models. This approach is similar to stacking (Wolpert,

1992), which involves training two or more base models, typically using complementary information and/or approaches, and then training a meta-model to combine their predictions, but differs from typical stacking procedures because BRCAPRO and BCRAT are pre-trained. For the ensemble approach, we consider a binary outcome of whether or not a proband develops breast cancer within τ years, where τ is fixed.

Let $F_B(\tau) = P(T \leq a + \tau, J = B | T > a, H, X)$. Let $F_B^1(\tau)$ be the τ-year BRCAPRO risk prediction and $F_B^2(\tau)$ the τ-year BCRAT risk prediction. For a given value of τ, we combine the τ-year BRCAPRO and BCRAT predictions using the model

$$\log \frac{F_B(\tau)}{1 - F_B(\tau)} = \beta_0 + \beta_1 F_B^1(\tau) + \beta_2 F_B^2(\tau) + \beta_3 F_B^1(\tau) F_B^2(\tau). \tag{1.19}$$

Other covariates and/or models can also be added to the logistic regression.

This combination approach is based on the assumption that predictions from BRCAPRO and BCRAT capture complementary risk information. We refer to the ensemble model as BRCAPRO + BCRAT (E).

In contrast to the penetrance modification model, the ensemble model needs to be trained using prospective follow-up data. The ensemble model should ideally be trained on a dataset representative of the target population. When the training data are not representative of the target population, reweighting methods can be used to account for differences in the covariate distributions. One widely used method is importance weighting (Sugiyama et al., 2007), which weights each training observation by the ratio of the joint probability distributions of the covariates in the target and training populations (Sugiyama et al., 2007). The importance weights can be estimated using kernel mean matching (Huang et al., 2007), Kullback-Leibler importance estimation (Sugiyama et al., 2008), or least squares importance fitting (Kanamori et al., 2009).

1.2.4 Model Evaluation Metrics

In the simulations and data application, we consider the binary outcome of whether a proband develops breast cancer within $\tau = 5$ years.

We evaluate model performance using four measures (Steyerberg et al., 2010): 1) the ratio

of observed (O) to expected (E) events (where E is calculated by summing everyone's predicted probabilities), a measure of calibration (with 1 indicating perfect calibration); 2) the area under the receiver operating characteristic curve (AUC) or concordance (C) statistic, which is the probability that an individual who experiences the event has a higher score than an individual who does not and is a measure of discrimination; 3) the Brier score, which is the mean squared difference between the predicted probabilities and actual outcomes; and 4) standardized net benefit (SNB) (Kerr et al., 2016), which is the difference between the true positive rate and a weighted false positive rate, based on a pre-specified risk threshold (the weight is the ratio of the odds of the threshold risk to the odds of the outcome). For SNB, we use a 5-year risk threshold of 1.67% (the clinical risk threshold for eligibility for chemoprevention). We report the Brier score in terms of relative difference with respect to BRCAPRO since this metric is prevalence-dependent and therefore more difficult to interpret on its original scale.

Since there is censoring in the validation data for the data application (some individuals were followed for fewer than τ years and were breast-cancer free when they were lost to follow-up), we use inverse probability of censoring weights (IPCW) (Uno et al., 2007; Gerds and Schumacher, 2006) to calculate the O/E, AUC, and Brier score: individuals with observed outcomes are used to calculate the performance measures and are weighted by their inverse probability of not being censored by the minimum of 1) their age at the end of the projection period, and 2) the age at which they were diagnosed with breast cancer. Letting C denote censoring time, an individual is weighted by $1/P(C > a + \tau)$ if they did not develop cancer within τ years and $1/P(C > T_B)$ otherwise. Individuals who are censored are not directly used to calculate the performance measures, but are used to estimate the censoring distribution, $P(C > t)$. We estimate the censoring distribution by fitting a Kaplan-Meier curve.

1.3 Simulations

We compared the performance of the combination models to BRCAPRO and BCRAT in data simulated under the assumptions of the penetrance modification model.

1.3.1 Data Generation

We first generated each proband's baseline family history, consisting of 1) the family structure, 2) dates of birth, 3) genotypes, and 4) cancer ages and death ages.

We simulated pedigrees to mimic family structures observed in real families from the CGN database (described in Section 1.4.1), including the number of sisters, number of brothers, and so on. We restricted the family members to first- and second-degree relatives of the proband.

For probands, dates of birth and baseline dates for risk assessment were also sampled from the CGN database. For non-probands, dates of birth were generated relative to the proband's date of birth by assuming that the age difference between a parent and a child has mean 27 and standard deviation 6. We generated the birth dates of the proband's parents and children based on the proband's birth date, then the birth dates of the proband's grandparents and siblings based on the birth dates of the parents, then the birth dates of the proband's aunts and uncles based on the birth dates of the grandmothers.

Next, we generated the BRCA genotypes for each family member. We first generated the genotypes of the grandparents using the default Ashkenazi Jewish allele frequencies for BRCA1 and BRCA2 in BRCAPRO to mimic a higher-risk population (CGN participants represent a higher-risk population than the general population since they were selected for family history of cancer). For individuals in subsequent generations, we generated genotypes according to Mendelian inheritance.

For all individuals, we generated baseline breast and ovarian cancer phenotypes conditional on genotype. Ages of onset were sampled from $\{1, 2, \ldots, \text{current age}\}$, with probabilities given by the genotype-specific penetrance functions from BRCAPRO, and the probability of being unaffected at baseline given by one minus the cumulative penetrance up to the current age. Probands were assumed to be alive at baseline, but we generated a death age for each non-proband from a distribution with mean 80 and standard deviation of 15. If an individual had cancer with an age of onset greater than their age at death, then the individual's cancer status was changed to unaffected. Probands with breast cancer at baseline were excluded from the analyses.

We then generated baseline BCRAT covariates (excluding number of affected first-degree relatives

and age at first live birth, which were calculated from the baseline family history), from the CGN, sampling different covariates independently of each other. The BCRAT covariates were used to transform the BRCAPRO non-carrier penetrance into the BRCAPRO+BCRAT (M) non-carrier penetrance.

For probands who did not have breast cancer at baseline, future ages of onset were generated from the BRCAPRO+BCRAT (M) penetrances (for carriers, the BRCAPRO+BCRAT (M) penetrances are the same as in BRCAPRO), which were rescaled to be conditional on not having developed cancer by the baseline age. Cases were defined as probands who developed breast cancer within 5 years of their baseline age.

After excluding 4,443 probands who had breast cancer at baseline, we used 50,000 probands to train the ensemble model and the remaining 45,557 for validation. There were 751 cases in the training set and 717 cases in the validation set.

1.3.2 Results

BRCAPRO+BRCAT (M), the true model, had the best performance (Table 1.1). The ensemble model performed nearly as well as the true model and outperformed BRCAPRO and BCRAT. Both combination models were well-calibrated, with O/E=1.00 (95% CI 0.93-1.07) for BRCAPRO+BCRAT (M) and O/E=1.05 (95% CI 0.98-1.12) for BRCAPRO+BCRAT (E), while BRCAPRO and BCRAT underpredicted the number of cases, with O/E=1.14 (95% CI 1.07-1.22) for BRCAPRO and O/E=1.13 (95% CI 1.05-1.20) for BCRAT. The combination models had slightly higher AUCs than BRCAPRO and BCRAT: 0.70 (95% CI 0.68-0.72 for BRCAPRO+BCRAT (M), 0.69 (95% CI 0.67-0.71) for BRCAPRO+BCRAT (E), 0.68 (95% CI 0.66-0.70) for BCRAT, and 0.67 (95% CI 0.66-0.69) for BRCAPRO. Also, the combination models performed better than BRCAPRO and BCRAT with respect to the Brier score and SNB. Across bootstrap 1000 replicates of the validation dataset, both combination models outperformed BRCAPRO and BCRAT with respect to all performance measures in more than 96% of the replicates. BRCAPRO+BCRAT(M) outperformed BRCAPRO+BCRAT(E) more than 99% of the time with respect to AUC and Brier score, 75% of the time with respect to calibration, and 59% of the time with respect to SNB.

With a training set of 50,000, the ensemble model was able to achieve similar performance to the true model and performance gains over BRCAPRO and BCRAT.

Table 1.1: 5-year performance in a simulated dataset with 45,557 probands (717 cases). B+B: BRCAPRO+BCRAT. ΔBS: % relative improvement in Brier Score compared to BRCAPRO. The "Comparisons Across Bootstrap Replicates" section shows pairwise comparisons involving the combination models across 1000 bootstrap replicates of the validation dataset; the row for $A > B$ shows the proportion of bootstrap replicates where model A outperformed model B with respect to each metric.

	O/E	AUC	SNB	ΔBS
Performance Metrics				
B+B (M)	1.00 (0.93, 1.07)	0.70 (0.68, 0.72)	0.26 (0.21, 0.30)	0.36 (0.19, 0.52)
B+B (E)	1.05 (0.98, 1.12)	0.69 (0.67, 0.71)	0.25 (0.21, 0.30)	0.17 (0.03, 0.31)
BRCAPRO	1.14 (1.07, 1.22)	0.67 (0.66, 0.69)	0.22 (0.18, 0.26)	0.00 (0.00, 0.00)
BCRAT	1.13 (1.05, 1.20)	0.68 (0.66, 0.70)	0.21 (0.17, 0.26)	-0.06 (-0.35, 0.22)
Comparisons Across Bootstrap Replicates				
B+B(M)>B+B(E)	0.746	0.997	0.585	0.999
B+B(M)>BRCAPRO	0.973	1.000	0.989	1.000
B+B(M)>BCRAT	0.961	1.000	0.999	1.000
B+B(E)>BRCAPRO	0.991	1.000	0.998	0.990
B+B(E)>BCRAT	0.990	0.996	0.997	0.986

1.4 Data Application

We trained an ensemble model for 5-year risk of breast cancer using data from the Newton-Wellesley Hospital (NWH) and validated it, along with BRCAPRO+BCRAT (M), BRCAPRO, BCRAT, and IBIS, on data from the CGN.

In the analyses, we excluded women with invasive breast cancer/ductal carcinoma in situ/lobular carcinoma in situ/bilateral mastectomy/bilateral oophorectomy prior to baseline, women who tested positive for BRCA1/2 prior to baseline (BCRAT requirement), women < 20 years old at baseline (BCRAT requirement), and women with projection age > 85 years old (IBIS requirement).

1.4.1 Datasets

The characteristics of the training and validation datasets are summarized in Table 1.2.

NWH

After applying the exclusion criteria, the training cohort consisted of 37,881 women who visited the breast imaging department of the NWH in Newton, Massachusetts for screening or diagnostic imaging from February 2007 through December 2009. During the initial (baseline) visit, information was collected on personal and family history of cancer, reproductive history, sociodemographic factors, and lifestyle factors. Family history was limited to relatives with cancer. Breast cancer diagnoses through 2015 were determined from the Massachusetts State Cancer Registry, Partners Hospital Cancer Registries, and patient self-reporting. The median age of the probands was 49, with an inter-quartile range (IQR) of 43-58. 30,758 (81.2%) of the probands were white. 5,684 (15.0%) had at least one affected first- or second-degree relative. The median follow-up time was 6.7 years (IQR 6.3-7.2), and 5-year outcomes were observed for all probands. 495 probands were diagnosed with breast cancer within 5 years of baseline.

Since the NWH cohort represents a general screening population while the CGN validation cohort (described below) represents a higher-risk population enriched for family history of cancer, we applied importance weights to the training data based on the distributions of the BCRAT covariates, 5-year BCRAT predictions, and 5-year BRCAPRO predictions. To estimate the weights, we used least squares importance fitting. Since the distributions of the BRCAPRO and BCRAT predictions were highly right-skewed, we used log-transformed versions of predictions in the logistic regression model.

CGN

The validation cohort consists of 7,314 women who enrolled in the CGN, a national research network consisting of 15 academic medical centers that was established for the purpose of studying inherited predisposition to cancer. Enrollment began in 1999 and ended in 2010. One of the criteria for enrollment was a personal and/or family history of cancer. Participants provided information on personal and family history of cancer, sociodemographic factors and lifestyle factors through an initial (baseline) phone interview and annual follow-up updates. From 2009 onward, information was also collected on reproductive history, cancer treatments, cancer screening results, and genetic

testing results.

The median age of the probands was 47 (IQR 38-57); 6,104 (83.5%) of the probands were white. 3,143 (42.9%) had at least one female first-degree relative with breast cancer. The median follow-up time was 7.3 years (IQR 6.0-8.3) and 934 (12.8%) probands were censored within 5 years without being diagnosed with breast cancer. 112 probands developed breast cancer during the first 5 years of follow-up. Demographic characteristics stratified by center are given in Supplementary Section S.1.3. Since follow-up times and breast cancer incidence rates varied by center, we estimated the censoring distribution separately for each center.

Information on some risk factors was missing or incomplete. We did not have information on atypical hyperplasia (used in BCRAT and IBIS), breast density (used in IBIS), polygenic risk scores (used in IBIS, or hormone replacement therapy (used in IBIS). Participants were asked whether they had ever had a benign breast biopsy but were not asked about the number of biopsies (categorized as 0, 1, or ≥ 2 in BCRAT). Since participants were asked about reproductive history starting in 2009, 4,157 (56.8%) were missing age at menarche (used in BCRAT and IBIS). Ashkenazi Jewish status (used in BRCAPRO and IBIS) was not available for the UWASH center. We coded the missing variables according to the specifications of the software for each model. Number of breast biopsies was coded as 1 for participants who indicated that they had previously had a biopsy.

Table 1.2: NWH and CGN cohort characteristics.

Variable	Category	NWH	CGN
N		37881	7314
Age (median [IQR])		49 [43, 58]	47 [38, 57]
Race (%)	White	30758 (81.2)	6104 (83.5)
	Black	479 (1.3)	257 (3.5)
	Hispanic	548 (1.4)	694 (9.5)
	Asian	1228 (3.2)	160 (2.2)
	Native American	25 (0.1)	29 (0.4)
	Unknown	4843 (12.8)	70 (1.0)
Affected 1st-degree Relatives (%)	0	32197 (85.0)	4171 (57.0)
	1	5277 (13.9)	2496 (34.1)
	2+	407 (1.1)	647 (8.8)
Follow-up Time (median [IQR])		6.7 [6.3, 7.2]	7.3 [6.0, 8.3]
Censored in <5 Years (%)		0 (0.0)	934 (12.8)
5-year Cases (%)		495 (1.3)	112 (1.5)

1.4.2 Results

The performance measures are shown in Table 1.3. Calibration plots are provided in Figure 1.2 and scatter plots, density plots, and correlations are provided in Figure 1.3. The weights from the ensemble model are provided in Supplementary Section S.1.2. BRCAPRO+BCRAT (M) (O/E=1.06, 95% CI 0.9-1.29), IBIS (O/E=1.01, 95% CI 0.86-1.124), and BCRAT (O/E=1.18, 95% CI 1-1.44) were well-calibrated overall while BRCAPRO+BCRAT (E) (O/E=1.24, 95% CI 1.05-1.51) and BRCAPRO (O/E=1.34, 95% CI 1.14-1.64) underestimated risk. BRCAPRO+BCRAT (M), BCRAT, and IBIS overestimated risk in the top decile of risk (Figure 1.2). The AUCs were 0.68 (95% CI 0.64-0.72) for BRCAPRO+BCRAT (M), 0.68 (95% CI 0.64-0.72) for BRCAPRO+BCRAT (E), 0.67 (95% CI 0.63-0.71) for BCRAT, 0.67 (95% CI 0.63-0.71) for IBIS, and 0.65 (95% CI 0.6-0.69) for BRCAPRO. IBIS had the highest SNB (SNB=0.3, 95% CI 0.18-0.4), followed by BRCAPRO+BCRAT (M) (SNB=0.25, 95% CI 0.16-0.35). All models performed similarly with respect to the Brier score. Predictions from BRCAPRO+BCRAT (M) were highly correlated with predictions from each of the other models in the entire cohort (Figure 1.3), with Pearson correlation coefficients ranging from $\rho = 0.77$ with BRCAPRO to $\rho = 0.91$ with BRCAPRO+BCRAT (E). BRCAPRO+BCRAT (E), which assigned a higher weight to BCRAT than to BRCAPRO (see Supplementary Section S.1.2), was very highly correlated with BCRAT ($\rho = 0.95$) and moderately correlated with BRCAPRO ($\rho = 0.66$).

In probands who met the NCCN criteria for further genetic risk evaluation (Table 1.3), BRCAPRO+BCRAT (M) and IBIS had the highest AUCs and SNBs, but both models overestimated risk. BRCAPRO and BCRAT had the lowest AUCs and SNBs. In probands who did not meet the NCCN criteria, all models underestimated risk. BRCAPRO had a slightly lower AUC than the other models and BRCAPRO+BCRAT (M) and IBIS had slightly higher SNBs than the other models.

To quantify the improvement of the combination models over the other models, we also looked at pairwise performance comparisons with respect to AUC, Brier score, and SNB across 1000 bootstrap replicates of the CGN cohort (Table 1.3). In the overall cohort, both combination models outperformed BRCAPRO and BCRAT in the majority of the replicates. The combination

models also outperformed IBIS with respect to AUC and Brier score, but IBIS almost always had a better SNB. BRCAPRO+BCRAT (M) outperformed BRCAPRO+BCRAT (E) with respect to AUC and SNB, while the two models performed similarly with respect to the Brier score. In probands meeting the NCCN criteria, BRCAPRO+BCRAT (M) outperformed each of the other models in the majority of the replicates.

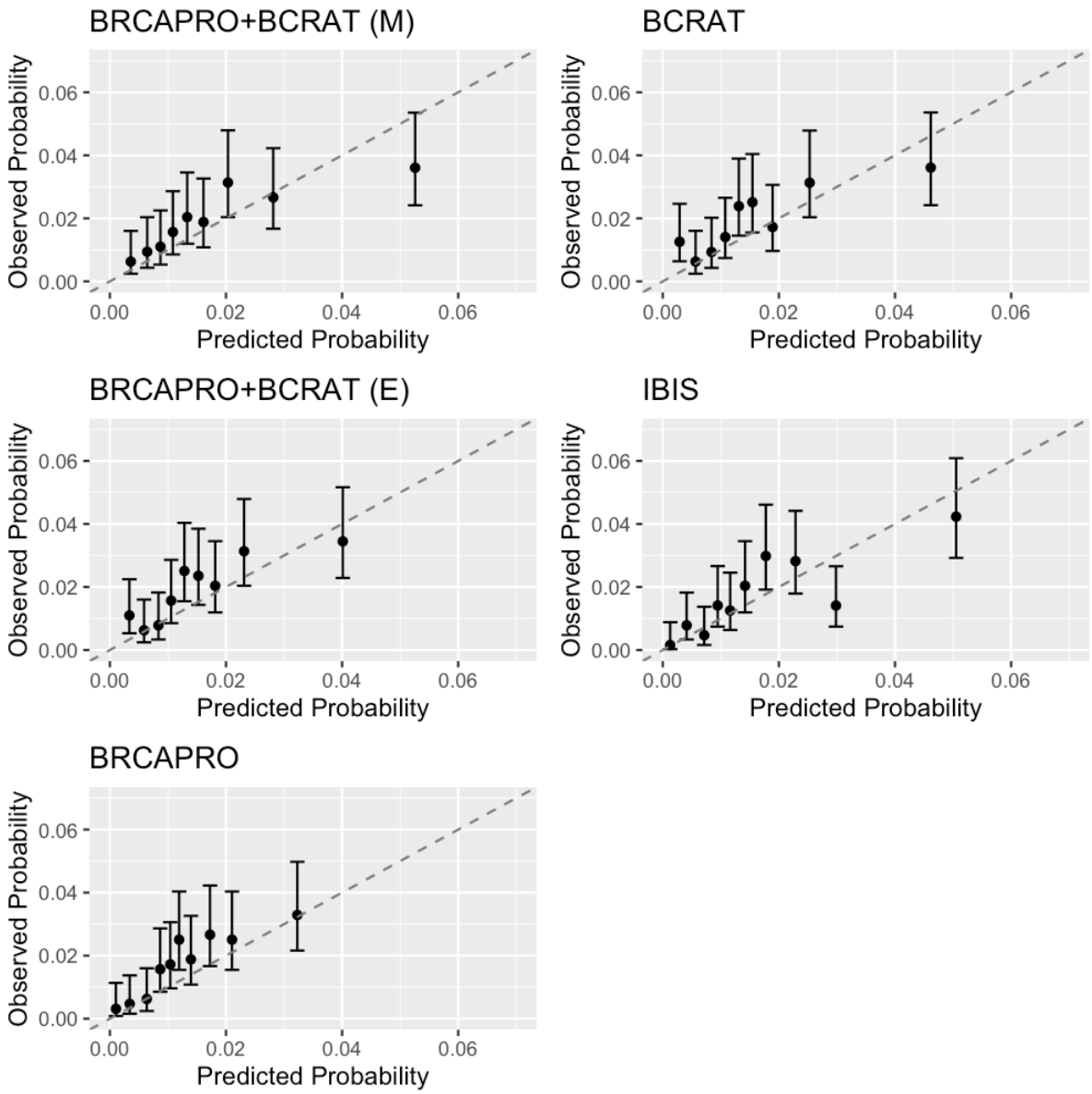

Figure 1.2: Calibration plots by deciles of predicted risk.

Table 1.3: 5-year performance in the CGN cohort, overall and stratified by family history (whether or not the proband met the NCCN criteria for further genetic risk evaluation National Comprehensive Cancer Network (2019); in applying the criteria, we only used information on breast and ovarian cancer diagnoses in relatives). B+B: BRCAPRO+BCRAT. ΔBS: % relative improvement in Brier Score compared to BRCAPRO. The "Comparisons Across Bootstrap Replicates" section shows pairwise comparisons involving the combintion models across 1000 bootstrap replicates of the validation dataset; the row for $A > B$ shows the proportion of bootstrap replicates where model A outperformed model B with respect to each metric.

	O/E	AUC	SNB	ΔBS
Overall (112 cases)				
Performance Metrics				
B+B (M)	1.06 (0.90, 1.29)	0.68 (0.64, 0.72)	0.25 (0.16, 0.35)	0.14 (-0.41, 0.65)
B+B (E)	1.24 (1.05, 1.51)	0.68 (0.64, 0.72)	0.21 (0.11, 0.31)	0.29 (-0.31, 0.78)
BRCAPRO	1.34 (1.14, 1.64)	0.65 (0.60, 0.69)	0.17 (0.06, 0.27)	0.00 (0.00, 0.00)
BCRAT	1.18 (1.00, 1.44)	0.67 (0.63, 0.71)	0.19 (0.08, 0.30)	0.20 (-0.63, 0.86)
IBIS	1.01 (0.86, 1.24)	0.67 (0.63, 0.71)	0.30 (0.18, 0.40)	-0.15 (-0.68, 0.40)
Comparisons Across Bootstrap Replicates				
B+B(M)>B+B(E)	928	575	904	234
B+B(M)>BRCAPRO	969	933	956	667
B+B(M)>BCRAT	881	749	980	372
B+B(M)>IBIS	365	730	171	844
B+B(E)>BRCAPRO	996	911	774	866
B+B(E)>BCRAT	24	831	701	739
B+B(E)>IBIS	112	634	52	922
Strong Family History (34 cases)				
Performance Metrics				
B+B (M)	0.83 (0.55, 1.14)	0.71 (0.63, 0.77)	0.45 (0.19, 0.62)	0.62 (-1.68, 1.8)
B+B (E)	1.16 (0.77, 1.6)	0.68 (0.59, 0.76)	0.41 (0.13, 0.55)	0.77 (-0.93, 1.87)
BRCAPRO	1.36 (0.89, 1.87)	0.66 (0.57, 0.74)	0.3 (0.04, 0.45)	0 (0, 0)
BCRAT	1.06 (0.7, 1.46)	0.66 (0.56, 0.75)	0.32 (0.07, 0.48)	0.78 (-1.43, 2.16)
IBIS	0.76 (0.5, 1.04)	0.69 (0.61, 0.76)	0.43 (0.12, 0.57)	-0.41 (-3.14, 1.25)
Comparisons Across Bootstrap Replicates				
B+B(M)>B+B(E)	472	899	714	364
B+B(M)>BRCAPRO	677	947	935	741
B+B(M)>BCRAT	348	903	953	373
B+B(M)>IBIS	927	709	714	909
B+B(E)>BRCAPRO	883	673	866	831
B+B(E)>BCRAT	304	855	855	516
B+B(E)>IBIS	613	412	442	882
Less Family History (78 cases)				
Performance Metrics				
B+B (M)	1.2 (0.97, 1.47)	0.66 (0.61, 0.71)	0.17 (0.04, 0.28)	-0.05 (-0.39, 0.34)
B+B (E)	1.27 (1.03, 1.56)	0.66 (0.61, 0.72)	0.12 (-0.02, 0.24)	0.09 (-0.2, 0.37)
BRCAPRO	1.32 (1.06, 1.61)	0.63 (0.59, 0.69)	0.11 (0, 0.24)	0 (0, 0)
BCRAT	1.23 (1, 1.51)	0.67 (0.62, 0.72)	0.14 (0, 0.26)	-0.03 (-0.41, 0.33)
IBIS	1.17 (0.95, 1.43)	0.66 (0.61, 0.71)	0.24 (0.11, 0.36)	-0.03 (-0.42, 0.3)
Comparisons Across Bootstrap Replicates				
B+B(M)>B+B(E)	952	346	910	196
B+B(M)>BRCAPRO	967	850	819	387
B+B(M)>BCRAT	939	251	900	411
B+B(M)>IBIS	90	398	111	428
B+B(E)>BRCAPRO	983	922	572	742
B+B(E)>BCRAT	36	320	312	967
B+B(E)>IBIS	51	472	33	763

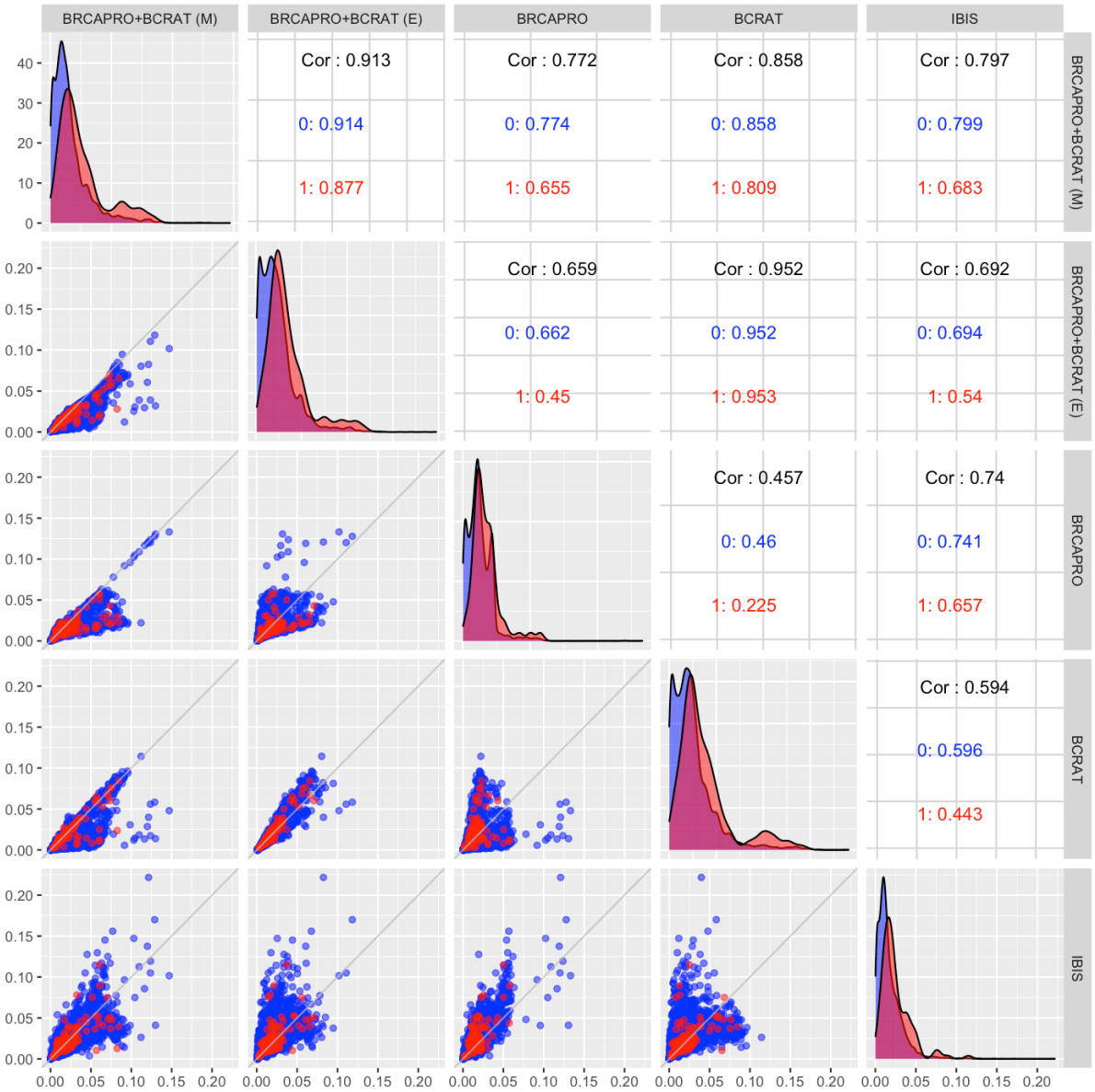

Figure 1.3: Scatter plots, density plots, and correlations (stratified by outcome). Red corresponds to cases and blue corresponds to non-cases. LR: logistic regression. TC: IBIS.

1.5 Discussion

The availability and use of multiple risk prediction models can lead to confusion in clinical practice. Combining models addresses this problem and also provides a way to develop more comprehensive models without building new models from the ground up. We validated two approaches for combining BRCAPRO, a Mendelian model based on detailed family history information, with BCRAT, an empirical model based on a simple summary of family history and various non-genetic risk factors. We compared the performance of the combination models to BRCAPRO, BCRAT, and IBIS, a family history-based model that also accounts for non-genetic risk factors. Another family history-based model, BOADICEA, was recently extended to incorporate non-genetic risk factors, but the software for the updated version (`https://canrisk.org/`) was not publicly available for research use when we performed our analyses. The penetrance modification model, BRCAPRO+BCRAT (M), achieved comparable performance to IBIS in the CGN cohort, outperforming BRCAPRO overall and outperforming BCRAT among women with a strong family history of breast/ovarian cancer. The ensemble model, BRCAPRO+BCRAT (E), showed similar discrimination to the penetrance modification model but had worse calibration and lower net benefit overall. These results suggest that there is value in combining BRCAPRO and BCRAT, though the CIs for the performance measures were overlapping and there was a relatively small number of cases in the subset with a strong family history. Further validation in independent prospective studies is needed.

While BCRAT performed well in the entire CGN cohort, it had lower discriminatory accuracy and net benefit than IBIS and the combination models in probands with a strong family history. This suggests that detailed family history information is important for achieving reliable risk stratification among higher-risk subgroups for which early screening and prevention measures can substantially reduce cancer risk and mortality (Metcalfe et al., 2008). Furthermore, BCRAT is not suitable for known BRCA1/2 carriers, while the other four models all take into account genetic testing results (the ensemble model does so indirectly through the BRCAPRO risk prediction).

Missing information on BRCAT and IBIS risk factors in the CGN dataset (atypical hyperplasia, age at menarche, and, for IBIS, mammographic density, hormone replacement therapy, and polygenic risk scores) could potentially have affected the discrimination of BCRAT, IBIS, and

the combination models, but the models still had relatively good discrimination. The CGN also did not collect genetic testing information for non-probands. Genetic testing information could considerably improve the discrimination of BRCAPRO, IBIS, and the combination models (Gail, 2019).

One limitation regarding the ensemble model applied to the CGN cohort is that the training data from NWH was not representative of the validation data from the CGN. The NWH cohort was a lower-risk cohort (as seen in Table 1.4.1–it had a lower proportion of women with a first-degree family history of breast cancer) and the family history information available for the NWH cohort was less detailed than that for the CGN cohort. We used importance weighting to address this limitation, but since this approach relies on accurate estimation of the probability distributions of risk factors in the training and target populations, the ensemble model we trained may have been sub-optimal for CGN participants. The performance of the ensemble approach could potentially be improved by training on data that is more representative of the validation data. In simulations, where the training and validation datasets were both generated under BRCAPRO+BCRAT (M), the ensemble model performed as well as BRCAPRO+BCRAT (M).

The two combination approaches each have their strengths and limitations. Ensembling via logistic regression calibrates the model to the training data, which can be a strength or a limitation depending on how well the training data represents the target population. The model might require recalibration in order to be suitable for a population different from the training population. Differences between the training and validation populations could also negatively affect discrimination and other performance measures. The penetrance modification approach, on the other hand, does not require training, but relies on accurate estimates of prevalence, penetrance, and relative risks. These estimates should be updated as new information becomes available. Another disadvantage of ensembling via logistic regression is the need to dichotomize the time-to-event outcome. An ensemble model for variable-year risk prediction would require additional assumptions on how the ensemble weights vary over time and/or prospective training data with long-term follow-up. In contrast, the penetrance modification approach uses a survival model that automatically handles different risk prediction periods. One advantage of ensembling is its greater flexibility compared to the penetrance modification approach. Ensembling can easily handle any number of models

that can be of any form. Including additional risk factors is also straightforward. The penetrance modification model requires more assumptions because it specifically combines a penetrance-based model with a relative risk model via proportional hazards. Additional risk modifiers can be incorporated as new relative risk estimates become available, but it is important to properly scale the relative risks to be compatible with the hazard functions they are meant to modify and to consider whether the effects of the risk modifiers differ by carrier status. BRCAPRO+BCRAT (M) currently modifies only the non-carrier hazard function using the BCRAT relative risk. Future work is needed to extend the model to include modifiers of the carrier hazard functions. One more advantage of ensembling is that once the model is trained, it only requires the predictions from the models being combined and not the raw inputs (besides any additional risk factors that are explicitly included in the ensemble model), which are potentially less accessible than the predictions.

Given our validation results, the penetrance modification approach seems more promising than ensembling in the context of breast cancer risk prediction. BRCAPRO+BCRAT (M) achieved competitive performance by leveraging the strengths of BRCAPRO and BCRAT, improving on aspects of both models.

2

Prediction of Hereditary Breast Cancer Using Neural Networks

2.1 Introduction

Family history is a major risk factor for many types of cancer, including breast, colorectal, and pancreatic cancer. Various family history-based cancer risk prediction models have been developed (Berry et al., 1997; Chen et al., 2006; Wang et al., 2010, 2007) and are used in clinical practice to guide decisions about screening and interventions. Existing models are primarily based on two approaches: 1) using Mendelian laws of inheritance to translate detailed family history information into risk predictions (Berry et al., 1997; Tyrer et al., 2004; Antoniou et al., 2004; Wang et al., 2007, 2010) and 2) using summaries of family history (for example, the number of relatives with a previous cancer diagnosis) as covariates in regression models (Gail et al., 1989; Balmaña et al., 2006).

Mendelian models take as input a pedigree (Figure 2.1) that reflects family history (including relatives' cancer diagnoses, ages at cancer onset, and current ages) and genealogical information. They estimate an individual's probability of carrying a mutation in a cancer susceptibility gene using Mendelian laws of inheritance, Bayes' Rule, and estimates of mutation prevalence and penetrance (probability of disease given genotype) from epidemiological literature (for example, see Chen and Parmigiani (2007)). The individual risk of cancer is then calculated as a weighted average of mutation carrier and non-carrier risks of developing cancer. Mendelian models are typically recommended over regression-based models for individuals with a strong family history of cancer, since Mendelian models use more detailed family history information (Quante et al., 2012; Pichert et al., 2003). However, they rely on explicit assumptions about cancer susceptibility genes, some of which may be unrealistic or restrictive. Known susceptibility genes account for a limited proportion of familial risk (Easton, 1999). Existing Mendelian models consider only a small subset of

these genes and do not incorporate information on shared environmental risk factors. Furthermore, Mendelian models are sensitive to misreporting of family history (Braun et al., 2014; Katki, 2006) and rely on accurate estimation of mutation prevalence and penetrance, which is challenging due to low mutation prevalence, heterogeneity of prevalence across populations, and the presence of genetic variants of unknown clinical significance.

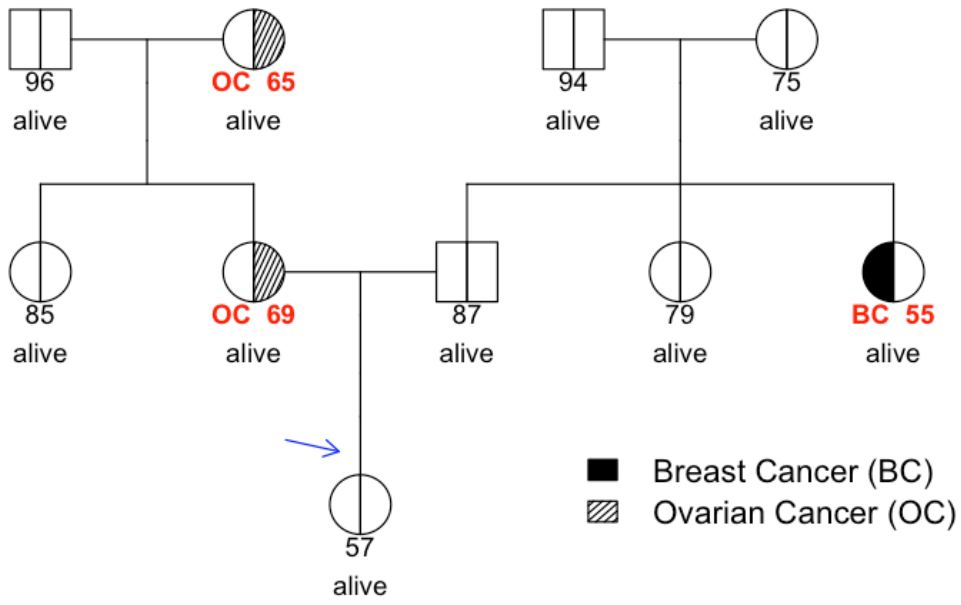

Figure 2.1: Example of a pedigree with family history of breast and ovarian cancers. Circles represent females and squares represent males. The arrow indicates the proband, the individual undergoing risk assessment. Numbers below each family member represent the individual's current age if alive and unaffected, age at death if dead, and age of diagnosis if affected by breast or ovarian cancer.

The main limitations of Mendelian models can be overcome by neural networks (NNs) that eliminate the need to explicitly specify the effects of cancer susceptibility genes. A NN consists of layers of nodes that convert inputs into predictions through a series of non-linear transformations. Under mild assumptions, NNs are theoretically capable of approximating any continuous function with arbitrary precision (Cybenko, 1989; Hornik, 1991; Leshno et al., 1993), and in practice they have achieved state-of-the-art performance in many tasks, such as image recognition (Krizhevsky

et al., 2012) and natural language processing (Hinton et al., 2012). The flexibility of NNs combined with large databases can potentially lead to accuracy gains over Mendelian models. However, while the literature on NNs is extensive, little work has been done to evaluate their performance in the context of family history-based cancer risk prediction. Kokuer et al. (2006) trained a NN to classify families into risk categories for hereditary colorectal cancer, but they used simple summaries of family history and cross-validated their model on a relatively small dataset with 313 pedigrees. To the best of our knowledge, there is no previous work leveraging large databases of pedigrees to develop NNs for cancer risk prediction.

In this paper, we develop new NN models to predict future risk of breast cancer based on pedigree data. We propose a method for mapping pedigrees to fixed-size NN inputs and apply two types of NNs : 1) standard fully-connected NNs (FCNNs), and 2) convolutional NNs (CNNs) that exploit pedigree structure. We compare the performance of the NNs to BRCAPRO (Parmigiani et al., 1998), a widely used Mendelian model, and logistic regression (LR). In our data application, we train NNs using over 200,000 families from the Risk Service database and validate the models on data from the Cancer Genetics Network (CGN). Although we focus on breast cancer risk prediction in our simulations and data application, the proposed approach can also be applied to other cancers.

2.2 Methods

2.2.1 Notation

Notation used throughout the paper is summarized in Table S.2.1 in Supplementary Section S.2.1. For a given type of cancer, consider a proband (someone who presents for risk assessment) who has not previously been diagnosed with that cancer. Let t be a pre-specified number of years. Let $Y_0 = 1$ if the proband develops the cancer of interest within t years and $Y_0 = 0$ otherwise. The goal is to estimate $P(Y_0 = 1|H)$, where H represents family history and is described below.

Family history can be visualized using a pedigree (Figure 2.1), a directed graph where nodes correspond to family members and edges flow from parents to offspring. The pedigree graph can be represented as a matrix where each row corresponds to a family member, containing their features

as well as the indices of their parents. Let H denote this matrix.

Let R be the number of relatives in the pedigree besides the proband. The family members are indexed by $r = 0, 1, \ldots, R$, where $r = 0$ corresponds to the proband. We have K features for each family member r: $H_{r1}, \ldots, H_{rK}$. In this paper, we will consider the following $K = 6$ features for breast cancer risk prediction: H_{r1} = current age or age at death, H_{r2} = breast cancer status (1 if affected, 0 otherwise), H_{r3} = ovarian cancer status (1 if affected, 0 otherwise), H_{r4} = age at onset of breast cancer (0 if unaffected), H_{r5} = age at onset of ovarian cancer (0 if unaffected), and H_{r6} = sex (0 if female, 1 if male). Furthermore, let A_{r1} be the index of r's father and A_{r2} the index of r's mother (either of which can be unknown). Let $H_r = (H_{r1}, \ldots, H_{rK}, A_{r1}, A_{r2}) \in \mathbb{R}^{K+2}$. H is a matrix with $R + 1$ rows and $k + 2$ columns, where for $r = 0, \ldots, R$, row $r + 1$ contains the information for family member r.

2.2.2 Fully-Connected Neural Networks

A NN (Bishop et al., 1995) is a directed graph consisting of an input layer, hidden layers, and an output layer. The nodes perform computations that transform input features into a prediction or classification. Each node receives a set of inputs via incoming edges, computes a function of its inputs, and propagates the result via outgoing edges.

A FCNN is a NN where in each layer, every node is connected to every node in the previous layer. FCNNs take as input a fixed-length vector X. In the context of cancer risk prediction, we propose a FCNN where X is a vector representation of the pedigree H and the output is a predicted probability for $Y_0 = 1$. We describe how H is mapped to X in Section 2.2.3.

Let L be the number of hidden layers in the FCNN. Let $l = 0$ and $l = L + 1$ correspond to the input and output layers respectively. Let N_l be the number of nodes in layer l, where N_0 is the length of X and $N_{L+1} = 1$. Each hidden layer node takes a weighted sum of its inputs, adds a bias term, and then applies a nonlinear activation function $\phi : \mathbb{R} \to \mathbb{R}$. Commonly used activation function include the logistic function $\phi(z) = \sigma(z) = \frac{\exp(z)}{1+\exp(z)}$ and the rectifier function $\phi(z) = ReLU(z) = \max(0, z)$.

The outputs of the layers are

$$a^0 = X \in \mathbb{R}^{N_0},$$
$$a^l = \phi^l(W^l a^{l-1} + b^l) \in \mathbb{R}^{N_l}, \qquad (l = 1, \ldots, L+1),$$

where $W^l \in \mathbb{R}^{N_l \times N_{l-1}}$ is the matrix of weights for layer l with row i consisting of the weights of node i, $b^l \in \mathbb{R}^{N_l}$ is the bias vector for layer l, and $\phi^l : \mathbb{R}^{N_l} \to \mathbb{R}^{N_l}$ represents the component-wise application of the activation function ϕ. The output layer ($l = L+1$) consists of a single node that uses the logistic activation function, outputting the predicted probability

$$\hat{Y}_0 = a^{L+1} = \sigma(w^{L+1} a^L + b^{L+1}).$$

Given a cost function C and N_T training observations (X_n, Y_{0n}), $n = 1, \ldots, N_T$, the weight and bias parameters are randomly initialized and iteratively updated to minimize $\sum_{n=1}^{N_T} C(Y_{0n}, \hat{Y}_{0n})$ using an optimization method such as stochastic gradient descent. Examples of cost functions include mean squared error, $C(y, z) = (y - z)^2$ and cross-entropy loss, $C(y, z) = -y \log(z) - (1 - y) \log(1 - z)$.

The number of parameters (W, b) in a FCNN grows quickly with the size of the input and the number and size of the hidden layers. Various regularization methods have been developed to avoid overfitting, including weight decay (Hinton, 1987; Krogh and Hertz, 1992) and dropout (Srivastava et al., 2014).

2.2.3 Standardizing and Flattening Pedigrees

Since FCNNs require a fixed-size input, they cannot be directly applied to pedigrees, which vary in size and structure. It is possible to generate a fixed-size input based on simple summaries of family history, but this can result in substantial loss of information. Therefore, we propose the following approach: define a reference pedigree with pre-specified relatives (for example: proband, grandparents, parents, sister, brother) and map each actual pedigree H to a standardized version H' that matches the structure of the reference pedigree (each relative in the reference pedigree may

or may not be present in the actual pedigree), then flatten H' into a fixed-length vector input X for a FCNN.

We first describe the reference pedigree (see Figure 2.2(B) for an example of a reference structure; the choice of reference structure can vary depending on the family structures in the available data). Let the reference pedigree contain the proband and Q' other types of relatives (father, mother, brother, sister, etc). Let $q = 0, 1, \ldots, Q'$ index the relative types, with $q = 0$ corresponding to the proband. Let R'_q be the number of relatives of type q for $q \in \{0, 1, \ldots, Q'\}$. Let the family members be indexed by $r = 0, 1, \ldots, R'$, where $R' = \sum_{q=1}^{Q'} R'_q$, $r = 0$ corresponds to the proband, $r = 1, \ldots, R_1$ corresponds to relatives of type $q = 1$, $r = R_1+1, \ldots, R_1+R_2$ corresponds to relatives of type $q = 2$, and so on.

Now we consider an actual pedigree matrix H and describe how to standardize and flatten it (Figure 2.2). For $q = 0, 1, \ldots, Q'$, R_q is the number of relatives of type q in H ($R_0 = 1$). To construct a standardized pedigree matrix, H', with the same structure as the reference pedigree matrix, we compare the number of relatives of type q in the actual pedigree to the number in the reference pedigree for each $q \in \{0, 1, \ldots, Q'\}$. If the two numbers are the same ($R_q = R'_q$), then we include all of the R'_q actual relatives in H'. If the actual number is smaller than the reference number ($R_q < R'_q$), then we include the R_q actual relatives in H' and represent each of the $R'_q - R_q$ absent relatives using a vector of pre-specified null values (zeros). If the actual number is larger than the reference number ($R_q > R'_q$), then we randomly select R'_q of the actual relatives to include in H'. We also include a column in H' to indicate whether each row corresponds to a relative who is absent from the actual pedigree (0 if present, 1 if absent). Therefore, H' is an $R' + 1$ by $K + 1$ matrix where each row consists of a family member's K cancer history features, along with the presence/absence indicator. Let H'_r be the vector for relative r in H'. We flatten H' by concatenating its rows to get a vector, $X = ({H'_0}^T, {H'_1}^T, \ldots, {H'_{R'}}^T) \in \mathbb{R}^{(R'+1)(K+1)}$, which can be used as input to a FCNN.

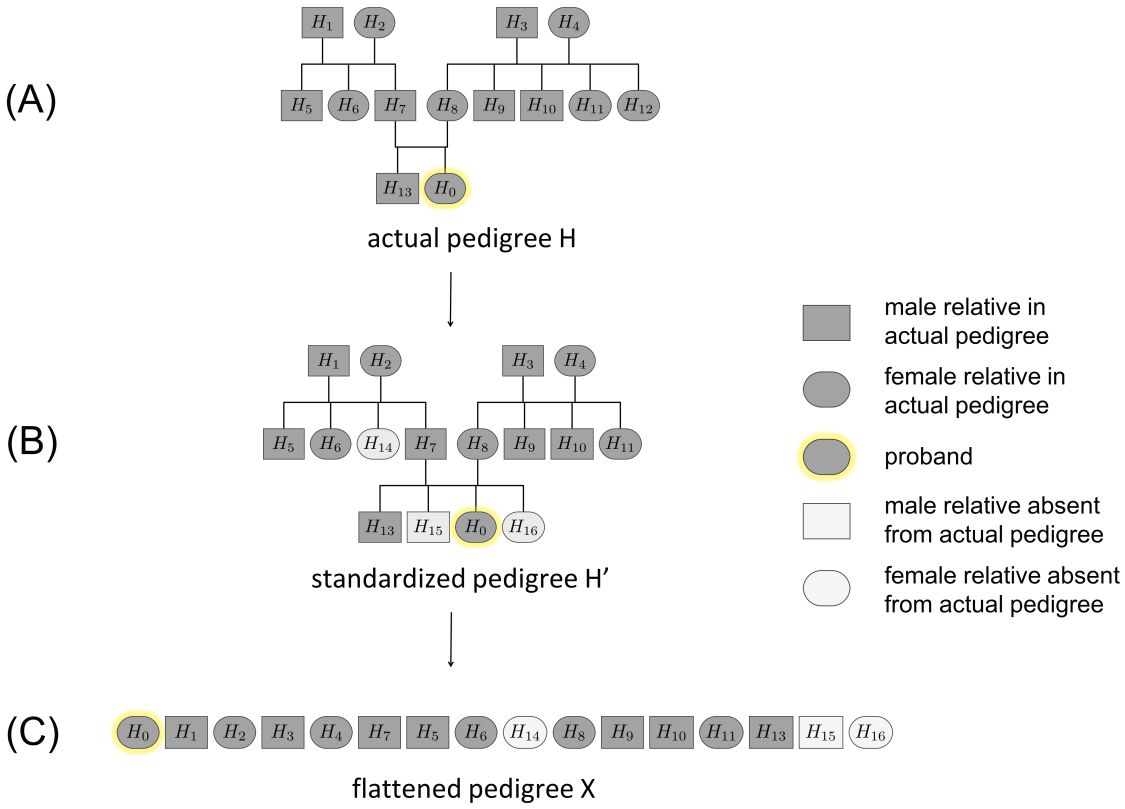

Figure 2.2: Consider a reference pedigree that includes the proband's grandparents, parents, uncles, aunts, and siblings, with each couple having $m = 2$ children of each sex. (A) Actual pedigree H. (B) Standardized pedigree H', obtained by mapping H to the reference structure. The actual pedigree has more maternal aunts than the reference pedigree, so we randomly select the desired number of maternal aunts to include in H'. The actual pedigree has fewer paternal aunts, brothers, and sisters than the reference pedigree, so in H', we use pre-specified non-informative values for the paternal aunts, brothers, and sisters absent from H. (C) Flattened pedigree X, which is used as input for a FCNN.

2.2.4 Convolutional Neural Networks

FCNNs can be inefficient to train and are prone to overfitting because the number of parameters grows quickly with network size (Geman et al., 1992). CNNs (LeCun et al., 1998), which are widely used in problems where the input has a spatial structure, such as image classification, reduce the number of parameters by using convolutional layers that enforce selective connections and weight sharing. A convolutional layer can be viewed as a fully-connected layer where certain weights are set to 0 and certain weights are constrained to have the same value. To exploit the correlation structure of the input (for example, pixels that are spatially close often have highly correlated values), a convolutional layer applies the same functions repeatedly to different fixed-

size neighborhoods of the input (for example, sets of neighboring pixels). These functions are called convolutional filters.

Analogous to neighboring pixels, closely related individuals are likely to have similar levels of susceptibility to cancer due to genetic similarity and shared environment. Therefore, we propose to adapt CNNs to pedigree data. For reference, a description of a standard CNN is provided in Supplementary Section S.2.1. While standard CNNs were designed for inputs that have a fixed size and structure, various generalizations have been proposed for graphs that vary in size and structure (Wu et al., 2019), such as molecular compounds. We follow the general steps used in the graph CNN framework proposed by Niepert et al. (2016): 1) standardize the graphs to have the same size and structure, then 2) define a sequence of neighborhoods within each standardized graph and apply convolutional filters to those neighborhoods. However, while Niepert et al.'s algorithm, which is based on graph isomorphism tests, is applicable to arbitrary graphs, our approach leverages the structure of pedigrees.

Like in the FCNN approach, we use a standardized and flattened pedigree X as the input (Figure 2.2). Prior to running the CNN, for each family member r in H', we define a fixed-size neighborhood centered at r consisting of r and r's first-degree relatives: mother, father, m_1 sisters, m_2 brothers, m_3 daughters, and m_4 sons. Similar to Figure 2.2, if r has more than m_1 sisters, then m_1 of them are randomly selected, and if r has fewer than m_1 sisters, then we use a pre-specified index representing an absent relative whose features are set to zero (analogous to zero padding in standard CNNs, as described in Supplementary Section S.2.1). The same approach is used for brothers, daughters, and sons. The neighborhood is represented by a vector $\mathcal{N}(r)$ of length $U = 3 + \sum_{i=1}^{4} m_i$. Within $\mathcal{N}(r)$, the individuals are ordered by relative type with respect to r.

We propose a CNN where all of the hidden layers are convolutional. There are L hidden layers. Hidden layer l applies M_l real-valued convolutional filters $f_1^l, \ldots, f_{M_l}^l$ to each of the $R'+1$ neighborhoods of the pedigree (Figure 2.3). For $i = 1, \ldots, M_l$, let $f_i^l : \mathbb{R}^{U(M_{l-1})} \to \mathbb{R}$ (let $M_0 = K + 1$ since each relative has $K + 1$ features in H' - see Section 2.2.3). Let $a_r^l \in \mathbb{R}^{M_l}$ be the output of layer l for neighborhood/family member r. Let $a_{\mathcal{N}(r)}^{l-1} \in \mathbb{R}^{U*M_{l-1}}$ be the vector obtained by concatenating

the layer inputs of the relatives in $\mathcal{N}(r)$. The output from applying filter i to r's neighborhood is

$$f_i^l(a_{\mathcal{N}(r)}^{l-1}) = \phi\left(w_i^l a_{\mathcal{N}(r)}^{l-1} + b_i^l\right),$$

where $w_i^l \in \mathbb{R}^{U*M_{l-1}}$ is the vector of weights for filter i and $b_i^l \in \mathbb{R}$ is the bias for filter i.

Let $f^l = (f_1^l, \ldots, f_{M_l}^l) : \mathbb{R}^{U*M_{l-1}} \to \mathbb{R}^{M_l}$. The layer outputs for relative r are

$$\begin{aligned} a_r^0 &= H_r' \in \mathbb{R}^{K+1}, \\ a_r^l &= f^l(a_{\mathcal{N}(r)}^{l-1}) \in \mathbb{R}^{M_l} \qquad (l = 1, \ldots, L) \end{aligned}$$

and the overall layer outputs are

$$a^l = \left[a_0^l, \ldots, a_{R'}^l\right] \in \mathbb{R}^{M_l*(R'+1)} \qquad (l = 0, 1, \ldots, L)$$

The final output is a transformation of a_0^L using a logistic activation function:

$$\hat{Y}_0 = \sigma(w^{L+1} a_0^L + b^{L+1})$$

where $w^{L+1} \in \mathbb{R}^{M_L}$ and $b^{L+1} \in \mathbb{R}$.

As in FCNNs, the weight and bias parameters are optimized with respect to $\sum_{n=1}^{N_T} C(Y_{0n}, \hat{Y}_{0n})$ and the optimization can be carried out using stochastic gradient descent.

Model Space

The proposed CNN for pedigrees satisfies a universal approximation property analogous to that of FCNNs (Cybenko, 1989; Hornik, 1991; Leshno et al., 1993). Fix a reference pedigree H^* of size $R' + 1$ containing relatives of up to degree d of the proband. Let Q' be the number of relative types in H^* besides the proband and let $m = \max_{q=0,1,\ldots,Q'} R_q'$. Let $\mathcal{X}^* \subset \mathbb{R}^{(R'+1)(K+1)}$ be the space of standardized and flattened pedigrees with the same structure as H^*. We consider the CNN's ability to approximate functions from $\mathcal{X}^*$ to $[0, 1]$. We first state the universal approximation theorem for standard FCNNs (Cybenko, 1989; Hornik, 1991; Leshno et al., 1993) and then verify that the same

Figure 2.3: The neighborhoods centered at relatives 0, 2, and 7 are shown above using shaded boxes. The same convolutional filters are applied to all neighborhoods of the pedigree.

property extends to CNNs (proof provided in Supplementary Section S.2.1).

Universal Approximation Theorem for FCNN (forward direction of Theorem 1 from (Leshno et al., 1993)) Let k be a positive integer and I a compact subset of $\mathbb{R}^k$. Let $g : I \to \mathbb{R}$ be continuous. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a piecewise continuous, locally bounded, and non-polynomial activation function. Then given $\epsilon > 0$, there exists a positive integer N, and for $i = 1, \ldots, N$, constants $\alpha_i, b_i \in \mathbb{R}$ and vectors $w_i \in \mathbb{R}^k$ such that

$$F(X) = \sum_{i=1}^{N} \alpha_i \phi(w_i^T X + b_i),$$

satisfies $|F(X) - g(X)| < \epsilon \ \forall X \in I$.

Theorem 2.1. *[Universal Approximation Theorem for Pedigree CNNs] Assume that the elements of $H'_r \in \mathbb{R}^{K+1}$ are bounded for $r = 0, 1, \ldots, R'$. Let $g : \mathcal{X}^* \to [0, 1]$ be continuous. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a continuous and invertible activation function. Let the fixed-size neighborhood about each relative contain $m_1 = \cdots = m_4 = m$ sisters/brothers/daughters/sons. Then given $\epsilon > 0$, there exists a pedigree CNN of the form described in subsection 2.2.4 with d hidden layers with activation function ϕ, M_l convolutional filters for hidden layer l, bias terms $b_i^l \in \mathbb{R}$ $(i = 1, \ldots, M_l;\ l = 1, \ldots, L + 1)$, and weight vectors $w_i^l \in \mathbb{R}^{U * M_{l-1}}$ $(i = 1, \ldots, M_l;\ l = 1, \ldots, L + 1)$, such that the final output*

$$F(X) = \sigma(w^{L+1} a_0^L(X) + b^{L+1})$$

satisfies $|F(X) - g(X)| < \epsilon \ \forall X \in \mathcal{X}^*$.

2.2.5 Benchmark Methods

In our simulations and data application, we focused on breast cancer risk prediction and compared NNs to the Mendelian BRCAPRO model and to LR, which is equivalent to a single-node FCNN with a logistic activation function. For LR, we used the flattened pedigree X as the input.

BRCAPRO (Berry et al., 1997; Parmigiani et al., 1998) is widely used in clinical practice and has been validated in various populations (Berry et al., 2002; Euhus et al., 2002; Terry et al., 2019; McCarthy et al., 2019). It estimates the probability of carrying a germline mutation in breast/ovarian cancer susceptibility genes BRCA1 and BRCA2, as well as future risk of breast/ovarian cancer.

Let γ_r be the genotype of relative r (0 for non-carrier of a pathogenic BRCA1 or BRCA2 mutation, 1 for BRCA1 carrier, 2 for BRCA2 carrier, and 3 for BRCA1 and BRCA2 carrier). BRCAPRO first estimates $P(\gamma_0 | H_0, \ldots, H_R)$, the conditional distribution of proband's genotype given the family history, using Bayes' rule:

$$P(\gamma_0 | H_0, \ldots, H_R) = \frac{P(\gamma_0) P(H_0, \ldots, H_R | \gamma_0)}{\sum_{\gamma=0}^{3} P(\gamma) P(H_0, \ldots, H_R | \gamma)}.$$

The prevalence $P(\gamma)$ is obtained from the literature and, under the assumption of conditional

independence of phenotypes given genotypes,

$$P(H_0, \ldots, H_R|\gamma_0) = \sum_{\gamma_1,\ldots,\gamma_R} \left(\prod_{r=0}^{R} P(H_r|\gamma_r) \right) P(\gamma_1, \ldots, \gamma_R|\gamma_0),$$

where the penetrance $P(H_r|\gamma_R)$ is obtained from the literature. The conditional distribution $P(\gamma_1, \ldots, \gamma_R|\gamma_0)$ is derived assuming Mendelian inheritance of BRCA1 and BRCA2 mutations.

After estimating the carrier probabilities, BRCAPRO calculates future risk of breast cancer through a weighted average of the genotype-specific penetrance functions $P(Y_0 = 1|\gamma_0)$:

$$P(Y_0 = 1|H) = \sum_{\gamma_0} P(Y_0 = 1|\gamma_0) P(\gamma_0|H_0, \ldots, H_R).$$

2.2.6 Implementation

We ran BRCAPRO using the BayesMendel R package (version 2.1-6) (Chen et al., 2004a). The NNs were implemented in Python using Keras (`https://github.com/keras-team/keras`) with the Theano backend (Team et al., 2016). For the CNNs, we used code from Hechtlinger et al. (2017) (`https://github.com/hechtlinger/graph_cnn`).

In the simulations and data application, 10% of the training set was held out for tuning NN hyperparameters. In the simulations, the final FCNN architecture consisted of 2 hidden layers with sizes 30 and 10 respectively, and the final CNN architecture consisted of 2 convolutional layers with 10 and 5 filters respectively. In the data application, the final FCNN architecture consisted of 2 hidden layers of size 30 and the final CNN architecture consisted of 2 convolutional layers with 5 filters each. We also used a dropout layer following the first hidden layer in each NN, with a dropout rate of 20%. We used the Exponential Linear Unit (ELU) activation function (Clevert et al., 2015), the Adam optimizer (Kingma and Ba, 2014), and the mean squared error loss function.

In the simulations, we used a reference pedigree of size 26 containing the proband's grandparents, parents, aunts (2 maternal, 3 paternal), uncles (3 maternal, 2 paternal), siblings (2 sisters, 3 brothers), and children (2 daughters, 2 sons). This was chosen based on the distribution of family structures in the CGN (see Supplementary Section S.2.2). We used $m_1 = m_2 = 3$ and $m_3 = m_4 = 2$ for the CNN neighborhoods. In the data application, we used a reference pedigree of size 19 with

the same relative types as in the simulations, but restricted to 2 relatives of each type and omitted sons and daughters due to the smaller family sizes in the training dataset (see Supplementary Section S.2.2). We used $m_1 = m_2 = 2$ and $m_3 = m_4 = 1$ for the CNN neighborhoods.

2.2.7 Model Evaluation

We evaluated model performance using three metrics (Steyerberg et al., 2010): 1) the ratio of observed (O) to expected (E) events (where E is the sum of the predicted probabilities for the validation set), a measure of calibration, 2) the area under the receiver operating characteristic curve (AUC), which is the probability that an individual who experiences the event has a higher score than an individual who does not and is a measure of discrimination, and 3) the Brier score, which is the mean squared difference between the predicted probabilities and actual outcomes. We obtained 95% confidence intervals for the performance measures by bootstrapping the validation set.

2.3 Simulations

We evaluated the performance of the proposed NN approaches in predicting 10-year risk of breast cancer in two simulation settings: one where BRCAPRO is the true model and one where the data do not satisfy the assumptions of BRCAPRO.

2.3.1 Data Generation

We simulated 1,000,000 pedigrees using the generating model assumed by BRCAPRO. To simulate each family, we first sampled a family structure (number of sisters, brothers, etc) from the CGN dataset (described in Section 2.4.1). For probands, we also sampled dates of birth and baseline dates for risk assessment from the CGN. For non-probands, dates of birth were generated relative to the proband's date of birth by assuming that the age difference between a parent and a child has mean 27 and standard deviation 6.

Next, we generated the genotypes for each family member. We first generated the genotypes of the proband's grandparents (the oldest generation) using the default Ashkenazi Jewish allele frequencies in BRCAPRO (0.014 for BRCA1 and 0.012 for BRCA2) to mimic a higher-risk population. For individuals in subsequent generations, we generated genotypes according to Mendelian inheritance.

We generated ages of onset for breast and ovarian cancer conditional on the genotypes. Each age of onset was randomly generated from $\{1, \ldots, 94\}$, with probabilities given by the genotype-specific penetrance functions from BRCAPRO (the cumulative lifetime probability of breast cancer ranges from 0.12 for non-carriers to 0.79 for carriers of mutations in both BRCA1 and BRCA2). We also generated a death age for each individual from a distribution with mean 80 and standard deviation 15. If an individual's age of onset was greater than their baseline age or death age, then their cancer status at baseline was set to 0.

We excluded probands who died or were diagnosed with breast cancer prior to baseline. For the remaining probands ($n = 887,353$), we predicted 10-year risk of breast cancer using the baseline family history.

2.3.2 Results

We used 800,000 families for training and the other 87,353 for testing. In the training set, 23,606 probands developed breast cancer within 10 years, while in the test set, 2570 probands developed breast cancer within 10 years.

We investigated how much training data is needed for the performance of the NNs to approach that of the true model by training NNs on increasingly large subsets of the entire training set, with sample sizes ranging from 50,000 to 800,000 (Figure 2.4). As the sample size increased, the NNs achieved higher AUCs and their predictions became highly correlated with those from BRCAPRO, the true model in this setting. For sample sizes under 100,000, the CNN had a higher AUC than the FCNN, though, as expected, the differences between the two approaches decreased with increasing sample size. With 200,000 or more training examples, both the FCNN and CNN achieved AUCs similar to that of BRCAPRO. For all sample sizes, the correlation between the CNN predictions

and BRCAPRO predictions was higher than that between the FCNN predictions and BRCAPRO predictions.

When the entire training set was used, the FCNN and CNN had correlations of 0.9 and 0.92 with BRCAPRO, while the LR model trained on the data had a correlation of 0.82 with BRCAPRO (Table 2.1). Across 1000 bootstrap replicates of the test set, the NNs outperformed LR with respect to AUC and Brier score more than 99% of time.

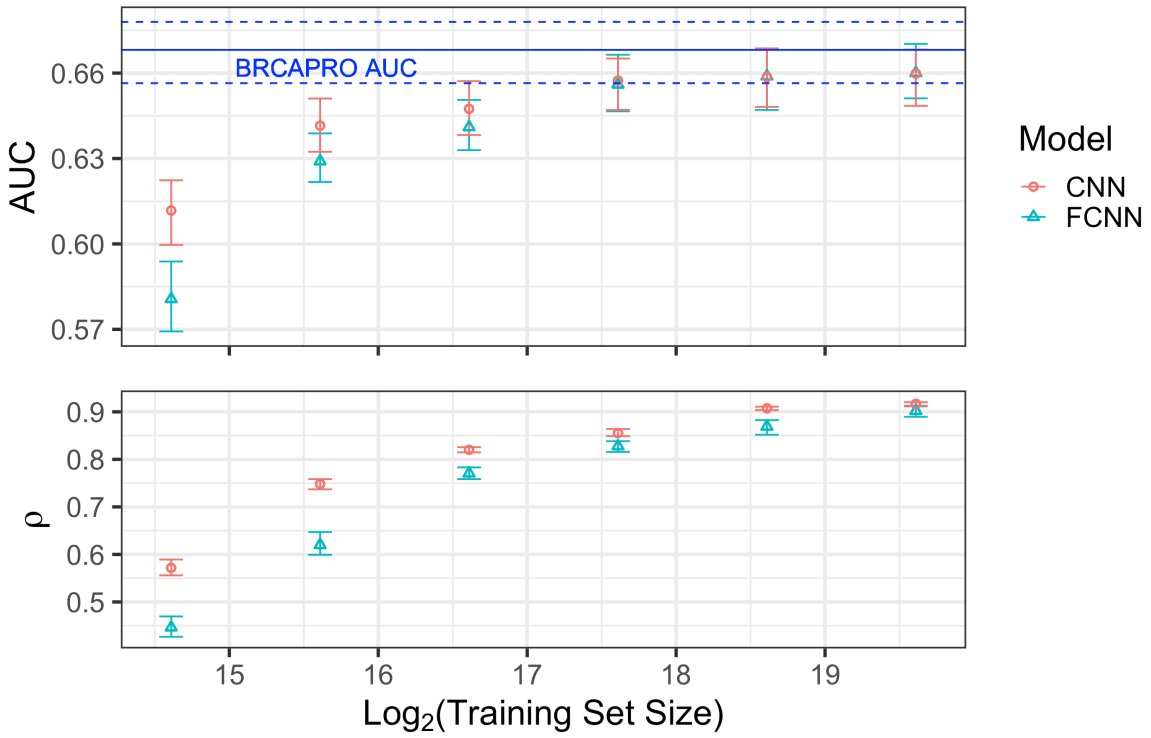

Figure 2.4: AUC and correlation (ρ) of 10-year risk predictions from NN with predictions from BRCAPRO (true model) as a function of training sample size in simulations.

Under the true model, the proband's risk of breast cancer increases with more affected relatives and earlier diagnosis ages. To assess whether NN and LR predictions reflected these trends, we fixed a family structure and varied the phenotypes of the mother and maternal grandmother (Figure 2.5). We considered five scenarios that are ordered by increasing risk with respect to the true model: (A) no affected relatives, (B) grandmother with breast cancer, (C) grandmother with breast cancer at an earlier age, (D) grandmother with breast cancer, mother with breast cancer, and (E) grandmother with breast cancer, mother with breast and ovarian cancer. While the NNs gave similar predictions to BRCAPRO across all scenarios (Figure 2.5), LR slightly underpredicted risk in Scenario (D)

Table 2.1: 10-year performance in 87,353 simulated families (training set of 800,000) based on true and misreported family history. ρ denotes the correlation with predictions from BRCAPRO. The "Comparisons Across Bootstrap Replicates" section shows pairwise comparisons between the NN models and the other models across 1000 bootstrap replicates of the test set; the row for $A > B$ shows the proportion of bootstrap replicates where model A outperformed model B with respect to each metric.

	O/E	AUC	sqrt(Brier Score)	ρ
True Family History				
Absolute Performance				
FCNN	0.93 (0.89, 0.96)	0.66 (0.65, 0.67)	0.168 (0.165, 0.171)	0.90 (0.89, 0.91)
CNN	0.99 (0.95, 1.03)	0.66 (0.65, 0.67)	0.168 (0.165, 0.171)	0.92 (0.91, 0.92)
BRCAPRO	1.02 (0.98, 1.06)	0.67 (0.66, 0.68)	0.168 (0.164, 0.171)	1.00 (1.00, 1.00)
LR	1.00 (0.96, 1.04)	0.65 (0.64, 0.66)	0.168 (0.165, 0.171)	0.82 (0.81, 0.83)
Comparisons Across Bootstrap Replicates				
FCNN>CNN	0.021	0.582	0.038	0.000
FCNN>BRCAPRO	0.083	0.000	0.000	0.000
FCNN>LR	0.025	1.000	0.991	1.000
CNN>BRCAPRO	0.691	0.000	0.000	0.000
CNN>LR	0.465	1.000	1.000	1.000
Misreported Family History				
Absolute Performance				
FCNN	1.06 (1.03, 1.10)	0.64 (0.64, 0.65)	0.168 (0.165, 0.171)	
CNN	1.01 (0.97, 1.05)	0.64 (0.63, 0.65)	0.168 (0.165, 0.172)	
BRCAPRO	0.81 (0.78, 0.84)	0.63 (0.62, 0.64)	0.169 (0.166, 0.172)	
LR	1.00 (0.96, 1.03)	0.64 (0.63, 0.65)	0.168 (0.165, 0.171)	
Comparisons Across Bootstrap Replicates				
FCNN>CNN	0.033	0.665	0.038	
FCNN>BRCAPRO	1.000	1.000	1.000	
FCNN>LR	0.061	0.968	0.486	
CNN>BRCAPRO	1.000	1.000	1.000	
CNN>LR	0.406	0.900	0.996	

and severely underpredicted risk in Scenario (E). LR assumes a restrictive functional form for the relationship between the features and the outcome, and this functional form does not match that of BRCAPRO, so the LR model is misspecified in these simulations. NNs with multiple hidden nodes are more flexible than LR and therefore less susceptible to misspecification.

Perturbations of Mendelian Models Misreported cancer diagnoses can considerably distort predictions from Mendelian models (Katki, 2006; Braun et al., 2014). In the second simulation setting, we introduced noise to the simulated family histories through incorrectly reported diagnoses, diagnosis ages, and current ages for non-probands. For diagnoses, we used false negative and false positive rates for self-reported family history of breast and ovarian cancer from Ziogas and Anton-Culver (2003). For ages, we used rates from Braun et al. (2017) (see Supplementary Section S.2.2 for more details).

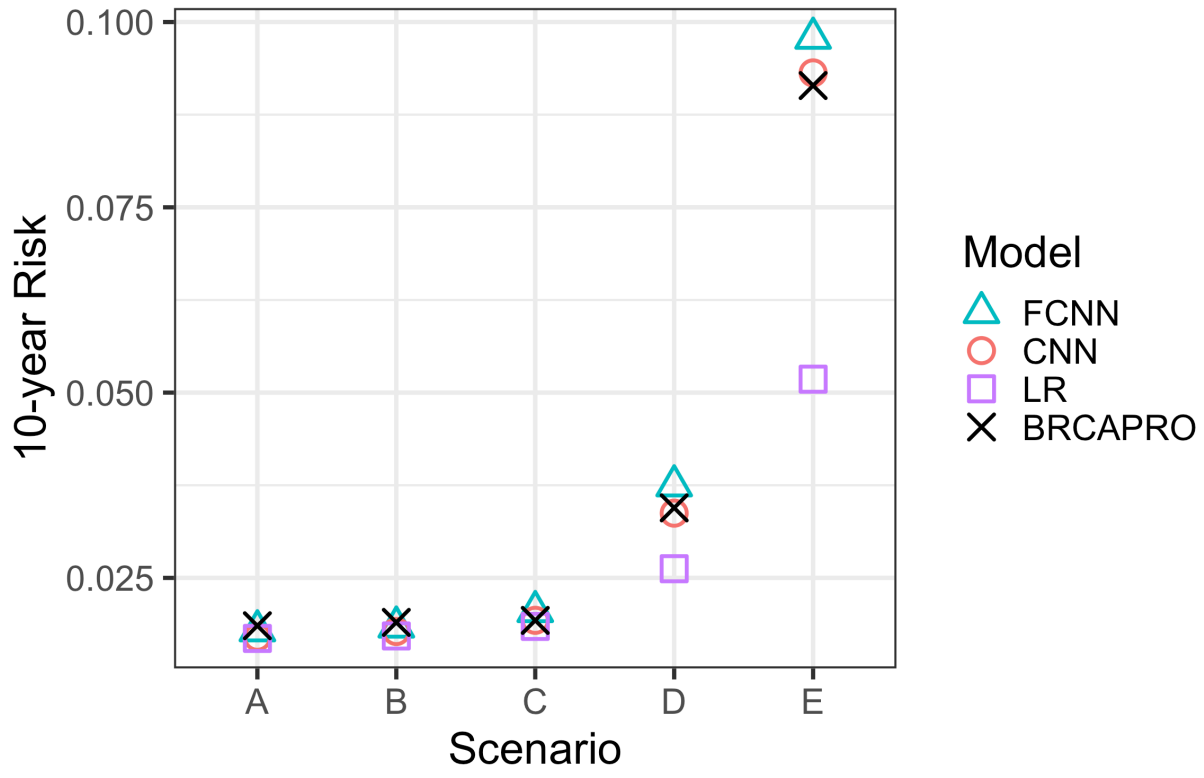

Figure 2.5: Under the first simulation settting, we fixed a simulated family structure (proband, grandparents, parents, 1 paternal aunt, 1 maternal aunt, 2 maternal uncles) for a 40-year-old proband and varied the level of family history across 5 scenarios (ordered by increasing risk with respect to BRCAPRO, the true model):

1. no affected relatives
2. maternal grandmother diagnosed with breast cancer at age 80
3. maternal grandmother diagnosed with breast cancer at age 60
4. maternal grandmother diagnosed with breast cancer at age 60, mother diagnosed with breast cancer at age 50
5. maternal grandmother diagnosed with breast cancer at age 60, mother diagnosed with breast cancer at age 50 and ovarian cancer at age 60. We calculated 10-year risk predictions for each scenario using each model

Under misreporting, BRCAPRO overestimated risk and had a worse AUC and Brier score than the NNs across 1000 bootstrap replicates of the test set (Table 2.1), showing that NNs are able to outperform BRCAPRO when Mendelian assumptions are not fully satisfied.

2.4 Data Application

We trained FCNN, CNN, and LR models on the Risk Service cohort to predict 5-year risk of breast cancer and compared their performance to BRCAPRO on an independent dataset from the CGN. We excluded male probands, probands who had breast cancer/bilateral mastectomy/bilateral oophorectomy before baseline, probands under 18 years old, and probands for whom we could not run BRCAPRO (probands over 89 years old at baseline and probands with related parents in their pedigree).

2.4.1 Datasets

Risk Service

The Risk Service (Chipman et al., 2013) is a web service for family history-based cancer risk prediction. It has been used in a wide range of clinics, including primary care, breast imaging, and genetic counseling clinics, to run various family history-based cancer risk prediction models, including BRCAPRO. Patient-reported model inputs are saved to a database. As of January 2018, the database contained over 450,000 families, with 285,161 probands consenting to the use of their data for research.

Training a future risk prediction model requires baseline and follow-up data, but the Risk Service does not follow probands over time. We therefore defined each proband's baseline date to be 5 years prior to the date at which they used the Risk Service and the follow-up date to be the date at which they used the Risk Service. We retrospectively reconstructed the family history at the baseline date based on the ages and diagnosis ages of the family members. However due to a considerable amount of missing age information for non-probands (74% of first- and second-degree relatives were missing age and 34% of affected first- and second-degree relatives were missing age at diagnosis), we decided not to use ages or diagnosis ages of non-probands for training and we imputed baseline cancer status for non-probands with missing diagnosis ages (see Supplementary Section S.2.2 for more details).

The training set consisted of 279,460 probands (Table 2.2). The median age was 45 and the median family size was 8. Also, 36,783 probands (13.2%) had at least one affected first-degree

relative and 13,307 (4.8%) developed breast cancer during the follow-up period.

CGN

The CGN is a national consortium of 15 academic medical centers that was established for the purpose of studying inherited predisposition to cancer (Anton-Culver et al., 2003). Between 1999 and 2010, 26,941 participants with cancer or a family history of cancer were recruited through population-based registries, high-risk clinics, and self-referral. They provided information on personal and family history of cancer, sociodemographic factors and lifestyle factors through an initial (baseline) phone interview and annual follow-up updates.

The validation cohort consisted of 7,489 probands. The median age was 47 and the median family size was 16. The majority (54.1%) of probands were recruited from population-based cancer registries. Also, 42.9% of probands had at least one female first-degree relative with breast cancer (a much higher proportion than in the Risk Service), 114 (1.5%) probands developed breast cancer within 5 years of baseline, and 1017 probands (13.6%) were lost to follow-up within 5 years without being diagnosed with breast cancer (Table 2.2). To adjust for censoring, we used inverse probability of censoring weights (Uno et al., 2007; Gerds and Schumacher, 2006): individuals with observed outcomes were used to calculate the performance measures and were weighted by their inverse probability of not being censored by the minimum of 1) the length of the risk prediction period (5 years) and 2) the time to breast cancer diagnosis. Censored individuals were not directly used to calculate the performance measures, but were used to estimate the censoring distribution. We assumed independent censoring and estimated the censoring distribution using the Kaplan-Meier estimator.

2.4.2 Training and Validation Populations

There are many differences between the Risk Service and CGN cohorts (Table 2.2). Since CGN participants were recruited based on family history of cancer, the CGN cohort represents a higher-risk population and has more probands with a positive family history (Table 2.2). Due to different data collection and ascertainment procedures, the family history information available in the CGN

Table 2.2: Risk Service and CGN cohort characteristics.

Variable	Category	Risk Service	CGN
N (probands)		279460	7489
Age (median [IQR])		45 [39, 55]	47 [38, 57]
Family Size (median [IQR])		8 [7, 14]	16 [12, 21]
Affected 1st-degree Relatives (%)	0	242677 (86.8)	4277 (57.1)
	1	35241 (12.6)	2549 (34.0)
	2+	1542 (0.6)	663 (8.9)
Ascertainment (%)	Population-Based	—	4050 (54.1)
	Clinic-Based	—	2187 (29.2)
	Self-Referral	—	1247 (16.7)
	Unknown	—	5 (0.1)
Censored (%)		0 (0.0)	1017 (13.6)
Cases (%)		13307 (4.8)	114 (1.5)

is more detailed than in the Risk Service. To handle the considerable amount of missing age information in the Risk Service data, we did not use current or diagnosis ages of non-proband relatives in the NN features and used only their breast and ovarian cancer affection statuses (we still used the proband's age).

Moreover, the Risk Service cohort is affected by selection bias because individuals who are diagnosed with breast cancer often seek genetic counseling shortly after diagnosis. Therefore, the breast cancer incidence rate among Risk Service probands during the defined follow-up period is higher than in the CGN cohort (Table 2.2) even though the CGN cohort represents a higher-risk population. To address this issue, we re-calibrated the models trained on the Risk Service to general population incidence rates adjusted for family history. We calculated age-specific 5-year risks based on 2012-2016 incidence rates in the general U.S. population from the Surveillance, Epidemiology, and End Results (SEER) program (Horner et al., 2009). We then modified the risk based on the number of affected first-degree relatives using relative risk estimates from on Hormonal Factors in Breast Cancer (2001). To re-calibrate each model, we used the Risk Service data to fit a linear regression with the family history-adjusted 5-year SEER risk as the outcome and the 5-year risk from the model as the predictor. Similar approaches have been used in regression calibration problems (Carroll, 2006). While BRCAPRO was not trained on the Risk Service and is calibrated to general population incidence rates, we also evaluated a re-calibrated version of BRCAPRO obtained via the SEER re-calibration approach.

2.4.3 Results

Overall, FCNN, CNN, LR, and BRCAPRO had similar AUCs in the CGN data (Table 2.3): 0.65 (95% CI 0.61-0.69) for BRCAPRO, 0.64 (95% CI 0.59-0.68) for the CNN, 0.63 (95% CI 0.58-0.68) for the FCNN, and 0.62 (95% CI 0.57-0.67) for LR. All models underpredicted risk except for LR, which was well-calibrated.

Among probands recruited from population-based registries, the NNs and LR had slightly higher AUCs than BRCAPRO. The CNN had the highest AUC (0.69, 95% CI 0.62-0.74), outperforming BRCAPRO with respect to AUC in 97% of bootstrap replicates. Among probands recruited from high-risk clinics, all models underpredicted risk and BRCAPRO had a slightly higher AUC (0.62, 95% CI 0.52-0.7) than the other models.

2.5 Discussion

The main contributions of our paper are 1) adapting FCNNs and CNNs to family history data and 2) investigating the potential of NNs for learning genetic susceptibility to cancer from family history data. Our simulations and data application show that NNs are a promising approach for developing new risk prediction models.

In simulations under the assumptions of BRCAPRO, we examined how much training data is required for NNs to achieve comparable performance to BRCAPRO. The FCNNs and CNNs trained on 200,000 or more families were highly correlated with BRCAPRO and had AUCs similar to that of BRCAPRO. With training set sizes under 200,000, the CNN performed better than the FCNN, showing that leveraging pedigree structure via convolutions can lead to more efficient training. Moreover, the NNs had higher accuracy than LR and were able to recognize rare patterns that are strongly indicative of hereditary cancer, such as the presence of multiple affected individuals on the same side of the family and the presence of multiple cancers in the same individual. In the setting where family history was subject to misreporting, the NNs outperformed BRCAPRO.

In our data application, we trained NNs on over 200,000 families from the Risk Service database and validated the models on families from the CGN. In the CGN, the NNs achieved competitive

Table 2.3: 5-year performance in the CGN cohort, overall and stratified by ascertainment mode. BRCAPRO^C: Re-calibrated version of BRCAPRO. The "Comparisons Across Bootstrap Replicates" section shows pairwise comparisons between the NN models and the other models across 1000 bootstrap replicates of the test set; the row for $A > B$ shows the proportion of bootstrap replicates where model A outperformed model B with respect to each metric.

	O/E	AUC	sqrt(Brier Score)
Overall (114 cases)			
Absolute Performance			
FCNN	1.14 (0.94, 1.33)	0.63 (0.58, 0.68)	0.129 (0.117, 0.140)
CNN	1.08 (0.89, 1.26)	0.64 (0.59, 0.68)	0.129 (0.117, 0.139)
BRCAPRO	1.27 (1.05, 1.47)	0.65 (0.61, 0.69)	0.129 (0.117, 0.139)
BRCAPRO^C	1.14 (0.94, 1.33)	0.65 (0.61, 0.69)	0.129 (0.117, 0.139)
LR	1.01 (0.84, 1.18)	0.62 (0.57, 0.67)	0.129 (0.118, 0.140)
Comparisons Across Bootstrap Replicates			
FCNN>CNN	0.148	0.383	0.500
FCNN>BRCAPRO^C	0.582	0.264	0.496
FCNN>LR	0.209	0.710	0.986
CNN>BRCAPRO^C	0.854	0.301	0.486
CNN>LR	0.292	0.766	0.989
Population-Based (63 cases)			
Absolute Performance			
FCNN	1.11 (0.85, 1.39)	0.68 (0.62, 0.74)	0.127 (0.112, 0.142)
CNN	1.04 (0.80, 1.29)	0.69 (0.63, 0.75)	0.127 (0.112, 0.142)
BRCAPRO	1.35 (1.04, 1.69)	0.65 (0.59, 0.70)	0.128 (0.112, 0.143)
BRCAPRO^C	1.20 (0.92, 1.50)	0.65 (0.59, 0.70)	0.128 (0.112, 0.142)
LR	1.02 (0.78, 1.27)	0.67 (0.60, 0.73)	0.128 (0.112, 0.143)
Comparisons Across Bootstrap Replicates			
FCNN>CNN	0.316	0.365	0.524
FCNN>BRCAPRO^C	0.841	0.862	0.779
FCNN>LR	0.334	0.805	0.963
CNN>BRCAPRO^C	0.766	0.970	0.797
CNN>LR	0.441	0.800	0.944
Clinic-Based (39 cases)			
Absolute Performance			
FCNN	1.46 (1.03, 1.89)	0.58 (0.47, 0.68)	0.144 (0.122, 0.164)
CNN	1.40 (1.00, 1.82)	0.59 (0.49, 0.69)	0.144 (0.122, 0.164)
BRCAPRO	1.36 (0.96, 1.76)	0.62 (0.52, 0.70)	0.145 (0.122, 0.164)
BRCAPRO^C	1.25 (0.88, 1.61)	0.62 (0.52, 0.70)	0.145 (0.122, 0.164)
LR	1.21 (0.85, 1.56)	0.57 (0.47, 0.66)	0.145 (0.122, 0.164)
Comparisons Across Bootstrap Replicates			
FCNN>CNN	0.023	0.355	0.328
FCNN>BRCAPRO^C	0.045	0.246	0.643
FCNN>LR	0.049	0.568	0.911
CNN>BRCAPRO^C	0.050	0.307	0.686
CNN>LR	0.055	0.696	0.961

performance compared to BRCAPRO in the overall cohort. They had slightly better discriminatory accuracy than BRCAPRO in population-based probands but performed worse than BRCAPRO in clinic-based probands with a stronger family history. These results are promising because BRCAPRO is based on domain knowledge accumulated over two decades of epidemiological studies (including Miki et al. (1994); Wooster et al. (1995); Easton et al. (1995); Antoniou et al. (2002); Chen and Parmigiani (2007)) while the NNs were trained on a single dataset. The poorer performance of the NNs in clinic-based probands may partly be explained by the fact that the NNs used less detailed family history information than BRCAPRO. Due to missing data in the training set, we did not include age information on non-probands in the NN inputs. This information could potentially improve the accuracy of the NNs. The performance of the NNs could also be improved by considering risk factors besides family history. Since NNs are empirical models, they can easily be extended to handle additional features by adding the features to the input vector. It is less straightforward to incorporate additional risk factors into Mendelian models because explicit assumptions need to be made about how the risk factors modify the genotype-specific risks.

In addition to pursuing accuracy gains through additional data and risk factors, an important direction for future work is to establish ways of measuring and promoting interpretability. While NNs allow for greater flexibility than Mendelian models and regression models, their predictions are challenging to interpret. Various methods have recently been proposed for developing more interpretable NN models, especially for image classification (Dong et al., 2017; Zhang et al., 2018; Chen et al., 2019), but further investigation is needed in the context of family history-based cancer risk prediction.

While NNs require further development and validation before they can be considered as a viable competitor to existing family history-based risk prediction models, our work indicates that they can potentially be a helpful tool for investigating and assessing familial risk.

3

Merging versus Ensembling in Multi-Study Machine Learning: Theoretical Insight from Random Effects

3.1 Introduction

Prediction and classification models trained on a single study often perform considerably worse in external validation than in cross-validation (Castaldi et al., 2011; Bernau et al., 2014). Their generalizability is compromised by overfitting, but also by various sources of study heterogeneity, including differences in study design, data collection and measurement methods, unmeasured confounders, and study-specific sample characteristics (Zhang et al., 2018). Using multiple training studies can potentially address these challenges and lead to more replicable prediction models. In many settings, such as precision medicine, multi-study learning is motivated by systematic data sharing and data curation initiatives. For example, the establishment of gene expression databases such as Gene Expression Omnibus (Edgar et al., 2002) and ArrayExpress (Parkinson et al., 2010) and neuro-imaging databases such as OpenNeuro (Gorgolewski et al., 2017) has facilitated access to sets of studies that provide comparable measurements of the same outcome and predictors. Even if the original measurements are not be comparable, they can often be made comparable through preprocessing (Lazar et al., 2012; Benito et al., 2004). For problems where such a set of studies is available, it is important to systematically integrate information across datasets when developing prediction and classification models.

One common approach is to merge all of the datasets and treat the observations as if they are all from the same study (for example, see Xu et al. (2008); Jiang et al. (2004)). The resulting increase in sample size can lead to better performance when the datasets are relatively homogeneous. Also, the merged dataset is often representative of a broader reference population than any of the individual

datasets. Xu et al. (2008) showed that a prognostic test for breast cancer metastases developed from merged data performed better than the prognostic tests developed using individual studies. Zhou et al. (2017) proposed hypothesis tests for determining when it is beneficial to pool data across multiple sites for linear regression, compared to using data from a single site.

Another approach is to combine results from separately trained models. Meta-analysis and ensembling both fall under this approach. Meta-analysis combines summary measures from multiple studies to increase statistical power (for example, see Riester et al. (2012); Tseng et al. (2012); Rashid et al. (2019)). A common combination strategy is to take a weighted average of the study-specific summary measures. In fixed effects meta-analysis, the weights are based on the assumption that there is a single true parameter underlying all of the studies, while in random effects meta-analysis, the weights are based on a model where the true parameter varies across studies according to a probability distribution. When learners are indexed by a finite number of common parameters, meta-analysis applied to these parameters can be used for multi-study learning, with useful results (Riester et al., 2012). Various studies have compared meta-analysis to merging. For effect size estimation, Bravata and Olkin (2001) showed that merging heterogeneous datasets can lead to spurious results while meta-analysis protects against such problematic effects. Taminau et al. (2014) and Kosch and Jung (2018) found that merging with batch effect removal had higher sensitivity than meta-analysis in gene expression analysis, while Lagani et al. (2016) found that the two approaches performed comparably in reconstruction of gene interaction networks. Ensemble learning methods (Dietterich, 2000a), which combine predictions from multiple models, can also be used to leverage information from multiple studies. Ensembling can lead to lower variance and higher accuracy, and is applicable to more general classes of learners than meta-analysis. Patil and Parmigiani (2018) proposed multi-study ensembles, which are weighted averages of prediction models trained on different studies, as an alternative to merging. They showed empirically that when the datasets are heterogeneous, multi-study ensembling can lead to improved generalizability and replicability compared to merging and meta-analysis.

In this paper, we provide explicit conditions under which merging outperforms multi-study ensembling and vice versa. We study merged and ensemble learners based on ordinary least squares and ridge regression. We hypothesize a mixed effects model for heterogeneity and show that merg-

ing has lower prediction error than ensembling when heterogeneity is low, but as heterogeneity increases, there exists a transition point beyond which ensembling outperforms merging. We characterize this transition point analytically under a fixed weighting scheme for the ensemble. We also discuss optimal ensemble weights for least squares and ridge regression. We study the transition point via simulations and compare merging and ensembling in practice, using microbiome data.

3.2 Problem Definition

We will use the following matrix notation: $\boldsymbol{I_N}$ is the $N \times N$ identity matrix, $\boldsymbol{0_{N \times M}}$ is an $N \times M$ matrix of 0's, $\boldsymbol{0_N}$ is a vector of 0's of length N, $tr(\boldsymbol{A})$ is the trace of matrix $\boldsymbol{A}$, $diag(\boldsymbol{u})$ is a diagonal matrix with $\boldsymbol{u}$ along its diagonal, and $(\boldsymbol{A})_{ij}$ is the entry in row i and column j of matrix $\boldsymbol{A}$. Other notation introduced throughout the paper is summarized in Supplementary Table S.3.1.

Suppose we have K studies that measure the same outcome and the same p predictors, and the datasets have been harmonized so that measurements across studies are on the same scale. For study k, let n_k denote the number of observations, $\boldsymbol{Y_k} \in \mathbb{R}^{n_k}$ the outcome vector, and $\boldsymbol{X_k} \in \mathbb{R}^{n_k \times p}$ the design matrix, where the first column of $\boldsymbol{X_k}$ is a vector of 1's if there is an intercept. Assume the data are generated from the linear mixed effects model

$$\boldsymbol{Y_k} = \boldsymbol{X_k}\boldsymbol{\beta} + \boldsymbol{Z_k}\boldsymbol{\gamma_k} + \boldsymbol{\epsilon_k} \tag{3.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of fixed effects, $\boldsymbol{Z_k} \in \mathbb{R}^{n_k \times q}$ is the design matrix for the random effects obtained by subsetting $\boldsymbol{X_k}$, $\boldsymbol{\gamma_k} \in \mathbb{R}^q$ is the vector of random effects with $E[\boldsymbol{\gamma_k}] = \boldsymbol{0_{n_k}}$ and $cov(\boldsymbol{\gamma_k}) = \boldsymbol{G} = diag(\sigma_1^2, \ldots, \sigma_q^2)$ where $\sigma_i^2 \geq 0$, $\boldsymbol{\epsilon_k} \in \mathbb{R}^{n_k}$ is the vector of residual errors with $E[\boldsymbol{\epsilon_k}] = \boldsymbol{0_{n_k}}$ and $cov(\boldsymbol{\epsilon_k}) = \sigma_\epsilon^2 \boldsymbol{I_{n_k}}$, and $cov((\boldsymbol{\gamma_k})_i, \boldsymbol{\epsilon_k}) = \boldsymbol{0_{n_k}}$ for $i = 1, \ldots, q$. For $j = 1, \ldots, q$, if $\sigma_j^2 > 0$, then the effect of the corresponding predictor differs across studies, and if $\sigma_j^2 = 0$, then the predictor has the same effect in each study.

The relationship between the predictors and the outcome in a given study can be seen as a perturbation of the population-level effect vector $\boldsymbol{\beta}$. The degree of heterogeneity in predictor-outcome relationships across studies can be summarized by the sum of the variances of the random

effects divided by the number of fixed effects: $\overline{\sigma^2} = tr(\boldsymbol{G})/p$. We are interested in comparing the performance of two multi-study learning approaches as $\overline{\sigma^2}$ varies: 1) merging all of the studies and fitting a single linear regression model, and 2) fitting a linear regression model on each study and forming a multi-study ensemble by taking a weighted average of the predictions.

Learners. For low-dimensional settings where $p < n_k$ for all k, we consider merged and ensemble learners based on least squares. The least squares estimator of $\boldsymbol{\beta}$ based on the merged data is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{3.2}$$

where $\boldsymbol{Y} = (\boldsymbol{Y_1^T}, \ldots, \boldsymbol{Y_K^T})^T \in \mathbb{R}^{\sum_{k=1}^K n_k}$ and $\boldsymbol{X} = \left[\boldsymbol{X_1^T} | \ldots | \boldsymbol{X_K^T}\right]^T \in \mathbb{R}^{\sum_{k=1}^K n_k \times p}$. The least squares estimator based on study k is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,k}} = (\boldsymbol{X_k^T X_k})^{-1}\boldsymbol{X_k^T Y_k} \tag{3.3}$$

and the least squares ensemble estimator is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}} = \sum_{k=1}^K w_k \hat{\boldsymbol{\beta}}_{\boldsymbol{LS,k}} \tag{3.4}$$

where $\sum_{k=1}^K w_k = 1$.

For settings where $\boldsymbol{X_k^T X_k}$ is not invertible, such as high-dimensional settings where $p > n_k$ for some k, we consider merged and ensemble learners based on ridge regression. In ridge regression, the predictors are typically standardized prior to fitting the model (Hoerl and Kennard, 1970). The coefficient estimates based on the standardized data are then transformed back to the original scale. Ridge regression is location-invariant (Brown, 1977), so without loss of generality, we assume that the predictors are scaled but not centered prior to applying ridge regression. We first provide the form of the ridge regression estimators in the case where an intercept is included. Let the scaled versions of $\boldsymbol{X_k}$ and $\boldsymbol{X}$ be denoted by $\tilde{\boldsymbol{X}}_{\boldsymbol{k}} = \boldsymbol{X_k S_k}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{XS}$, where $\boldsymbol{S_k}, \boldsymbol{S} \in \mathbb{R}^{p\times p}$ are positive definite scaling matrices. If scaling is not necessary or desirable (for example, if the predictors are measured in the same units), then set $\boldsymbol{S_k} = \boldsymbol{S} = \boldsymbol{I_p}$. Otherwise, let $\boldsymbol{S_k}$ be diagonal

with $(\boldsymbol{S_k})_{11} = 1$ and $(\boldsymbol{S_k})_{jj}$ equal to the inverse standard deviation of column j of $\boldsymbol{X_k}$ for $j > 1$ and let $\boldsymbol{S}$ be diagonal with $(\boldsymbol{S})_{11} = 1$ and $(\boldsymbol{S})_{jj}$ equal to the inverse standard deviation of column j of $\boldsymbol{X}$ for $j > 1$. The merged ridge regression estimator of $\boldsymbol{\beta}$ can be written as

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}} = \boldsymbol{S}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + \lambda \boldsymbol{I_p^-})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I_p^-}\boldsymbol{S}^{-2})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \tag{3.5}$$

where $\lambda \geq 0$ is the regularization parameter and $\boldsymbol{I_p^-}$ is obtained from $\boldsymbol{I_p}$ by setting $(\boldsymbol{I_p})_{11} = 0$, so that the intercept is not regularized (Brown, 1977). The estimator of $\boldsymbol{\beta}$ from study k is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{R,k}} = \boldsymbol{S_k}(\tilde{\boldsymbol{X}}_{\boldsymbol{k}}^T\tilde{\boldsymbol{X}}_{\boldsymbol{k}} + \lambda_k \boldsymbol{I_p^-})^{-1}\tilde{\boldsymbol{X}}_{\boldsymbol{k}}^T\boldsymbol{Y_k} = (\boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{X_k} + \lambda_k \boldsymbol{I_p^-}\boldsymbol{S}_{\boldsymbol{k}}^{-2})^{-1}\boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{Y_k}, \tag{3.6}$$

where $\lambda_k \geq 0$ is the regularization parameter for study k, and the ensemble estimator is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{R,E}} = \sum_{k=1}^{K} w_k \hat{\boldsymbol{\beta}}_{\boldsymbol{R,k}} \tag{3.7}$$

If there is no intercept, then we set $(\boldsymbol{S_k})_{11}$ and $(\boldsymbol{S})_{11}$ to be the inverse standard deviations of the first columns of $\boldsymbol{X_k}$ and $\boldsymbol{X}$ respectively and replace $\boldsymbol{I_p^-}$ with $\boldsymbol{I_p}$ in the expressions above.

To make progress analytically, we assume that the weights w_k and the regularization parameters λ and λ_k are predetermined (for example, using independent data (Tsybakov, 2014)). For linear regression, averaging predictions across study-specific learners is equivalent to averaging the estimated coefficient vectors across study-specific learners and then computing predictions. Multi-study ensembling based on linear regression on scaled variables has points in common with meta-analysis of effect sizes. In the univariate case, a standard meta-analysis is also a weighted average of β_k's, though typically weights reflect precision of estimates rather than cross-study features of predictions. When $p > 1$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}}$ weights each dimension of the coefficient vector equally in a given study while meta-analytic approaches, which involve either performing separate univariate meta-analyses for each predictor or performing a multivariate meta-analysis (for example, see Jackson et al. (2010, 2011)), do not impose this constraint.

Performance Comparison. Given a test set with design matrix $\boldsymbol{X_0}$ and outcome vector $\boldsymbol{Y_0}$, the goal is to identify conditions under which multi-study ensembling has lower mean squared

prediction error than merging, i.e.

$$E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\cdot,\boldsymbol{E}}\|_2^2] \leq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\cdot,\boldsymbol{M}}\|_2^2]$$

where the expectations are taken with respect to $\boldsymbol{Y_0}$ and $\|(u_1, \ldots, u_m)^T\|_2 = \sqrt{\sum_{i=1}^m u_i^2}$.

3.3 Theoretical Results

3.3.1 Overview

We consider two cases for the structure of $\boldsymbol{G}$: equal variances and unequal variances. Let $\sigma^2_{(1)}, \ldots, \sigma^2_{(r)}$ be the distinct values on the diagonal of $\boldsymbol{G}$ and let m_j be the number of random effects with variance $\sigma^2_{(j)}$. In the equal variances case where $r = 1$ and $\sigma_j^2 = \sigma^2$ for $j = 1, \ldots, q$, we provide a necessary and sufficient condition for the ensemble learner to outperform the merged learner. In the unequal variances case, we provide sufficient conditions under which the ensemble learner outperforms the merged learner and vice versa. These conditions allow us to characterize a transition point in terms of the heterogeneity measure $\overline{\sigma^2}$ between a regime that favors merging and a regime that favors ensembling.

Theorems 3.1 and 3.2 state the equal variances and unequal variances results for least squares. Theorems 3.3 and 3.4 state the analogous results for ridge regression. Theorems 3.1-3.3 are special cases of Theorem 3.4 (proved in Supplementary Section S.3.2), but for clarity of presentation, we start with the simplest case and then present the more general cases. We also consider an asymptotic version of the transition point (Corollary 3.1), provide optimal ensemble weights for least squares (Proposition 3.1) and ridge regression (Proposition 3.2), and discuss how to interpret the results.

To present the results more concisely, let $\boldsymbol{R} = \boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{M} = \boldsymbol{R} + \lambda \boldsymbol{I_p^-}\boldsymbol{S}^{-2}$. For $k = 0, 1, \ldots, K$, let $\boldsymbol{R_k} = \boldsymbol{X_k}^T\boldsymbol{X_k}$ and $\boldsymbol{M_k} = \boldsymbol{R_k} + \lambda_k \boldsymbol{I_p^-}\boldsymbol{S_k^{-2}}$. Let $\boldsymbol{\Gamma_{(j)}} \in \mathbb{R}^{q \times m_j}$ be a matrix where $(\boldsymbol{\Gamma_{(j)}})_{il} = 1$ if random effect i is the lth random effect with variance $\sigma^2_{(j)}$ and $(\boldsymbol{\Gamma_{(j)}})_{il} = 0$ otherwise, so that $\boldsymbol{G\Gamma_{(j)}}$ subsets $\boldsymbol{G}$ to the columns corresponding to $\sigma^2_{(j)}$.

3.3.2 Least Squares

For the least squares results, assume $n_k > p$ for all k.

Theorem 3.1. *Suppose* $\sigma_j^2 = \sigma^2$ *for* $j = 1, \ldots, q$ *and*

$$tr(\boldsymbol{R}^{-1} \sum_{k=1}^{K} \boldsymbol{X_k^T Z_k Z_k^T X_k R^{-1} R_0}) - \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{Z_0^T Z_0}) > 0. \tag{3.8}$$

Define

$$\tau_{LS} = \sigma_\epsilon^2 \times \frac{q}{p} \times \frac{\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{R_k^{-1} R_0}) - tr(\boldsymbol{R^{-1} R_0})}{tr(\boldsymbol{R}^{-1} \sum_{k=1}^{K} \boldsymbol{X_k^T Z_k Z_k^T X_k R^{-1} R_0}) - \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{Z_0^T Z_0})} \tag{3.9}$$

Then $E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}}\|_2^2] \leq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}\|_2^2]$ *if and only if* $\overline{\sigma^2} \geq \tau_{LS}$.

By Theorem 3.1, for any fixed weighting scheme that satisfies Condition 3.8, τ_{LS} represents a transition point from a regime where merging outperforms ensembling to a regime where ensembling outperforms merging. Assuming $\boldsymbol{R_k}$ is not identical for all k, if equal weights $w_k = 1/K$ are used, then Condition 3.8 is satisfied (by Jensen's operator inequality (Hansen and Pedersen, 2003)) and $\tau_{LS} > 0$, so the transition point always exists.

Theorem 3.2.

1. *Suppose*

$$\max_{j=1,\ldots,r} tr(\boldsymbol{R}^{-1} \sum_{k=1}^{K} \boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k R^{-1} R_0}) - \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}) > 0 \tag{3.10}$$

and define

$$\tau_{LS,1} = \frac{\sigma_\epsilon^2}{p} \frac{\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{R_k^{-1} R_0}) - tr(\boldsymbol{R^{-1} R_0})}{\max_{j=1,\ldots,r} \frac{1}{m_j}(tr(\boldsymbol{R}^{-1} \sum_{k=1}^{K} \boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k R^{-1} R_0}) - \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}))}. \tag{3.11}$$

Then $E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}}\|_2^2] \geq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}\|_2^2]$ *when* $\overline{\sigma^2} \leq \tau_{LS,1}$.

2. *Suppose*

$$\min_{j=1,\ldots,r} tr(\boldsymbol{R}^{-1}\sum_{k=1}^{K}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma_{(j)}}\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{R}^{-1}\boldsymbol{R_0}) - \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_0}^T\boldsymbol{Z_0}\boldsymbol{\Gamma_{(j)}}) > 0 \quad (3.12)$$

and define

$$\tau_{LS,2} = \frac{\sigma_\epsilon^2}{p}\frac{\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{R_k}^{-1}\boldsymbol{R_0}) - tr(\boldsymbol{R}^{-1}\boldsymbol{R_0})}{\min_{j=1,\ldots,r}\frac{1}{m_j}(tr(\boldsymbol{R}^{-1}\sum_{k=1}^{K}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma_{(j)}}\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{R}^{-1}\boldsymbol{R_0}) - \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_0}^T\boldsymbol{Z_0}\boldsymbol{\Gamma_{(j)}}))}. \quad (3.13)$$

Then $E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}}\|_2^2] \leq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}\|_2^2]$ *when* $\overline{\sigma^2} \geq \tau_{LS,2}$.

Theorem 3.2 shows that in the more general scenario where the random effects do not necessarily have the same variance, there is a transition interval such that the merged learner outperforms the ensemble learner when $\overline{\sigma^2}$ is smaller than the lower bound of the interval and the ensemble learner outperforms the merged learner when $\overline{\sigma^2}$ is greater than the upper bound of the interval.

Corollary 3.1. *Suppose there exist positive definite matrices* $\boldsymbol{A_1}, \boldsymbol{A_2}, \boldsymbol{A_{(j)}} \in \mathbb{R}^{p\times p}$ *such that as* $K \to \infty$,

1. $\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{R_k} \to \boldsymbol{A_1}$
2. $\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{R_k}^{-1} \to \boldsymbol{A_2}$
3. $\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma_{(j)}}\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_k}^T\boldsymbol{X_k} \to \boldsymbol{A_{(j)}}$ *for* $j = 1, \ldots, r$

where $\to$ *denotes almost sure convergence. Let* $w_k = 1/K$.

1. *If*

$$\max_{j=1,\ldots,r}(tr(\boldsymbol{A_1}^{-1}\boldsymbol{A_{(j)}}\boldsymbol{A_1}^{-1}\boldsymbol{R_0}) - tr(\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_0}^T\boldsymbol{Z_0}\boldsymbol{\Gamma_{(j)}})) > 0 \quad (3.14)$$

then

$$\tau_{LS,1} \to \frac{\sigma_\epsilon^2}{p} \times \frac{tr(\boldsymbol{A_2 R_0}) - tr(\boldsymbol{A_1^{-1} R_0})}{\max\limits_{j=1,\ldots,r} \frac{1}{m_j}(tr(\boldsymbol{A_1^{-1} A_{(j)} A_1^{-1} R_0}) - tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}))} \tag{3.15}$$

2. If

$$\min_{j=1,\ldots,r}(tr(\boldsymbol{A_1^{-1} A_{(j)} A_1^{-1} R_0}) - tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}})) > 0 \tag{3.16}$$

then

$$\tau_{LS,2} \to \frac{\sigma_\epsilon^2}{p} \times \frac{tr(\boldsymbol{A_2 R_0}) - tr(\boldsymbol{A_1^{-1} R_0})}{\min\limits_{j=1,\ldots,r} \frac{1}{m_j}(tr(\boldsymbol{A_1^{-1} A_{(j)} A_1^{-1} R_0}) - tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}))} \tag{3.17}$$

Corollary 3.1 presents an asymptotic version of Theorem 3.2 as the number of studies goes to infinity. If all study sizes are equal to n and the predictors are independent and identically distributed within and across studies, then $\boldsymbol{A_1} = E[\boldsymbol{R_k}]$, $\boldsymbol{A_2} = E[\boldsymbol{R_k^{-1}}]$, and $\boldsymbol{A_{(j)}} = E[\boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k}]$. In the special case where $p = q = 1$ and the predictor follows $N(0, v)$, $\tau_{LS,1} = \tau_{LS,2}$ converges to

$$\frac{\sigma_\epsilon^2}{(n-2)v}, \tag{3.18}$$

so asymptotically the transition point is controlled simply by the variance of the residuals, the variance of the predictor, and the study sample size.

Proposition 3.1. *The optimal weights for the least squares ensemble are*

$$w_k = \frac{(tr(\boldsymbol{G Z_0^T Z_0}) + \sigma_\epsilon^2 tr(\boldsymbol{R_k^{-1} R_0}))^{-1}}{\sum_{k=1}^K (tr(\boldsymbol{G Z_0^T Z_0}) + \sigma_\epsilon^2 tr(\boldsymbol{R_k^{-1} R_0}))^{-1}}. \tag{3.19}$$

The optimal weight for study k is proportional to the inverse mean squared prediction error of the least squares learner trained on that study. In the equal variances setting, $tr(\boldsymbol{G Z_0^T Z_0}) = \sigma^2 tr(\boldsymbol{Z_0^T Z_0})$. The optimal weights depend on σ^2, so τ_{LS} depends on σ^2 under the optimal weighting scheme. Thus, it is difficult to obtain a closed-form expression for the transition point. However, the transition point under any fixed weighting scheme provides an upper bound for the transition

point under the optimal weighting scheme, and the optimal weights transition point ($\overline{\sigma^2}$ such that $\overline{\sigma^2} = \tau_{LS}$) can be computed numerically. In practice, σ^2 and σ_ϵ^2 are not known, so the optimal weights need to be estimated. Weight estimation increases the variance of the ensemble learner and may result in a higher transition point than when the optimal weights are known. However, our simulations suggest that if σ^2 and σ_ϵ^2 can be reasonably estimated (for example, via a linear mixed effects model), then the estimation will have little impact on the transition point (see Supplementary Figure S.3.4).

3.3.3 Ridge Regression

Next, we present results for ridge regression that generalize the linear regression results and are applicable to both low- and high-dimensional settings. Let $\boldsymbol{b_E} = -\sum_k w_k \lambda_k \boldsymbol{X_0 M_k^{-1} I_p^- S_k^{-2} \beta}$ be the bias of the ensemble predictions in the test set and $\boldsymbol{b_M} = -\lambda \boldsymbol{X_0 M^{-1} I_p^- S^{-2} \beta}$ the bias of the merged predictions in the test set.

Theorem 3.3. *Suppose $\sigma_j^2 = \sigma^2$ for $j = 1, \ldots, q$ and*

$$tr(\boldsymbol{M^{-1}} \sum_{k=1}^{K} \boldsymbol{X_k^T Z_k Z_k^T X_k M^{-1} R_0}) - \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{M_k^{-1} X_k^T Z_k Z_k^T X_k M_k^{-1} R_0}) > 0. \tag{3.20}$$

Define

$$\tau_R = \frac{q}{p} \times \frac{\sigma_\epsilon^2 \left(\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{M_k^{-1} R_k M_k^{-1}} \mathbf{R_0}) - tr(\boldsymbol{M^{-1}} \sum_{k=1}^{K} \boldsymbol{R_k M^{-1} R_0}) \right) + \boldsymbol{b_E^T b_E} - \boldsymbol{b_M^T b_M}}{tr(\boldsymbol{M^{-1}} \sum_{k=1}^{K} \boldsymbol{X_k^T Z_k Z_k^T X_k M^{-1} R_0}) - \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{M_k^{-1} X_k^T Z_k Z_k^T X_k M_k^{-1} R_0})} \tag{3.21}$$

Then $E[\|\boldsymbol{Y_0} - \boldsymbol{X_0 \hat{\beta}_{R,E}}\|_2^2] \leq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0 \hat{\beta}_{R,M}}\|_2^2]$ if and only if $\overline{\sigma^2} \geq \tau_R$.

Theorem 3.4.

1. *Suppose*

$$\max_{j=1,\ldots,r} tr(\boldsymbol{M^{-1}} \sum_{k=1}^{K} \boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k M^{-1} R_0}) - \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{M_k^{-1} X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k M_k^{-1} R_0}) > 0 \tag{3.22}$$

and define

$$\tau_{R,1}=\frac{\sigma_\epsilon^2\left(\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{M_k}^{-1}\boldsymbol{R_k}\boldsymbol{M_k}^{-1}\mathbf{R_0})-tr(\boldsymbol{M}^{-1}\sum_{k=1}^{K}\boldsymbol{R_k}\boldsymbol{M}^{-1}\boldsymbol{R_0})\right)+\boldsymbol{b_E}^T\boldsymbol{b_E}-\boldsymbol{b_M}^T\boldsymbol{b_M}}{p\max_{j=1,\ldots,r}\frac{1}{m_j}(tr(\boldsymbol{M}^{-1}\sum_{k=1}^{K}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma_{(j)}}\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M}^{-1}\boldsymbol{R_0})-\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{M_k}^{-1}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma_{(j)}}\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M_k}^{-1}\boldsymbol{R_0}))}. \tag{3.23}$$

Then $E[\|\boldsymbol{Y_0}-\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,E}}\|_2^2]\geq E[\|\boldsymbol{Y_0}-\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}}\|_2^2]$ *when* $\overline{\sigma^2}\leq\tau_{R,1}$.

2. *Suppose*

$$\min_{j=1,\ldots,r} tr(\boldsymbol{M}^{-1}\sum_{k=1}^{K}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma_{(j)}}\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M}^{-1}\boldsymbol{R_0})-\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{M_k}^{-1}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma_{(j)}}\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M_k}^{-1}\boldsymbol{R_0})>0 \tag{3.24}$$

and define

$$\tau_{R,2}=\frac{\sigma_\epsilon^2\left(\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{M_k}^{-1}\boldsymbol{R_k}\boldsymbol{M_k}^{-1}\mathbf{R_0})-tr(\boldsymbol{M}^{-1}\sum_{k=1}^{K}\boldsymbol{R_k}\boldsymbol{M}^{-1}\boldsymbol{R_0})\right)+\boldsymbol{b_E}^T\boldsymbol{b_E}-\boldsymbol{b_M}^T\boldsymbol{b_M}}{p\min_{j=1,\ldots,r}\frac{1}{m_j}(tr(\boldsymbol{M}^{-1}\sum_{k=1}^{K}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma_{(j)}}\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M}^{-1}\boldsymbol{R_0})-\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{M_k}^{-1}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma_{(j)}}\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M_k}^{-1}\boldsymbol{R_0}))}. \tag{3.25}$$

Then $E[\|\boldsymbol{Y_0}-\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,E}}\|_2^2]\leq E[\|\boldsymbol{Y_0}-\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}}\|_2^2]$ *when* $\overline{\sigma^2}\geq\tau_{R,2}$.

Proposition 3.2. *Let* $v_k=tr(\boldsymbol{M_k}^{-1}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{G}\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M_k}\boldsymbol{R_0})+\sigma_\epsilon^2 tr(\boldsymbol{M_k}^{-1}\boldsymbol{X_k}^T\boldsymbol{X_k}\boldsymbol{M_k}^{-1}\mathbf{R_0})$ *and* $\boldsymbol{b_k}=\lambda_k\boldsymbol{X_0}\boldsymbol{M_k}^{-1}\boldsymbol{I_p^-}\boldsymbol{S_k}^{-2}\boldsymbol{\beta}$. *Then the optimal weights for the ridge regression ensemble are*

$$\boldsymbol{w_k}=\frac{\sum_{j=1}^{K}(\boldsymbol{C}^{-1})_{kj}}{\sum_{i=1}^{K}\sum_{j=1}^{K}(\boldsymbol{C}^{-1})_{kj}} \tag{3.26}$$

where $\boldsymbol{C}\in\mathbb{R}^{K\times K}$ *has entries* $(\boldsymbol{C})_{kk}=v_k+\boldsymbol{b_k}^T\boldsymbol{b_k}$ *and* $(\boldsymbol{C})_{jk}=\boldsymbol{b_j}^T\boldsymbol{b_k}$ *for* $j\neq k$.

Since $\hat{\boldsymbol{\beta}}_{\boldsymbol{R,k}}$ is biased for $\lambda_k>0$, the optimal weights for ridge regression depend on the true population-level effects $\boldsymbol{\beta}$ as well as $\boldsymbol{G}$ and σ_ϵ^2. Using a first-order Taylor approximation, the optimal w_k is approximately proportional to

$$(v_k+\boldsymbol{b_k}^T\boldsymbol{b_k})^{-1}-(v_k+\boldsymbol{b_k}^T\boldsymbol{b_k})^{-2}-\sum_{j\neq k}\boldsymbol{b_j}^T\boldsymbol{b_k}(v_j+\boldsymbol{b_j}^T\boldsymbol{b_j})^{-1}(v_k+\boldsymbol{b_k}^T\boldsymbol{b_k})^{-1},$$

where the first two terms depend on the inverse mean squared prediction error and the third

term depends on the covariances between the bias terms from study k and the bias terms from other studies, as well as the magnitudes of the studies' prediction errors. If the magnitudes of the prediction error and bias term for study k are held fixed and the other studies are held fixed, then the optimal weight for study k increases as it becomes less correlated with the other studies. This is similar to a result from Krogh and Vedelsby (1995) that decomposes the prediction error of an ensemble into a term that depends on the prediction errors of the individual learners and a correlation term that quantifies the disagreement across the individual learners; the decomposition implies that if the individual prediction errors are held fixed, then the performance of the ensemble improves as the correlation term decreases.

As with least squares, the transition point under any fixed weighting scheme provides an upper bound for the optimal weights transition point, which can be calculated numerically. Weight estimation can potentially shift this transition point, but we expect the shift to be small when the fixed effects and the variances of the random effects and residual errors can be estimated reasonably (see Supplementary Figure S.3.4).

3.3.4 Interpretation

The covariance matrices of linear regression coefficient estimators can be written as a sum of two components, one driven by between-study variability and one driven by within-study variability. For example, when all predictors have random effects, the covariance of the least squares ensemble is

$$cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}}) = \sum_{k=1}^{K} w_k^2 \boldsymbol{G} + \sigma_\epsilon^2 \sum_{k=1}^{K} w_k^2 \boldsymbol{R_k^{-1}}$$

and the covariance of the merged least squares learner is

$$cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1} \sum_{k=1}^{K} [\boldsymbol{R_k G R_k^T}] \boldsymbol{R}^{-1} + \sigma_\epsilon^2 \boldsymbol{R}^{-1}.$$

Since the merged learner ignores between-study heterogeneity, the trace of its first component is generally larger than that of the ensemble learner. However, since the merged learner is trained on a larger sample, the trace of its second component is generally smaller than that of the ensemble

learner. The merged and ensemble learners based on least squares are unbiased, so the transition point depends on the trade-off between these two components. When $p = 1$, Expression 3.18 shows that having a higher-variance predictor favors ensembling over merging, since increasing the variance of the predictor amplifies the impact of the random effect.

Unlike least squares estimators, ridge regression estimators are biased as a result of regularization. The transition point for ridge regression depends on the regularization parameters used on the merged and individual datasets. It also depends on the true coefficient vector $\boldsymbol{\beta}$ through the squared bias terms in the mean squared prediction errors of the merged and ensemble learners, so an estimate of $\boldsymbol{\beta}$ is needed to compute the expressions in Theorems 3.3 and 3.4. These expressions can vary considerably for different choices of regularization parameters and different values of $\boldsymbol{\beta}$. We did not provide the asymptotic results for ridge regression as $K \to \infty$ (with the study sizes held constant) because this scenario is not entirely fair to the ensemble learner. For $p > n_k$ and sufficiently large K, the merged learner will be in the low-dimensional setting while the ensemble learner will remain in the high-dimensional setting. As $K \to \infty$, the bias term approaches 0 for the merged learner (assuming $\lambda/K \to 0$) but not for the ensemble learner, which suggests that when K is sufficiently large, merging will always yield lower mean squared prediction error than ensembling.

In general, the transition points for least squares and ridge regression depend on the design matrix of the test set. However, the test design matrix drops out when it is a scalar multiple of an orthogonal matrix. For example, this occurs when $p = 1$.

3.4 Simulations

We conducted simulations to verify the theoretical results for least squares and ridge regression and to compare them to the empirical transition points for three methods for which we cannot derive a closed-form solution: lasso, single hidden layer neural network, and random forest. We also ran a linear mixed effects model, univariate random effects meta-analyses, and multivariate random effects meta-analysis. We used the R packages `glmnet, nnet, randomForest, nlme, metafor`, and `mvmeta` for ridge regression/lasso, neural networks, random forests, linear mixed effects models,

univariate meta-analyses, and multivariate meta-analyses respectively.

We considered four simulation scenarios corresponding to the settings in Theorems 3.1-3.4. We used 4 training studies and 4 test studies of size 40 and set $\sigma_\epsilon^2 = 1$ for all scenarios. For the low-dimensional settings, we set $p = 10$, $q = 5$, and generated 5 $\boldsymbol{\beta}$'s from $N(0, 1)$ and 5 from $N(0, 0.01^2)$, with 5 of the $\boldsymbol{\beta}$'s having random slopes. For the high-dimensional settings, we set $p = 100$, $q = 10$, and generated 30 $\boldsymbol{\beta}$'s from $N(0, 1)$ and 70 from $N(0, 0.01^2)$, with 10 of the $\boldsymbol{\beta}$'s having random slopes. For each simulation scenario, we fixed the predictor values in the training and test sets and the model hyperparameters. Predictor values were sampled from datasets in the `curatedOvarianData` R package (Ganzfried et al., 2013). Model hyperparameters were tuned once using 5-fold cross-validation with outcomes generated under $\sigma^2 = 0$. For various values of $\overline{\sigma^2}$, including 0 and the theoretical transition point, we generated random slopes, residual errors, and outcomes for each training and test study according to Model (3.1), then trained and tested the following approaches: linear mixed effects model, random effects meta-analysis of univariate least squares estimates, random effects multivariate meta-analysis, and merged and ensemble learners based on least squares, ridge regression, lasso, neural networks, and random forests. For ridge regression and lasso, the predictors were standardized prior to model fitting. For linear mixed effects, we fit the true model, using restricted maximum likelihood to estimate $\boldsymbol{G}$. For meta-analysis of univariate least squares estimates, we used the DerSimonian and Laird method. For multivariate meta-analysis, we used restricted maximum likelihood to estimate $\boldsymbol{G}$, constraining the covariance matrix to be diagonal. Least squares, linear mixed effects, and meta-analysis were only applied in the low-dimensional setting. We performed 1000 replicates for each value of $\overline{\sigma^2}$ and estimated the mean squared prediction error of each estimator by averaging the squared error across replicates.

As seen in Figures 3.1 and 3.2, the empirical transition points for least squares and ridge regression agree with the theoretical results from Theorems 3.1 and 3.3 (see Supplementary Section S.3.3 for plots for Theorems 3.2 and 3.4). The methods all have similar empirical transition points except for random forest, which performed considerably worse than all of the other approaches (see Figure S.3.3 in S.3.3). The poor performance of random forest could be because the data were generated from a linear model. The univariate meta-analysis approach also performed poorly, which is

unsurprising because the generating model is a multivariate model. The performance of the other models relative to the data generating model is summarized in Figure 3.3 for three values of $\overline{\sigma^2}$. When $\overline{\sigma^2} = 0$, the merged regression learners and multivariate meta-analysis perform as well as or slightly better than the mixed effects model and outperform the ensembles. The merged neural network does slightly worse than the regression learners. At the least squares transition point, all models perform similarly. Beyond the transition point, the models continue to perform similarly (when heterogeneity is high, all models perform poorly), with the ensembles slightly outperforming the merged learners and multivariate meta-analysis performing as well the mixed effects model. For each of the three values of $\overline{\sigma^2}$, lasso performed best, even slightly outperforming the mixed effects model and multivariate meta-analysis. This is likely because several of the true $\boldsymbol{\beta}$'s were close to 0.

3.5 Data Application

To illustrate in a practical example, we compared the performance of merging and ensembling used datasets from the `curatedMetagenomicData` R package (Pasolli et al., 2017), which contains a collection of curated, uniformly processed human microbiome data. We focused on three gut microbiome studies that measured cholesterol as well as gene marker abundance in stool, restricting to samples from female patients: 1) a study of Chinese type 2 diabetes patients and non-diabetic controls ($n_1 = 151$ samples from independent female patients) (Qin et al., 2012), 2) a study of middle-aged European women with normal, impaired or diabetic glucose control ($n_2 = 145$ samples from independent female patients) (Karlsson et al., 2013), and 3) a study of patients with a family history of type 1 diabetes ($n_0 = 32$ samples from 13 female patients) (Heintz-Buschart et al., 2017). We used merging and ensembling to train linear regression models to predict cholesterol, calculated the theoretical transition interval, and evaluated the performance of the two approaches.

We considered two scenarios: 1) training on different subsets of the same study and testing on a held out subset, and 2) training on different studies and testing on an independent study. In the first scenario, we randomly split the samples from Qin et al. (2012) into five datasets of approximately equal size, using four for training and the remaining one for testing. We used age and the top five marker abundances most correlated with the outcome in the training set as the predictors.

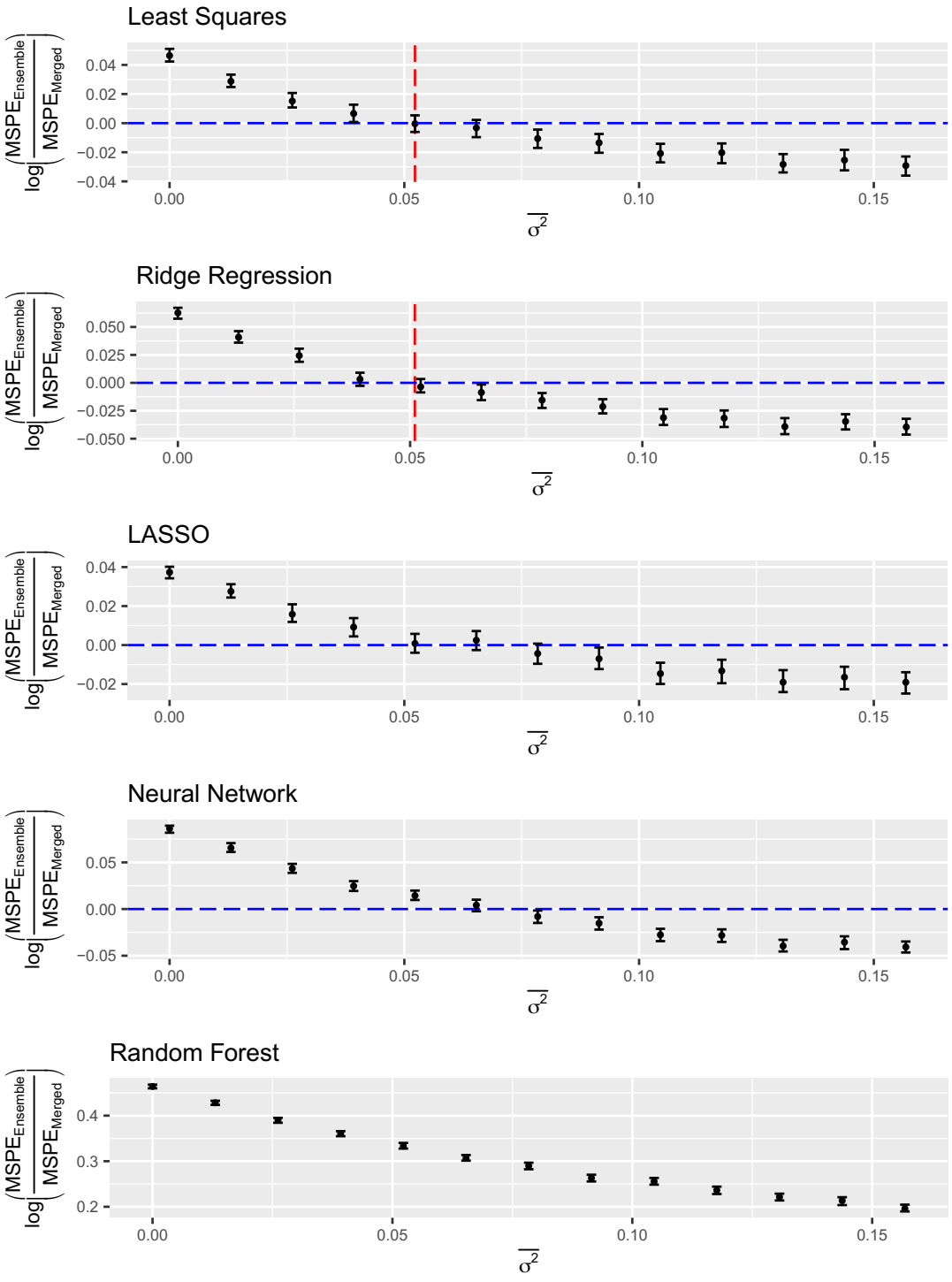

Figure 3.1: Relative performance of merging and ensembling as a function of heterogeneity when $p = 10$, $q = 5$, and the random effects have equal variances. MSPE: mean squared prediction error. The vertical dashed lines correspond to the theoretical transition points from Theorems 3.1 and 3.3. The empirical transition point occurs at the value of $\overline{\sigma^2}$ where the log ratio of the prediction errors for ensembling and merging is equal to 0.

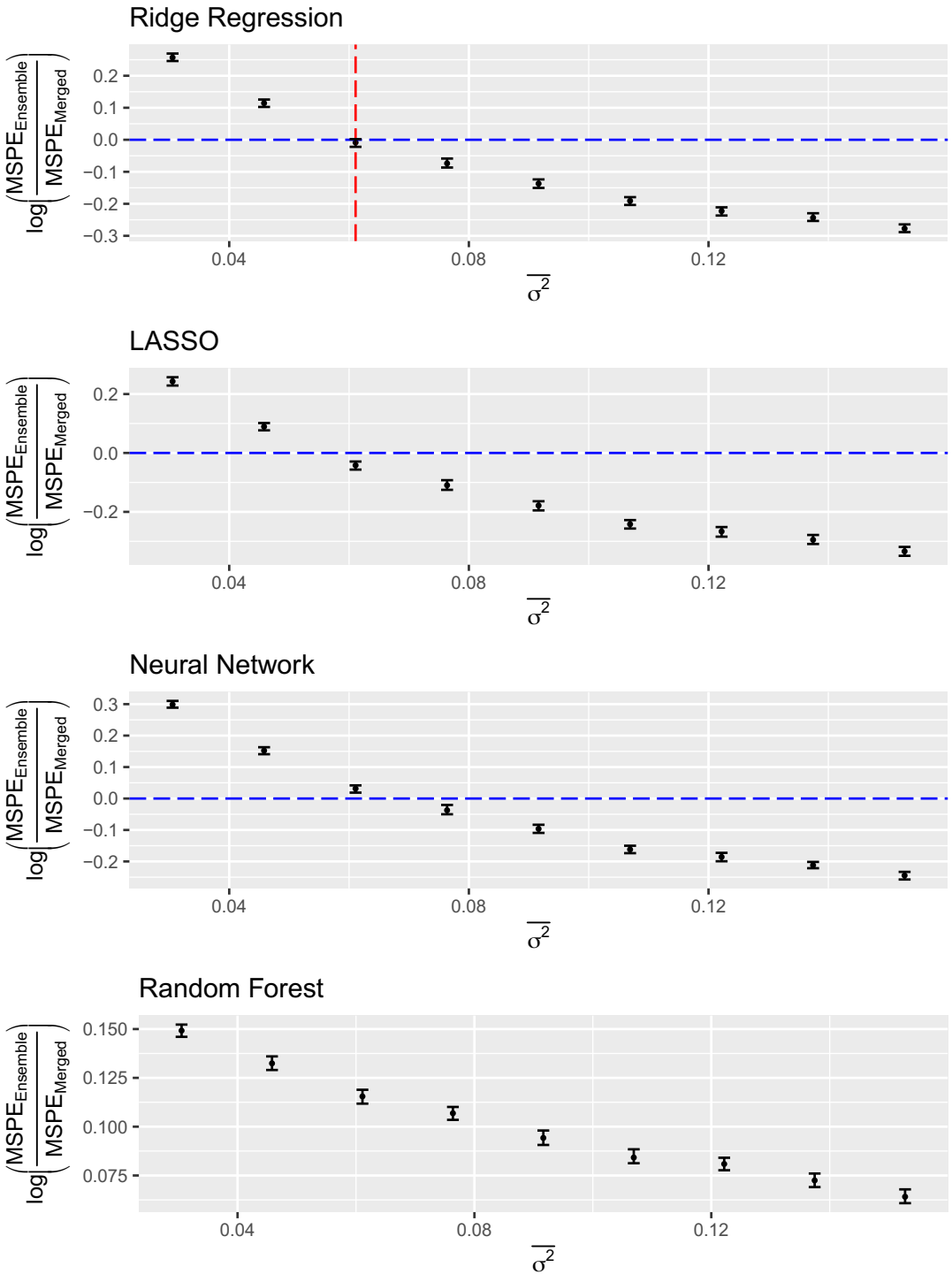

Figure 3.2: Relative performance of merging and ensembling as a function of heterogeneity when $p = 100$, $q = 10$, and the random effects have equal variances. MSPE: mean squared prediction error. The vertical dashed line corresponds to the theoretical transition point from Theorem 3.3. The empirical transition point occurs at the value of $\overline{\sigma^2}$ where the log ratio of the prediction errors for ensembling and merging is equal to 0.

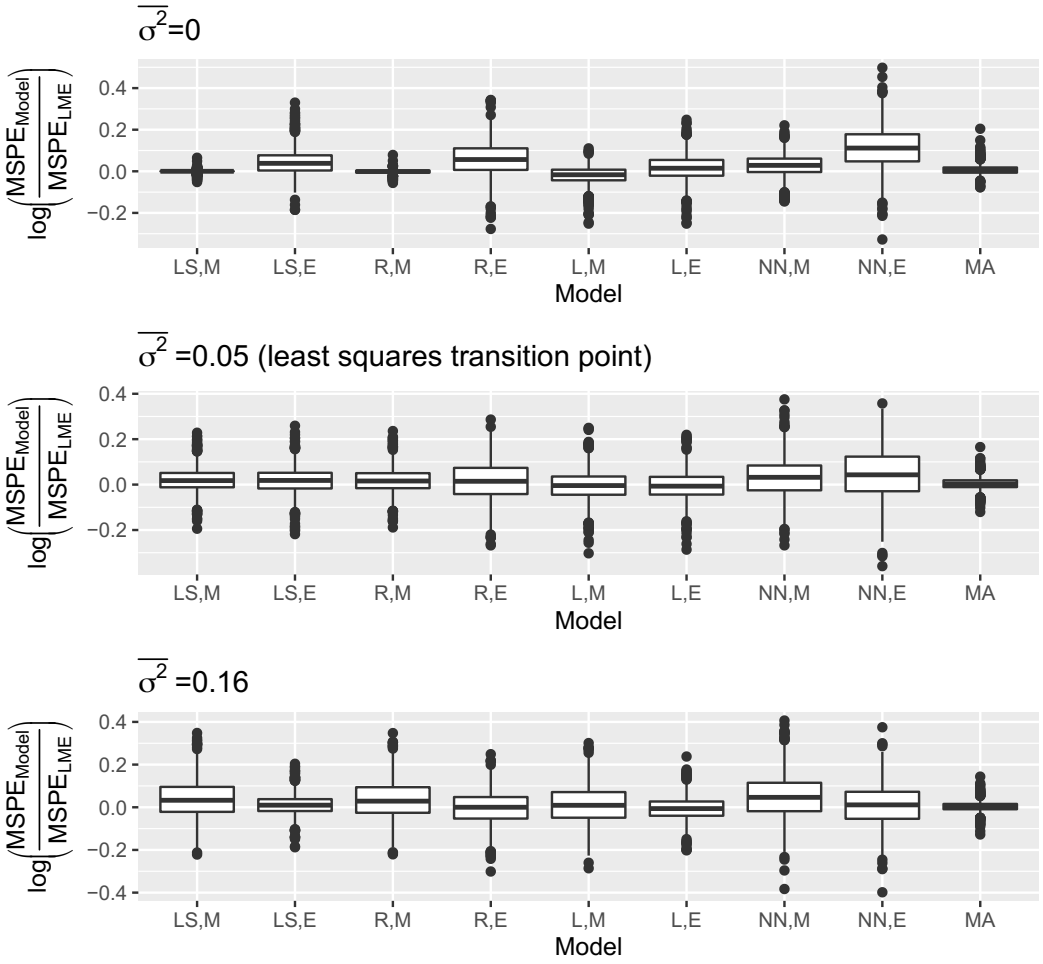

Figure 3.3: Performance comparisons for three values of $\overline{\sigma^2}$ when $p = 10$, $q = 5$, and the random effects have equal variances (the random forest and univariate meta-analysis results are omitted to avoid stretching the y-axis but are shown in Supplementary Section S.3.3). MSPE: mean squared prediction error; LME: linear mixed effects model; LS,M: merged least squares learner; LS,E: ensemble learner based on least squares; R,M: merged ridge regression learner; R,E: ensemble learner based on ridge regression; L,M: merged lasso learner; L,E: ensemble learner based on lasso; NN,M: merged neural network; NN,E: ensemble learner based on neural networks; MA: multivariate meta-analysis.

In second scenario, we used the datasets from Qin et al. (2012) and Karlsson et al. (2013) for training and the dataset from Heintz-Buschart et al. (2017) for testing. We used age and the top twenty marker abundances most correlated with the outcome in the training set as the predictors. In each scenario, we fit merged and ensemble versions of least squares and ridge regression. We estimated $\boldsymbol{G}$ by fitting a linear mixed effects model using residual maximum likelihood, allowing each predictor to have a random effect. Using equal weights for the ensemble learners, we calculated the theoretical transition bounds from Theorems 3.2 and 3.4 and compared them to the estimate

of $\overline{\sigma^2}$. We evaluated the performance of merging, ensembling with equal weights, and ensembling with optimal weights by calculating the prediction error on the test set. To estimate the optimal weights, we plugged in estimates of $\boldsymbol{G}$, σ^2_ϵ, and $\boldsymbol{\beta}$ from the mixed effects model in the expressions in Propositions 3.1 and 3.2.

In the first scenario, $\overline{\sigma^2}$ was estimated to be 0.005^2, $\tau_{LS,1}$ was 0.266^2, and $\tau_{R,1}$ was 0.050^2, so merging was expected outperform ensembling. In the test set, the merged versions of least squares and ridge regression had lower prediction error than the corresponding ensembles (Figure 3.4). In the second scenario, $\overline{\sigma^2}$ was estimated to be 18.322^2, $\tau_{LS,2}$ was 15.247^2, and $\tau_{R,2}$ was 4.423^2, so ensembling was expected to outperform merging. In the test set, the ensemble versions of least squares and ridge regression had lower prediction error than the corresponding merged learners (Figure 3.5).

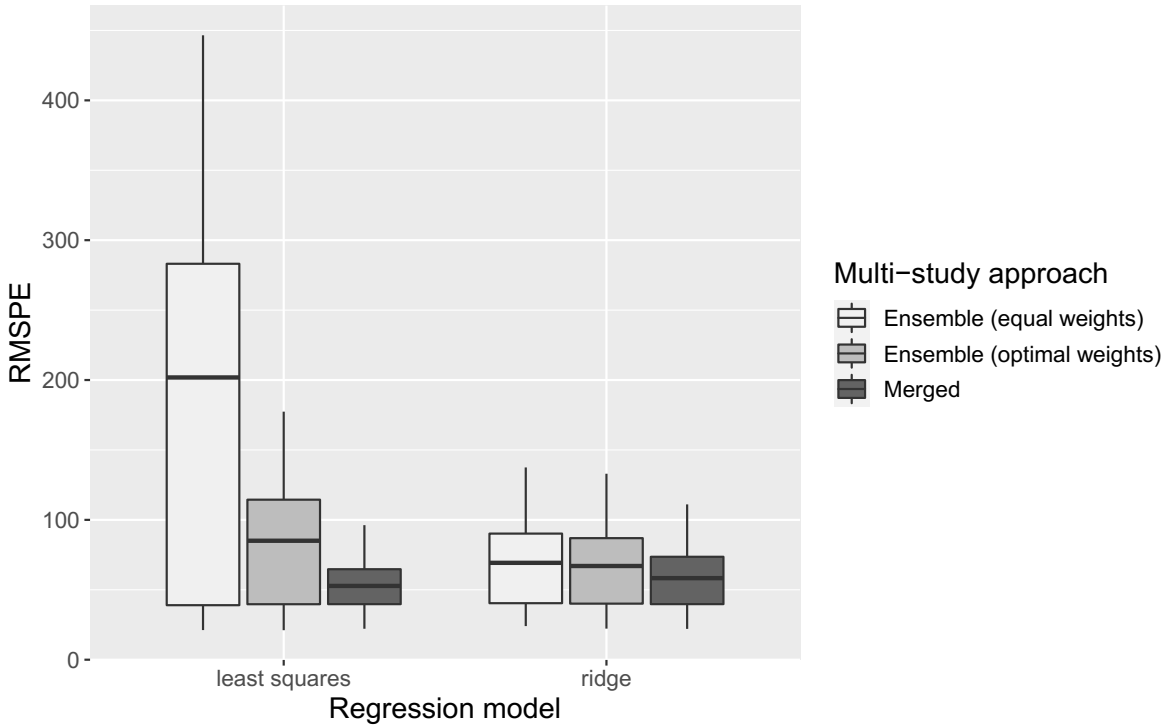

Figure 3.4: Root mean squared prediction error (RMSPE) for the first data illustration scenario with bootstrap confidence intervals.

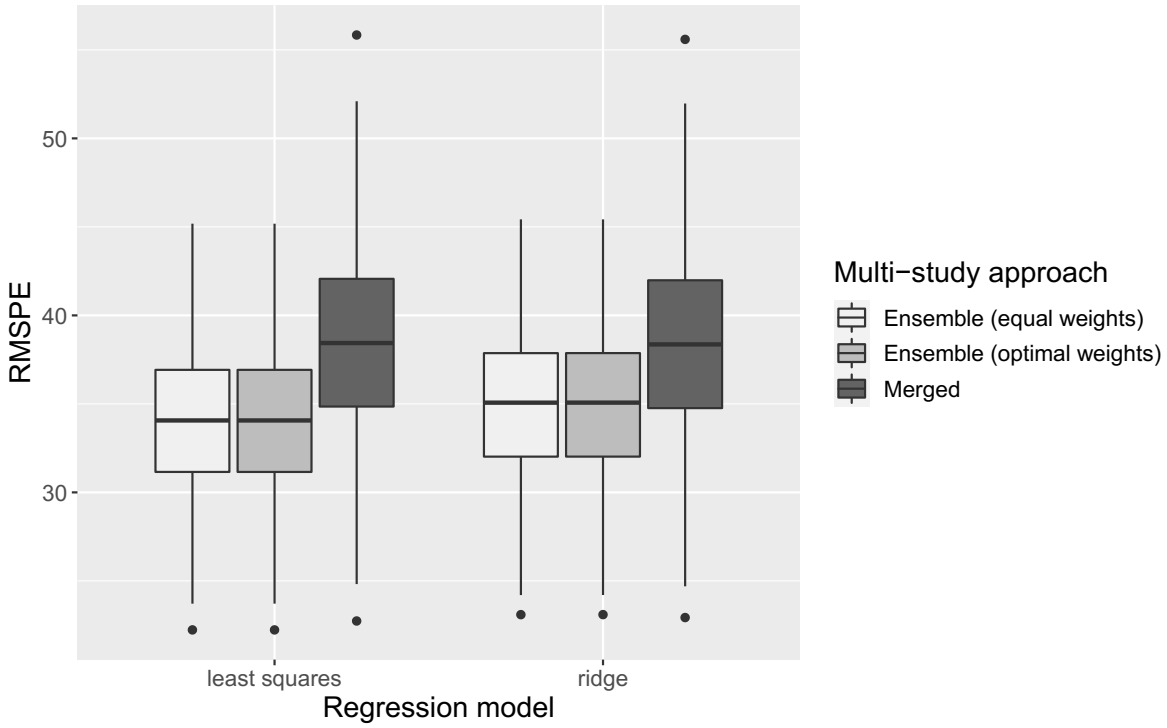

Figure 3.5: Root mean squared prediction error (RMSPE) for the second data illustration scenario with bootstrap confidence intervals.

3.6 Discussion

The availability of large and increasingly heterogeneous collections of data for training classifiers is challenging traditional approaches for training and validating prediction and classification algorithms. At the same time, it is creating opportunities for new and more general paradigms. One of these is multi-study ensemble learning, motivated by variation in the relation between predictors and outcomes across collections of similar studies. A natural benchmark for these methods is to combine all training studies to exploit the power of larger training sample sizes. In previous work (Patil and Parmigiani, 2018), merged learners were shown to perform better than ensemble learners in simulations under low-heterogeneity settings. As heterogeneity increased, however, there was a transition point in the heterogeneity scale beyond which acknowledging cross-study heterogeneity became preferable, and the ensemble learners outperformed the merged learners.

In this paper, we provide the first theoretical investigation of multi-study ensembling. In linear models, we have been able to prove that multi-study ensembling can achieve better optimality

properties than merging all studies prior to training under certain conditions, a question that was still open. We characterized the relative performance of the two approaches as a function of inter-study heterogeneity, explicitly identifying the transition point beyond which ensembling outperforms merging for least squares and ridge regression. We confirmed the analytic results in simulations and demonstrated that when the data are generated by a linear model, the least squares and ridge regression solutions can serve as proxies for the transition point under other learning strategies (lasso, neural network) for which closed-form derivation is difficult. Finally, we estimated the transition point in cases of low and high cross-study heterogeneity in microbiome data and showed how it can be used as a guide for deciding when and when not to merge studies together in the course of learning a prediction rule.

We focused on deriving analytic results for least squares and ridge regression because of the opportunity to pursue closed-form solutions. Other widely used methods such as lasso, neural networks, and random forests are not as easily amenable to a closed-form solution, so we used simulations to study the performance of merging versus ensembling for these methods. In our simulation settings, the merged learners based on least squares, ridge regression, lasso, and neural networks had comparable accuracy, as did the corresponding ensemble learners. The methods all had similar empirical transition points as well, with the exception of random forest, which did not reach a transition point within the specified heterogeneity levels, and also performed worse in general than the other methods, as is expected in data generated by linear models. The analytic results for least squares/ridge regression could potentially serve as an approximation for other methods that perform comparably, though it is important to consider how the reliability of such an approximation could be affected by the nature of the data and choice of model hyperparameters.

In practice, the analytic transition point and transition interval expressions could be used to help guide decisions about whether to merge data from multiple studies when there is potential heterogeneity in predictor-outcome relationships across the study populations. $\overline{\sigma^2}$ can be estimated from the training data and compared to the theoretical transition points or bounds for least squares and/or ridge regression. Various methods can be used to estimate $\overline{\sigma^2}$, including maximum likelihood and method of moments-based approaches used in meta-analysis (for example, see Jackson et al. (2016)), with the caveat that estimates will be imprecise when the number of studies is small.

Under Model (3.1), fitting a correctly specified mixed effects model will generally be more efficient than both the merged and ensemble versions of least squares. However, more flexible machine learning algorithms can potentially yield better prediction accuracy than the true model. For example, in the the low-dimensional simulations, the mixed effects model was outperformed by either the merged learner or ensemble learner based on lasso for most levels of heterogeneity. Moreover, fitting a mixed effects model can be computationally difficult when the number of predictors is large and standard mixed effects models are not appropriate for high-dimensional data, though there are methods for penalized mixed effects models (Bondell et al., 2010; Ibrahim et al., 2011; Schelldorfer et al., 2011; Fan and Li, 2012).

A limitation of our derivations is that they treat the following quantities as known: the subset of predictors with random effects, the ensemble weights, and the regularization parameters for ridge regression. In practice, these are usually selected using statistical procedures that introduce additional variability. Furthermore, we obtained closed-form transition point expressions for cases where the ensemble weighting scheme does not depend on the variances of the random effects. Such weighting schemes are generally sub-optimal (the optimal weights given by Propositions 3.1 and 3.2 depend on $\boldsymbol{G}$), so the closed-form results are based on a conservative estimate of the maximal performance of multi-study ensembling. Another limitation is the assumption that the random effects are uncorrelated, which is often not true in practice.

In summary, although this work is predicated upon the assumption that cross-study heterogeneity manifests as random effects and assumes that weights and regularization parameters are known, we believe it provides a theoretical rationale for multi-study ensembling and a strong foundation for developing practical rules and guidelines to implement it.

References

American Cancer Society (2020). Facts and figures 2020. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2020.html. Accessed: 2020-05-03.

Amir, E., Evans, D., Shenton, A., Lalloo, F., Moran, A., Boggis, C., Wilson, M., and Howell, A. (2003). Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. *Journal of medical genetics* **40,** 807–814.

Anton-Culver, H., Ziogas, A., Bowen, D., Finkelstein, D., Griffin, C., Hanson, J., Isaacs, C., Kasten-Sportes, C., Mineau, G., Nadkarni, P., et al. (2003). The cancer genetics network: recruitment results and pilot studies. *Public Health Genomics* **6,** 171–177.

Antoniou, A., Pharoah, P., McMullan, G., Day, N., Stratton, M., Peto, J., Ponder, B., and Easton, D. (2002). A comprehensive model for familial breast cancer incorporating brca1, brca2 and other genes. *British journal of cancer* **86,** 76–83.

Antoniou, A. C., Cunningham, A., Peto, J., Evans, D., Lalloo, F., Narod, S., Risch, H., Eyfjord, J., Hopper, J., Southey, M., et al. (2008). The boadicea model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *British journal of cancer* **98,** 1457–1466.

Antoniou, A. C., Pharoah, P., Smith, P., and Easton, D. F. (2004). The boadicea model of genetic susceptibility to breast and ovarian cancer. *British journal of cancer* **91,** 1580.

Balmaña, J., Stockwell, D. H., Steyerberg, E. W., Stoffel, E. M., Deffenbaugh, A. M., Reid, J. E., Ward, B., Scholl, T., Hendrickson, B., Tazelaar, J., et al. (2006). Prediction of mlh1 and msh2 mutations in lynch syndrome. *Jama* **296,** 1469–1478.

Banegas, M. P., John, E. M., Slattery, M. L., Gomez, S. L., Yu, M., LaCroix, A. Z., Pee, D., Chlebowski, R. T., Hines, L. M., Thompson, C. A., et al. (2016). Projecting individualized absolute invasive breast cancer risk in us hispanic women. *Journal Of The National Cancer Institute* **109,** djw215.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., and Marron, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics* **20,** 105–114.

Bernau, C., Riester, M., Boulesteix, A.-L., Parmigiani, G., Huttenhower, C., Waldron, L., and Trippa, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30,** i105–i112. PMCID: PMC4058929.

Berry, D., Iversen Jr, E., Gudbjartsson, D., Hiller, E., Garber, J., Peshkin, B., Lerman, C., Watson, P., Lynch, H., Hilsenbeck, S., et al. (2002). Brcapro validation, sensitivity of genetic testing of brca1/brca2, and prevalence of other breast cancer susceptibility genes. *Journal of Clinical Oncology* **20,** 2701–2712.

Berry, D., 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. *Journal of the National Cancer Institute* **89,** 227–237.

Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Biswas, S., Atienza, P., Chipman, J., Hughes, K., Barrera, A. M. G., Amos, C. I., Arun, B., and Parmigiani, G. (2013). Simplifying clinical use of the genetic risk prediction model brcapro. *Breast cancer research and treatment* **139,** 571–579.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66,** 1069–1077.

Braun, D., Gorfine, M., Katki, H. A., Ziogas, A., Anton-Culver, H., and Parmigiani, G. (2014). Extending mendelian risk prediction models to handle misreported family history.

Braun, D., Gorfine, M., Katki, H. A., Ziogas, A., and Parmigiani, G. (2017). Nonparametric adjustment for measurement error in time to event data: Application to risk prediction models. *Journal of the American Statistical Association* .

Bravata, D. M. and Olkin, I. (2001). Simple pooling versus combining in meta-analysis. *Evaluation & the health professions* **24,** 218–230.

Breiman, L. (1996). Bagging predictors. *Machine learning* **24,** 123–140.

Breiman, L. (2001). Random forests. *Machine learning* **45,** 5–32.

Brentnall, A. R. and Cuzick, J. (2019). Risk models for breast cancer and their validation. *arXiv preprint arXiv:1907.02829* .

Brown, P. J. (1977). Centering and scaling in ridge regression. *Technometrics* **19,** 35–36.

Carroll, R. (2006). *Measurement error in nonlinear models: a modern perspective*, volume 105. CRC Press.

Castaldi, P. J., Dahabreh, I. J., and Ioannidis, J. P. (2011). An empirical assessment of validation practices for molecular classifiers. *Briefings in bioinformatics* **12,** 189–202.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, pages 8928–8939.

Chen, J., Bae, E., Zhang, L., Hughes, K., Parmigiani, G., Braun, D., and Rebbeck, T. R. (2020). Penetrance of breast and ovarian cancer in women who carry a brca1/2 mutation and do not use risk-reducing salpingo-oophorectomy: An updated meta-analysis. *JNCI Cancer Spectrum* .

Chen, S. and Parmigiani, G. (2007). Meta-analysis of brca1 and brca2 penetrance. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* **25,** 1329.

Chen, S., Wang, W., Broman, K., Katki, H. A., and Parmigiani, G. (2004a). Bayesmendel: an r environment for mendelian risk prediction.

Chen, S., Wang, W., Broman, K. W., Katki, H. A., and Parmigiani, G. (2004b). Bayesmendel: an r environment for mendelian risk prediction. *Statistical applications in genetics and molecular biology* **3,** 1–19.

Chen, S., Wang, W., Lee, S., Nafa, K., Lee, J., Romans, K., Watson, P., Gruber, S., Euhus, D., Kinzler, K., et al. (2006). Prediction of germline mutations and cancer risk in the lynch syndrome. *JAMA: the journal of the American Medical Association* **296,** 1479.

Chipman, J., Drohan, B., Blackford, A., Parmigiani, G., Hughes, K., and Bosinoff, P. (2013). Providing access to risk prediction tools via the hl7 xml-formatted risk web service. *Breast cancer research and treatment* **140,** 187–193.

Cintolo-Gonzalez, J. A., Braun, D., Blackford, A. L., Mazzola, E., Acar, A., Plichta, J. K., Griffin, M., and Hughes, K. S. (2017). Breast cancer risk models: a comprehensive overview of existing models, validation, and clinical applications. *Breast Cancer Research and Treatment* pages 1–22.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289* .

Collins, I. M., Bickerstaffe, A., Ranaweera, T., Maddumarachchi, S., Keogh, L., Emery, J., Mann, G. B., Butow, P., Weideman, P., Steel, E., et al. (2016). iprevent®: a tailored, web-based, decision support tool for breast cancer risk assessment and management. *Breast cancer research and treatment* **156,** 171–182.

Cunningham, P. and Carney, J. (2000). Diversity versus quality in classification ensembles based on feature selection. In *European Conference on Machine Learning*, pages 109–116. Springer.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems* **2,** 303–314.

Dietterich, T. G. (2000a). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Dietterich, T. G. (2000b). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning* **40,** 139–157.

Dong, Y., Su, H., Zhu, J., and Bao, F. (2017). Towards interpretable deep neural networks by leveraging adversarial examples. *arXiv preprint arXiv:1708.05493* .

Easton, D., Ford, D., and Bishop, D. (1995). Breast and ovarian cancer incidence in brca1-mutation carriers. breast cancer linkage consortium. *American Journal of Human Genetics* **56,** 265.

Easton, D. F. (1999). How many more breast cancer predisposition genes are there? *Breast Cancer Research* **1,** 14.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research* **30,** 207–210.

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

Euhus, D. M., Smith, K. C., Robinson, L., Stucky, A., Olopade, O. I., Cummings, S., Garber, J. E., Chittenden, A., Mills, G. B., Rieger, P., 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,** 844–851.

Fan, Y. and Li, R. (2012). Variable selection in linear mixed effects models. *Annals of statistics* **40,** 2043.

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer.

Gail, M. H. (2019). Performance of bcrat in high-risk patients with breast cancer. *The Lancet Oncology* **20,** e285.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* **81,** 1879–1886.

Gail, M. H., Costantino, J. P., Pee, D., Bondy, M., Newman, L., Selvan, M., Anderson, G. L., Malone, K. E., Marchbanks, P. A., McCaskill-Stevens, W., et al. (2007). Projecting individualized absolute invasive breast cancer risk in african american women. *Journal of the National Cancer Institute* **99,** 1782–1792.

Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M. J., Parmigiani, G., Huttenhower, C., and Waldron, L. (2013). curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database (Oxford)* **2013,** bat013. PMCID: PMC3625954.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation* **4,** 1–58.

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal* **48,** 1029–1040.

Gorgolewski, K., Esteban, O., Schaefer, G., Wandell, B., and Poldrack, R. (2017). Openneuroa free online platform for sharing and analysis of neuroimaging data. *Organization for Human Brain Mapping. Vancouver, Canada* page 1677.

Hansen, F. and Pedersen, G. K. (2003). Jensen's operator inequality. *Bulletin of the London Mathematical Society* **35,** 553–564.

Hechtlinger, Y., Chakravarti, P., and Qin, J. (2017). A generalization of convolutional neural networks to graph-structured data. *arXiv preprint arXiv:1704.08165* .

Heintz-Buschart, A., May, P., Laczny, C. C., Lebrun, L. A., Bellora, C., Krishna, A., Wampach, L., Schneider, J. G., Hogan, A., de Beaufort, C., et al. (2017). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature microbiology* **2,** 16180.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* **29,** 82–97.

Hinton, G. E. (1987). Learning translation invariant recognition in a massively parallel networks. In *International Conference on Parallel Architectures and Languages Europe*, pages 1–13. Springer.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12,** 55–67.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1998). Bayesian model averaging. In *Proceedings of the AAAI workshop on integrating multiple learned models*, volume 335, pages 77–83. Citeseer.

Horner, M., Ries, L., Krapcho, M., Neyman, N., Aminou, R., Howlader, N., Altekruse, S., Feuer, E., Huang, L., Mariotto, A., et al. (2009). Seer cancer statistics review, 1975-2006, national cancer institute. bethesda, md.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks* **4,** 251–257.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67,** 495–503.

Jackson, D., Law, M., Barrett, J. K., Turner, R., Higgins, J. P., Salanti, G., and White, I. R. (2016). Extending dersimonian and laird's methodology to perform network meta-analyses with random inconsistency effects. *Statistics in medicine* **35,** 819–839.

Jackson, D., Riley, R., and White, I. R. (2011). Multivariate meta-analysis: potential and promise. *Statistics in medicine* **30,** 2481–2498.

Jackson, D., White, I. R., and Thompson, S. G. (2010). Extending dersimonian and laird's methodology to perform multivariate random effects meta-analyses. *Statistics in medicine* **29,** 1282–1297.

Jacobi, C. E., de Bock, G. H., Siegerink, B., and van Asperen, C. J. (2009). Differences and similarities in breast cancer risk assessment models in clinical practice: which model to choose? *Breast cancer research and treatment* **115,** 381–390.

Jiang, H., Deng, Y., Chen, H.-S., Tao, L., Sha, Q., Chen, J., Tsai, C.-J., and Zhang, S. (2004). Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC bioinformatics* **5,** 81.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research* **10,** 1391–1445.

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., Nielsen, J., and Bäckhed, F. (2013). Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature* **498,** 99.

Katki, H. (2006). Effect of misreported family history on mendelian mutation prediction models. *Biometrics* **62,** 478–487.

Kerr, K. F., Brown, M. D., Zhu, K., and Janes, H. (2016). Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *Journal of Clinical Oncology* **34,** 2534.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kleinberg, E. (1990). Stochastic discrimination. *Annals of Mathematics and Artificial intelligence* **1,** 207–239.

Kokuer, M., Naguib, R. N., Jancovic, P., Younghusband, H. B., and Green, R. (2006). A comparison of multi-layer neural network and logistic regression in hereditary non-polyposis colorectal cancer risk assessment. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 2417–2420. IEEE.

Kosch, R. and Jung, K. (2018). Conducting gene set tests in meta-analyses of transcriptome expression data. *Research synthesis methods* .

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Krogh, A. and Hertz, J. A. (1992). A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957.

Krogh, A. and Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. In *Advances in neural information processing systems*, pages 231–238.

Kuncheva, L. I. (2002). A theoretical study on six classifier fusion strategies. *IEEE Transactions on pattern analysis and machine intelligence* **24,** 281–286.

Lagani, V., Karozou, A. D., Gomez-Cabrero, D., Silberberg, G., and Tsamardinos, I. (2016). A comparative evaluation of data-merging and meta-analysis methods for reconstructing gene-gene interactions. *BMC bioinformatics* **17,** S194.

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Solís, D. Y., Duque, R., Bersini, H., and Nowé, A. (2012). Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in bioinformatics* **14,** 469–490.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86,** 2278–2324.

Lee, A., Mavaddat, N., Wilcox, A., Cunningham, A., Carver, T., Hartley, S., Babb, d. V. C., Izquierdo, A., Simard, J., Schmidt, M., et al. (2019). Boadicea: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genetics in medicine: official journal of the American College of Medical Genetics* .

Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks* **6,** 861–867.

Liu, Y., Horick, N., Blackford, A., Zhang, Z., Finkelstein, D., and Chen, S. (2013). Brcagail and its validation and comparison to three existing breast cancer risk projection models (brcapro, gail and ibis) in a large us cohort.

Matsuno, R. K., Costantino, J. P., Ziegler, R. G., Anderson, G. L., Li, H., Pee, D., and Gail, M. H. (2011). Projecting individualized absolute invasive breast cancer risk in asian and pacific islander american women. *Journal of the National Cancer Institute* **103,** 951–961.

McCarthy, A. M., Guan, Z., Welch, M., Griffin, M. E., Sippo, D. A., Deng, Z., Coopey, S. B., Acar, A., Semine, A., Parmigiani, G., et al. (2019). Performance of breast cancer risk assessment models in a large mammography cohort. *JNCI: Journal of the National Cancer Institute* .

Metcalfe, K. A., Birenbaum-Carmeli, D., Lubinski, J., Gronwald, J., Lynch, H., Moller, P., Ghadirian, P., Foulkes, W. D., Klijn, J., Friedman, E., et al. (2008). International variation in rates of uptake of preventive options in brca1 and brca2 mutation carriers. *International journal of cancer* **122,** 2017–2022.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M., Ding, W., et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science* pages 66–71.

Milne, R. L. and Antoniou, A. C. (2016). Modifiers of breast and ovarian cancer risks for brca1 and brca2 mutation carriers. *Endocrine-related cancer* **23,** T69–T84.

Ming, C., Viassolo, V., Probst-Hensch, N., Chappuis, P. O., Dinov, I. D., and Katapodi, M. C. (2019). Machine learning techniques for personalized breast cancer risk prediction: comparison with the bcrat and boadicea models. *Breast Cancer Research* **21,** 75.

National Comprehensive Cancer Network (2019). Genetic/familial high-risk assessment: Breast and ovarian (version 3.2019). `https://www2.tri-kobe.org/nccn/guideline/gynecological/english/genetic_familial.pdf`. Accessed: 2020-05-03.

Niepert, M., Ahmed, M., and Kutzkov, K. (2016). Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023.

on Hormonal Factors in Breast Cancer, T. C. G. (2001). Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58 209 women with breast cancer and 101 986 women without the disease. *The Lancet* **358,** 1389–1399.

Opitz, D. and Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research* **11,** 169–198.

Ozanne, E. M., Drohan, B., Bosinoff, P., Semine, A., Jellinek, M., Cronin, C., Millham, F., Dowd, D., Rourke, T., Block, C., et al. (2013). Which risk model to use? clinical implications of the acs mri screening guidelines. *Cancer Epidemiology and Prevention Biomarkers* **22,** 146–149.

Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., et al. (2010). Arrayexpress updatean archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic acids research* **39,** D1002–D1004.

Parmigiani, G., Berry, D., and Aguilar, O. (1998). Determining carrier probabilities for breast cancer-susceptibility genes brca1 and brca2. *The American Journal of Human Genetics* **62,** 145–158.

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., et al. (2017). Accessible, curated metagenomic data through experimenthub. *Nature methods* **14,** 1023.

Patil, P. and Parmigiani, G. (2018). Training replicable predictors in multiple studies. *Proceedings of the National Academy of Sciences* **115,** 2578–2583.

Perrone, M. P. and Cooper, L. N. (1992). When networks disagree: Ensemble methods for hybrid neural networks. Technical report, BROWN UNIV PROVIDENCE RI INST FOR BRAIN AND NEURAL SYSTEMS.

Phillips, K.-A., Liao, Y., Milne, R. L., MacInnis, R. J., Collins, I. M., Buchsbaum, R., Weideman, P. C., Bickerstaffe, A., Nesci, S., Chung, W. K., et al. (2019). Accuracy of risk estimates from the iprevent breast cancer risk assessment and management tool. *JNCI Cancer Spectrum* **3,** pkz066.

Pichert, G., Bolliger, B., Buser, K., and Pagani, O. (2003). Evidence-based management options for women at increased breast/ovarian cancer risk. *Annals of Oncology* **14,** 9–19.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55.

Quante, A. S., Whittemore, A. S., Shriver, T., Strauch, K., and Terry, M. B. (2012). Breast cancer risk assessment across the risk continuum: genetic and nongenetic risk factors contributing to differential model performance. *Breast Cancer Research: BCR* **14,** R144.

Rashid, N. U., Li, Q., Yeh, J. J., and Ibrahim, J. G. (2019). Modeling between-study heterogeneity for improved replicability in gene signature selection and clinical prediction. *Journal of the American Statistical Association* **0,** 1–14.

Riester, M., Taylor, J. M., Feifer, A., Koppie, T., Rosenberg, J. E., Downey, R. J., Bochner, B. H., and Michor, F. (2012). Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer. *Clinical Cancer Research* **18,** 1323–1333.

Rockhill, B., Spiegelman, D., Byrne, C., Hunter, D. J., and Colditz, G. A. (2001). Validation of the gail et al. model of breast cancer risk prediction and implications for chemoprevention. *Journal of the National Cancer Institute* **93,** 358–366.

Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S., et al. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics* **26,** 1651–1686.

Schelldorfer, J., Bühlmann, P., and DE GEER, S. V. (2011). Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics* **38,** 197–214.

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians* **70,** 7–30.

Spiegelman, D., Colditz, G. A., Hunter, D., and Hertzmark, E. (1994). Validation of the gail et al. model for predicting individual breast cancer risk. *JNCI: Journal of the National Cancer Institute* **86,** 600–607.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15,** 1929–1958.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)* **21,** 128.

Sugiyama, M., Krauledat, M., and MÃžller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research* **8,** 985–1005.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics* **60,** 699–746.

Taminau, J., Lazar, C., Meganck, S., and Nowé, A. (2014). Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis. *ISRN bioinformatics* **2014,**.

Team, T. T. D., Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., et al. (2016). Theano: A python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688* .

Terry, M. B., Liao, Y., Whittemore, A. S., Leoce, N., Buchsbaum, R., Zeinomar, N., Dite, G. S., Chung, W. K., Knight, J. A., Southey, M. C., et al. (2019). 10-year performance of four models of breast cancer risk: a validation study. *The Lancet Oncology* **20,** 504–517.

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research* **40,** 3785–3799.

Tsybakov, A. B. (2014). Aggregation and minimax optimality in high-dimensional estimation.

Tyrer, J., Duffy, S. W., and Cuzick, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine* **23,** 1111–1130.

Uno, H., Cai, T., Tian, L., and Wei, L. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102,** 527–537.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. Super learner. *Statistical applications in genetics and molecular biology* **6,**.

Wang, W., Chen, S., Brune, K., Hruban, R., Parmigiani, G., and Klein, A. (2007). Pancpro: risk assessment for individuals with a family history of pancreatic cancer. *Journal of clinical oncology* **25,** 1417–1422.

Wang, W., Niendorf, K. B., Patel, D., Blackford, A., Marroni, F., Sober, A. J., Parmigiani, G., and Tsao, H. (2010). Estimating cdkn2a carrier probability and personalizing cancer risk assessments in hereditary melanoma using melapro. *Cancer research* **70,** 552–559.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks* **5,** 241–259.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., Micklem, G., et al. (1995). Identification of the breast cancer susceptibility gene brca2. *Nature* **378,** 789–792.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2019). A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596* .

Xu, L., Tan, A. C., Winslow, R. L., and Geman, D. (2008). Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC bioinformatics* **9,** 125.

Zhang, Q., Nian Wu, Y., and Zhu, S.-C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836.

Zhang, Y., Bernau, C., Parmigiani, G., and Waldron, L. (2018). The impact of different sources of heterogeneity on loss of accuracy from genomic prediction models. *Biostatistics* page kxy044.

Zhou, H. H., Zhang, Y., Ithapu, V. K., Johnson, S. C., Wahba, G., and Singh, V. (2017). When can multi-site datasets be pooled for regression? hypothesis tests, l_2-consistency and neuroscience applications. *arXiv preprint arXiv:1709.00640* .

Ziogas, A. and Anton-Culver, H. (2003). Validation of family history data in cancer family registries. *American journal of preventive medicine* **24,** 190–198.

Supplementary Materials

S.1 Supplementary Materials for Chapter 1

S.1.1 Penetrance Modification Parameters

Hazard Functions in BRCAPRO and BCRAT

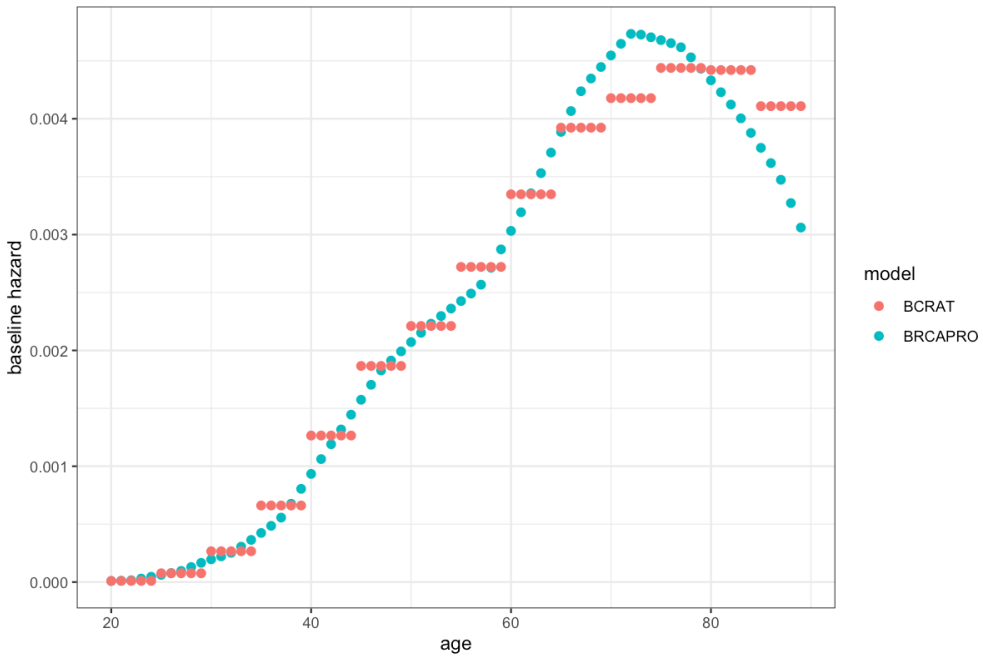

Figure S.1.1: BRCAPRO cause-specific hazard of breast cancer for White female non-carriers ($\lambda_B^0(t)$) and BCRAT cause-specific hazard of breast cancer for White women in the general population ($\tilde{\lambda}_B(t) = \tilde{\lambda}_{B,0}(t)/(1 - AR(t))$).

Population Attributable Risk Estimates for BCRAT Covariates

	White	Black	Hispanic	Asian	Native American
< 50	1.81	1.41	1.37	2.10	1.55
≥ 50	1.96	1.44	1.41	2.43	1.94

Table S.1.1: Estimates of $1 - AR(t)$ from NHIS 2015.

S.1.2 Ensemble Weights From NWH

We fit the model

$$\log \frac{F_B(\tau)}{1 - F_B(\tau)} = \beta_0 + \beta_1 \log(F_B^1(\tau)) + \beta_2 \log(F_B^2(\tau)) + \beta_3 \log(F_B^1(\tau)) \log(F_B^2(\tau)),$$

to the NWH cohort, where $\tau = 5$. The estimated weights are given in Table S.1.2.

Table S.1.2: Ensemble weights.

	Estimate	Standard Error
$\hat{\beta}_0$	2.55	1.31
$\hat{\beta}_1$	0.86	0.32
$\hat{\beta}_2$	1.21	0.28
$\hat{\beta}_3$	0.11	0.06

S.1.3 CGN Characteristics by Center

Table S.1.3: CGN cohort characteristics by center.

Variable	N	Age (median [IQR])	Affected 1st-degree Relatives (%)			Follow-up (median [IQR])	Censored (%)	Cases (%)
Category			0	1	2+			
BAYLOR	69	47 [38, 52]	34 (49.3)	28 (40.6)	7 (10.1)	7.5 [6.1, 8.6]	12 (17.4)	0 (0.0)
COLORADO	1198	51 [41, 64]	528 (44.1)	547 (45.7)	123 (10.3)	7.6 [6.4, 8.5]	85 (7.1)	23 (1.9)
DUKE	286	46 [38, 53]	134 (46.9)	116 (40.6)	36 (12.6)	7.2 [6.2, 8.2]	40 (14.0)	9 (3.1)
EMORY	136	44 [38.8, 51]	56 (41.2)	51 (37.5)	29 (21.3)	7.2 [6.6, 8.3]	29 (21.3)	3 (2.2)
GEORGETOWN	309	43 [35, 52]	107 (34.6)	147 (47.6)	55 (17.8)	7.8 [6.3, 8.5]	81 (26.2)	2 (0.6)
JH	469	47 [39, 56]	279 (59.5)	148 (31.6)	42 (9.0)	8.0 [6.1, 9.0]	73 (15.6)	10 (2.1)
MDAND	295	45 [37, 53]	215 (72.9)	64 (21.7)	16 (5.4)	6.1 [4.2, 7.0]	87 (29.5)	4 (1.4)
UCI	608	48 [37, 59]	352 (57.9)	223 (36.7)	33 (5.4)	5.3 [4.0, 7.1]	209 (34.4)	4 (0.7)
UNC	229	46 [39, 53]	80 (34.9)	111 (48.5)	38 (16.6)	8.0 [7.1, 9.0]	28 (12.2)	6 (2.6)
UNM	324	[41, 63]	160 (49.4)	123 (38.0)	41 (12.7)	6.6 [6.0, 7.5]	43 (13.3)	11 (3.4)
UPENN	540	45 [37, 53]	297 (55.0)	185 (34.3)	58 (10.7)	8.1 [6.6, 9.1]	76 (14.1)	8 (1.5)
UTAH	880	47 [35, 61]	522 (59.3)	298 (33.9)	60 (6.8)	7.3 [5.4, 8.0]	61 (6.9)	14 (1.6)
UTSA	92	43 [35.8, 52]	29 (31.5)	48 (52.2)	15 (16.3)	5.1 [4.0, 6.5]	36 (39.1)	1 (1.1)
UTSW	247	41 [33.5, 47]	88 (35.6)	116 (47.0)	43 (17.4)	6.6 [5.6, 7.6]	30 (12.1)	4 (1.6)
UWASH	1632	46 [37, 56]	1290 (79.0)	291 (17.8)	51 (3.1)	7.6 [6.9, 8.3]	44 (2.7)	13 (0.8)

S.2 Supplementary Materials for Chapter 2

S.2.1 Methods

Notation Table

Table S.2.1: Notation table for Chapter 2.

Variable	Description
R	number of relatives (besides the proband)
r	indexes family members ($r = 0$ for the proband)
H_{ri}	ith feature for family member r
H_r	vector of features for family member r
K	number of features per family member
A_r	indices of r's parents
H	family history matrix
Y_0	proband's outcome
Notation for NN	
X	vector input to NN
L	number of hidden layers in NN
l	indexes NN layers ($l = 0$ for input layer, $l = L + 1$ for output layer)
a^l	output of layer l
w	NN weight parameter
b	NN bias parameter
ϕ	activation function
σ	logistic function
C	cost function
N_T	training sample size
Q'	number of relative types in reference pedigree
R'_q	number of relatives of type q in reference pedigree
R'	number of relatives in reference pedigree (besides the proband)
H'	standardized version of H that has the same structure as the reference pedigree
R_q	number of relatives of type q in actual pedigree
N_l	number of nodes in fully-connected layer l
M_l	number of filters in layer l
$\mathcal{N}(r)$	neighborhood about member r, consisting of r and r's first-degree relatives
m_i	number of relatives of type i ($i = 1, 2, 3, 4$ for sisters, brothers, daughters, sons) in neighborhood
U	neighborhood size ($U = 3 + \sum_{i=1}^{4} m_i$)

Standard CNNs

A NN is a CNN if it contains at least one convolutional layer. A convolutional layer applies the same functions repeatedly to different fixed-size regions of the layer input (for example, sets of pixels in images). These functions are called convolutional filters.

Using the same notation as in Section 2.2.2 (see also Table S.2.1) unless otherwise specified, we describe a CNN that takes as input a fixed-length vector X and outputs a predicted probability. Suppose the L hidden layers are all convolutional layers. Let convolutional layer l have M_l real-valued convolutional filters $f_1^l, \ldots, f_{M_l}^l$ that act on sliding windows of the input to the layer. The output of layer l will be a matrix with M_l columns and N_l rows, where N_l is the number of windows to which the M_l filters are applied. N_l is determined by two pre-specified quantities: the length (number of rows) of each window, which we denote by U_l, and the distance by which we slide (along the row axis) to move from one window to the next, which we denote by S_l (typically called the stride). Therefore, $N_l = \lfloor (N_{l-1} - U_l)/S_l \rfloor$. Let N_0 be the length of X. Let $M_0 = 1$.

For $i = 1, \ldots, M_l$, $f_i^l : \mathbb{R}^{U_l * M_{l-1}} \to \mathbb{R}$. Let the flattened version of the jth window for layer l (obtained by concatenating the rows of the window) be denoted by $v_j^{l-1} \in \mathbb{R}^{U_l * M_{l-1}}$. Each filter f_i^l takes a weighted sum of the elements of v_j^{l-1} and adds a bias term, then applies an activation function. The output from applying filter i to the jth window is

$$f_i^l(v_j^{l-1}) = \phi\left(w_i^l v_j^{l-1} + b_i^l\right)$$

where $w_i^l \in \mathbb{R}^{U_l * M_{l-1}}$ is the vector of weights and $b_i^l \in \mathbb{R}$ is the bias for filter i.

Let $f^l = (f_1^l, \ldots, f_{M_l}^l) : \mathbb{R}^{U_l * M_{l-1}} \to \mathbb{R}^{M_l}$. Then the outputs of layers $0, \ldots, L$ are

$$\begin{aligned} a^0 &= X, \\ a^l &= \left[f^l(v_1^{l-1}), \ldots, f^l(v_{N_l}^{l-1})\right]^T, \qquad (l = 1, \ldots, L), \end{aligned}$$

where $a^l \in \mathbb{R}^{N_l \times M_l}$.

The output of the final ($l = L$) convolutional layer is flattened into a vector of length $M_L \times N_L$ before being passed to the fully-connected output layer. The calculation of the prediction in the output layer and the optimization of the weight and bias parameters proceed in the same way as for FCNNs.

When $S_l > 1$, the sliding windows might not exactly cover the length of the input. To resolve this issue, CNNs are often combined with zero padding, which involves adding zeros to the borders

of the layer input. We use a similar approach to apply convolutions to pedigrees (described in Section 2.2.4).

Proof of Theorem 2.1

Let N_r^d be the number of relatives of r of degree d. We first show by induction that for $d \geq 0$, there exists a CNN with d convolutional layers such that for any relative r, $a_r^d \in \mathbb{R}^{(K+1)\sum_{i=0^d} N_r^i}$ (r's output vector from layer d), is a continuous and invertible transformation of the original features H_s' of r's relatives s of degree$\leq d$.

For $d = 0$, the result holds trivially since $a_r^0 = H_r'$. Let $d > 0$ and consider a CNN with $d-1$ convolutional layers that satisfies the statement for $d-1$. Any degree d relative of r is a degree $d-1$ relative of some first-degree relative of r. Therefore, by the induction assumption, there is a subset of $a_{\mathcal{N}(r)}^{d-1}$ that is a continuous and invertible transformation of the original features of relatives of r of degree $\leq d$ (the subset has size $(K+1)\sum_{i=0^d} N_r^i$). Add a dth convolutional layer where $M_d = (K+1)\sum_{i=0^d} N_r^i$ and define the ith component of f^d to be the application of activation function ϕ to the ith element of the subset (setting the weight for the ith element to 1 and the remaining weights to 0). Since ϕ is continuous and invertible, $a_r^d = f^d(a_{\mathcal{N}(r)}^{d-1})$ is a continuous, invertible transformation of the original features of relatives of r of degree $\leq d$.

By the result above, there exists a CNN with $d-1$ convolutional layers such that a_r^{d-1} is a continuous and invertible transformation of the original features of relatives of r of degree $\leq d-1$. It follows that $V = a_{\mathcal{N}(0)}^{d-1} \in \mathbb{R}^{U \times M_{d-1}}$ (obtained by concatenating the outputs from layer $d-1$ for the proband's neighborhood) is a continuous and invertible transformation of the original features X, since any degree d relative is a degree $d-1$ relative of some first-degree relative. Let tau : $\mathbb{R}^{U(R'+1)(K+1)} \to \mathbb{R}^{M_{d-1}}$ denote the continuous and invertible transformation, so that $V = \tau(X)$.

Suppose we add another convolutional layer with M_d filters. Then computing a_0^d is equivalent to applying a fully-connected layer with UM_d nodes to V. Since the original features are bounded and τ is continuous, V is contained in a compact subset of $\mathbb{R}^{U \times M_{d-1}}$. The logistic function σ is uniformly continuous on any compact subset of $\mathbb{R}$, so there exists $\delta > 0$ such that

$$|X_1 - X_2| < \delta \implies |\sigma(X_1) - \sigma(X_2)| < \epsilon$$

By the universality theorem for FCNNs, there exists a value of M_d and parameters for layer d such that

$$|a_0^d - \sigma^{-1}(g(X))| < \delta$$

and

$$|\sigma(a_0^d) - g(X)| < \epsilon$$

While the Universal Approximation Theorem for FCNN states that a given continuous function with compact support can be approximated by a single-layer FCNN with a finite number of neurons, N, to any degree of precision, it does not provide an upper bound for N. In practice, N might be very large. Similarly, our theorem states that a given continuous function with compact support can be approximated by a CNN with d convolutional layers, but does not provide an upper bound for the number of convolutional filters M_d for the final convolutional layer d (we showed above that the theorem is satisfied by a CNN with M_d proportional to N).

S.2.2 Simulations and Data Application

Misreporting Rates

For the second simulation setting, we perturbed the family histories using misreporting rates for breast and ovarian cancer from Ziogas and Anton-Culver (2003) (Table S.2.2).

Table S.2.2: Accuracy of probands' reported family history of breast and ovarian cancer. FNR: False negative rate. FPR: False positive rate.

Degree	FNR	FPR
Breast Cancer		
1	0.05	0.03
2	0.18	0.03
Ovarian Cancer		
1	0.17	0.01
2	0.56	0.02

We perturbed diagnosis ages based on the rates reported in Braun et al. (2017): 3% of breast diagnosis ages and 4% of ovarian cancer diagnosis ages were misreported. We assumed the difference between the true age and misreported age has mean 4 and standard deviation 3.

Reference Pedigree Structure

In the simulations, we sampled family structures from the CGN, so we chose a reference pedigree structure based on the relative counts in the CGN. We only used first- and second-degree relatives since studies have shown that excluding third- and higher-degree relatives has little effect on the performance of family history-based models (Biswas et al., 2013; Terry et al., 2019). In the CGN, the third quartile for the number of children of the proband's maternal/paternal grandparents was 6, the third quartile for the number of children of the proband's parents was 5, and the third quartile for the number of children of the proband was 3. We used these numbers to define a symmetric family structure where each couple has an equal number of sons and daughters: the proband's grandparents each have 3 sons and 3 daughters, the proband's parents have 3 sons and 3 daughters, and the proband has 2 sons and 2 daughters. We used a slightly larger family size (26) than the third quartile for family size (21) in the CGN in order to capture more of the family history.

In the data application, we reduced the size of the reference family to 19 because the Risk Service families (median size 8, IQR 7-14) were smaller than the CGN families (median size 16, IQR 12-21). We excluded sons and daughters of the proband from the reference structure because only 18% of Risk Service probands had a son or daughter in their pedigree (versus 72% in the CGN).

Missing Age Information in the Risk Service

Ages and diagnoses ages were available for all but 185 probands, who were excluded from the training set. If an affected first- or second-degree relative was missing age or age at diagnosis, then we set their baseline cancer status as affected. One justification for this imputation approach is that probands may have more accurate memory of recent diagnoses (i.e. within the last 5 years) among relatives than less recent diagnoses. Therefore, the missing diagnoses ages are more likely to correspond to less recent diagnoses.

We considered including current age and age at diagnosis in the models trained on the Risk Service (and adding a missing indicator for each age variable), but in cross-validation on the Risk Service, the models without age information performed better than the models with age information, so we excluded the age information (other than the proband's current age).

S.3 Supplementary Materials for Chapter 3

S.3.1 Notation Table

Table S.3.1: Notation table for Chapter 3.

Variable	Value / Description
K	number of training studies
n_k	sample size in study k ($k = 0$ corresponds to test dataset)
$\boldsymbol{Y_k}$	outcome vector for study k
$\boldsymbol{X_k}$	fixed effects design matrix for study k
$\boldsymbol{Y}$	$(\boldsymbol{Y_1}^T, \ldots, \boldsymbol{Y_K}^T)^T$; outcome vector for merged dataset
$\boldsymbol{X}$	$\left[\boldsymbol{X_1}^T \vert \ldots \vert \boldsymbol{X_K}^T\right]^T$; design matrix for merged dataset
p	number of predictors (including the intercept)
q	number of predictors with random effects
r	number of unique variance values for random effects
$\boldsymbol{\Gamma}$	$(\boldsymbol{\Gamma})_{il} = 1$ if predictor i is the lth predictor with a random effect and $(\boldsymbol{\Gamma})_{il} = 0$ otherwise; matrix such that $\boldsymbol{X_k\Gamma}$ subsets $\boldsymbol{X_k}$ to the columns with random effects
$\mathbf{Z_k}$	$\mathbf{X_k}\Gamma$; random effects design matrix for study k
$\boldsymbol{\beta}$	vector of fixed effects
$\boldsymbol{\gamma_k}$	vector of random effects
σ_i^2	variance of random effect i
$\boldsymbol{G}$	$diag(\sigma_1^2, \ldots, \sigma_q^2)$; covariance matrix for random effects
$\overline{\sigma^2}$	$tr(\mathbf{G})/p$; heterogeneity summary measure
$\boldsymbol{\epsilon_k}$	vector of residual errors for study k
σ_ϵ^2	variance of residuals
$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}$	$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$; merged least squares estimator
w_k	ensemble weight for study k
$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}}$	$\sum_k w_k(\boldsymbol{X_k}^T\boldsymbol{X_k})^{-1}\boldsymbol{X_k}^T\boldsymbol{Y_k}$; ensemble least squares estimator
$\boldsymbol{S_k}$	$\boldsymbol{I_p}$ if scaling is unnecessary, diagonal matrix with $(\boldsymbol{S_k})_{jj}$ = inverse standard deviation of jth column of $\boldsymbol{X_k}$ otherwise; scaling matrix for study k
$\boldsymbol{S}$	$\boldsymbol{I_p}$ if scaling is unnecessary, diagonal matrix with $(\boldsymbol{S})_{jj}$ = inverse standard deviation of jth column of $\boldsymbol{X}$ otherwise; scaling matrix for merged data
$\tilde{\boldsymbol{X}}_{\boldsymbol{k}}$	$\boldsymbol{S_kX_k}$; ridge design matrix for study k
$\tilde{\boldsymbol{X}}$	$\boldsymbol{SX}$; ridge design matrix for merged dataset
$\boldsymbol{I_p^-}$	$\boldsymbol{I_p}$ with $(\boldsymbol{I_p})_{11} = 0$
$\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}}$	$\boldsymbol{S}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + \lambda\boldsymbol{I_p^-})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{Y}$; merged ridge estimator of $\boldsymbol{\beta}$
$\hat{\boldsymbol{\beta}}_{\boldsymbol{R,k}}$	$\boldsymbol{S_k}(\tilde{\boldsymbol{X}}_{\boldsymbol{k}}^T\tilde{\boldsymbol{X}}_{\boldsymbol{k}} + \lambda_k\boldsymbol{I_p^-})^{-1}\tilde{\boldsymbol{X}}_{\boldsymbol{k}}^T\boldsymbol{Y_k}$; ridge estimator of $\boldsymbol{\beta}$ based on study k
$\hat{\boldsymbol{\beta}}_{\boldsymbol{R,E}}$	$\sum_{k=1}^K w_k\hat{\boldsymbol{\beta}}_{\boldsymbol{R,k}}$; ensemble ridge estimator of $\boldsymbol{\beta}$
$\boldsymbol{R_k}$	$\boldsymbol{X_k}^T\boldsymbol{X_k}$
$\boldsymbol{R}$	$\boldsymbol{X}^T\boldsymbol{X}$
$\boldsymbol{M_k}$	$\boldsymbol{R_k} + \lambda_k\boldsymbol{I_p^-}\boldsymbol{S_k}^{-2}$
$\boldsymbol{M}$	$\boldsymbol{R} + \lambda\boldsymbol{I_p^-}\boldsymbol{S}^{-2}$
$\boldsymbol{\Gamma_{(j)}}$	$(\boldsymbol{\Gamma_{(j)}})_{il} = 1$ if random effect i is the lth random effect with variance $\sigma_{(j)}^2$ and $(\boldsymbol{\Gamma_{(j)}})_{il} = 0$ otherwise; matrix such that $\boldsymbol{G\Gamma_{(j)}}$ subsets $\boldsymbol{G}$ to the columns corresponding to $\sigma_{(j)}^2$
τ_{LS}	transition point for least squares (Theorem 3.1)
$\tau_{LS,1}$	lower bound of transition interval for least squares (Theorem 3.2)
$\tau_{LS,2}$	upper bound of transition interval for least squares (Theorem 3.2)
$\boldsymbol{b_E}$	$Bias(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,E}}) = -\sum_k w_k\lambda_k\boldsymbol{X_0}\boldsymbol{M_k}^{-1}\boldsymbol{I_p^-}\boldsymbol{S_k}^{-2}\boldsymbol{\beta}$
$\boldsymbol{b_k}$	$Bias(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,k}}) = \lambda_k\boldsymbol{X_0}\boldsymbol{M_k}^{-1}\boldsymbol{I_p^-}\boldsymbol{S_k}^{-2}\boldsymbol{\beta}$
$\boldsymbol{b_M}$	$Bias(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}}) = -\lambda\boldsymbol{X_0}\boldsymbol{M}^{-1}\boldsymbol{I_p^-}\boldsymbol{S}^{-2}\boldsymbol{\beta}$
τ_R	transition point for ridge regression (Theorem 3.3)
$\tau_{R,1}$	lower bound of transition interval for ridge regression (Theorem 3.4)
$\tau_{R,2}$	upper bound of transition interval for ridge regression (Theorem 3.4)
v_k	$tr(\boldsymbol{M_k}^{-1}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{G}\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M_k}\boldsymbol{R_0}) + \sigma_\epsilon^2 tr(\boldsymbol{M_k}^{-1}\boldsymbol{X_k}^T\boldsymbol{X_k}\boldsymbol{M_k}^{-1}\mathbf{R_0})$

S.3.2 Calculations and Proofs

Covariance Matrices and MSPEs

$$Cov(\boldsymbol{Y_k}) = \boldsymbol{Z_k G Z_k^T} + \sigma_\epsilon^2 \boldsymbol{I_{n_k}}$$

$$Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,k}}) = \boldsymbol{\Gamma G \Gamma}^T + \sigma_\epsilon^2 (\boldsymbol{X_k^T X_k})^{-1}$$

$$Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}}) = \sum_{k=1}^{K} w_k^2 \boldsymbol{\Gamma G \Gamma}^T + \sigma_\epsilon^2 \sum_{k=1}^{K} w_k^2 (\boldsymbol{X_k^T X_k})^{-1}$$

$$Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1} \sum_{k=1}^{K} [\boldsymbol{X_k^T Z_k G Z_k^T X_k}](\boldsymbol{X}^T\boldsymbol{X})^{-1} + \sigma_\epsilon^2 (\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$$\begin{aligned}
E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}}\|_2^2] &= tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}})) \\
&= \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{\Gamma G \Gamma}^T \boldsymbol{X_0^T X_0}) + \sigma_\epsilon^2 \sum_{k=1}^{K} w_k^2 tr((\boldsymbol{X_k^T X_k})^{-1} \boldsymbol{X_0^T X_0}) \\
&= \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{G Z_0^T Z_0}) + \sigma_\epsilon^2 \sum_{k=1}^{K} w_k^2 tr((\boldsymbol{X_k^T X_k})^{-1} \boldsymbol{X_0^T X_0}) \\
&= \sum_{j=1}^{r} \sigma_{(j)}^2 tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}) \sum_{k=1}^{K} w_k^2 + \sigma_\epsilon^2 \sum_{k=1}^{K} w_k^2 tr((\boldsymbol{X_k^T X_k})^{-1} \boldsymbol{X_0^T X_0})
\end{aligned}$$

$$
\begin{aligned}
E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}\|_2^2] &= tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}})) \\
&= tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\sum_{k=1}^{K}[\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{G}\boldsymbol{Z_k}^T\boldsymbol{X_k}](\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0}^T\boldsymbol{X_0})+ \\
&\quad \sigma_\epsilon^2 tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0}^T\boldsymbol{X_0}) \\
&= tr(\boldsymbol{G}\sum_{k=1}^{K}\boldsymbol{Z_k}^T\boldsymbol{X_k}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0}^T\boldsymbol{X_0}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_k}^T\boldsymbol{Z_k})+ \\
&\quad \sigma_\epsilon^2 tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0}^T\boldsymbol{X_0}) \\
&= \sum_{j=1}^{r}\sigma_{(j)}^2 tr(\sum_{k=1}^{K}\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0}^T\boldsymbol{X_0}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma_{(j)}})+ \\
&\quad \sigma_\epsilon^2 tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0}^T\boldsymbol{X_0}) \\
&= \sum_{j=1}^{r}\sigma_{(j)}^2 tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\sum_{k=1}^{K}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma_{(j)}}\boldsymbol{\Gamma_{(j)}}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0}^T\boldsymbol{X_0})+ \\
&\quad \sigma_\epsilon^2 tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0}^T\boldsymbol{X_0})
\end{aligned}
$$

$$
Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{R,k}}) = \boldsymbol{M_k}^{-1}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{G}\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M_k}^{-1} + \sigma_\epsilon^2\boldsymbol{M_k}^{-1}\boldsymbol{R_k}\boldsymbol{M_k}^{-1}
$$

$$
Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{R,E}}) = \sum_k w_k^2\boldsymbol{M_k}^{-1}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{G}\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M_k}^{-1} + \sigma_\epsilon^2\sum_k w_k^2\boldsymbol{M_k}^{-1}\boldsymbol{R_k}\boldsymbol{M_k}^{-1}
$$

$$
Bias(\hat{\boldsymbol{\beta}}_{\boldsymbol{R,E}}) = -(\sum_k w_k\lambda_k\boldsymbol{M_k}^{-1}\boldsymbol{I_p^-}\boldsymbol{S_k}^{-2})\boldsymbol{\beta}
$$

$$
Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}}) = \boldsymbol{M}^{-1}\sum_k \boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{G}\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M}^{-1} + \sigma_\epsilon^2\boldsymbol{M}^{-1}\boldsymbol{R}\boldsymbol{M}^{-1}
$$

$$
Bias(\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}}) = -\lambda(\boldsymbol{M}^{-1}\boldsymbol{I_p^-}\boldsymbol{S}^{-2})\boldsymbol{\beta}
$$

$$
\begin{aligned}
E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,E}}\|_2^2] &= tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,E}})) + \boldsymbol{b}_{\boldsymbol{E}}^T\boldsymbol{b}_{\boldsymbol{E}} \\
&= \sum_k w_k^2 tr(\boldsymbol{M}_{\boldsymbol{k}}^{-1}\boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{Z}_{\boldsymbol{k}}\boldsymbol{G}\boldsymbol{Z}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}_{\boldsymbol{k}}\boldsymbol{R_0}) + \\
&\quad \sigma_\epsilon^2 \sum_k w_k^2 tr(\boldsymbol{M}_{\boldsymbol{k}}^{-1}\boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}_{\boldsymbol{k}}^{-1}\mathbf{R_0}) + \boldsymbol{b}_{\boldsymbol{E}}^T\boldsymbol{b}_{\boldsymbol{E}} \\
&= \sum_k w_k^2 tr(\boldsymbol{G}\boldsymbol{Z}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}_{\boldsymbol{k}}^{-1}\boldsymbol{R_0}\boldsymbol{M}_{\boldsymbol{k}}^{-1}\boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{Z}_{\boldsymbol{k}}) + \\
&\quad \sigma_\epsilon^2 \sum_k w_k^2 tr(\boldsymbol{M}_{\boldsymbol{k}}^{-1}\boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}_{\boldsymbol{k}}^{-1}\mathbf{R_0}) + \boldsymbol{b}_{\boldsymbol{E}}^T\boldsymbol{b}_{\boldsymbol{E}} \\
&= \sum_{j=1}^r \sigma_{(j)}^2 \sum_k w_k^2 tr(\boldsymbol{\Gamma}_{\boldsymbol{(j)}}^T\boldsymbol{Z}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}_{\boldsymbol{k}}^{-1}\boldsymbol{R_0}\boldsymbol{M}_{\boldsymbol{k}}^{-1}\boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{Z}_{\boldsymbol{k}}\boldsymbol{\Gamma}_{\boldsymbol{(j)}}) + \\
&\quad \sigma_\epsilon^2 \sum_k w_k^2 tr(\boldsymbol{M}_{\boldsymbol{k}}^{-1}\boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}_{\boldsymbol{k}}^{-1}\mathbf{R_0}) + \boldsymbol{b}_{\boldsymbol{E}}^T\boldsymbol{b}_{\boldsymbol{E}}
\end{aligned}
$$

$$
\begin{aligned}
E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}}\|_2^2] &= tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}})) + \boldsymbol{b}_{\boldsymbol{M}}^T\boldsymbol{b}_{\boldsymbol{M}} \\
&= tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{Z}_{\boldsymbol{k}}\boldsymbol{G}\boldsymbol{Z}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}^{-1}\boldsymbol{R_0}) + \\
&\quad \sigma_\epsilon^2 tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}^{-1}\boldsymbol{R_0}) + \boldsymbol{b}_{\boldsymbol{M}}^T\boldsymbol{b}_{\boldsymbol{M}} \\
&= \sum_k tr(\boldsymbol{G}\boldsymbol{Z}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}^{-1}\boldsymbol{R_0}\boldsymbol{M}^{-1}\boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{Z}_{\boldsymbol{k}}) + \\
&\quad \sigma_\epsilon^2 tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}^{-1}\boldsymbol{R_0}) + \boldsymbol{b}_{\boldsymbol{M}}^T\boldsymbol{b}_{\boldsymbol{M}} \\
&= \sum_{j=1}^r \sigma_{(j)}^2 \sum_k tr(\boldsymbol{\Gamma}_{\boldsymbol{(j)}}^T\boldsymbol{Z}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}^{-1}\boldsymbol{R_0}\boldsymbol{M}^{-1}\boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{Z}_{\boldsymbol{k}}\boldsymbol{\Gamma}_{\boldsymbol{(j)}}) + \\
&\quad \sigma_\epsilon^2 tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}^{-1}\boldsymbol{R_0}) + \boldsymbol{b}_{\boldsymbol{M}}^T\boldsymbol{b}_{\boldsymbol{M}} \\
&= \sum_{j=1}^r \sigma_{(j)}^2 tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{Z}_{\boldsymbol{k}}\boldsymbol{\Gamma}_{\boldsymbol{(j)}}\boldsymbol{\Gamma}_{\boldsymbol{(j)}}^T\boldsymbol{Z}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}^{-1}\boldsymbol{R_0}) + \\
&\quad \sigma_\epsilon^2 tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{M}^{-1}\boldsymbol{R_0}) + \boldsymbol{b}_{\boldsymbol{M}}^T\boldsymbol{b}_{\boldsymbol{M}}
\end{aligned}
$$

Proof of Theorem 3.4

Theorems 3.1-3.3 are special cases of Theorem 3.4, so we provide a proof for Theorem 3.4.

$$\overline{\sigma^2}=\frac{\sum_{j=1}^{r} m_j\sigma^2_{(j)}}{p}$$

$$\leq\frac{\sigma^2_\epsilon\left(\sum_k w_k^2 tr(\boldsymbol{M_k^{-1}R_kM_k^{-1}R_0})-tr(\boldsymbol{M^{-1}}\sum_k\boldsymbol{R_kM^{-1}R_0})\right)+\boldsymbol{b_E^Tb_E}-\boldsymbol{b_M^Tb_M}}{p\max_{j=1,\ldots,r}\frac{1}{m_j}(tr(\boldsymbol{M^{-1}}\sum_k\boldsymbol{X_k^TZ_k\Gamma_{(j)}\Gamma_{(j)}^TZ_k^TX_kM^{-1}R_0})-\sum_k w_k^2 tr(\boldsymbol{\Gamma_{(j)}^TZ_k^TX_kM_kR_0M_k^{-1}X_k^TZ_k\Gamma_{(j)}})}$$

$$=\tau_{R,1}$$

$$\Longrightarrow\sum_{j=1}^{r}\sigma^2_{(j)}(tr(\boldsymbol{M^{-1}}\sum_k\boldsymbol{X_k^TZ_k\Gamma_{(j)}\Gamma_{(j)}^TZ_k^TX_kM^{-1}R_0})-\sum_k w_k^2 tr(\boldsymbol{M_k^{-1}X_k^TZ_k\Gamma_{(j)}\Gamma_{(j)}^TZ_k^TX_kM_k^{-1}R_0}))$$

$$\leq\sigma^2_\epsilon\left(\sum_k w_k^2 tr(\boldsymbol{M_k^{-1}R_kM_kR_0})-tr(\boldsymbol{M^{-1}}\sum_k\boldsymbol{R_kM^{-1}R_0})\right)+\boldsymbol{b_E^Tb_E}-\boldsymbol{b_M^Tb_M}$$

(assuming Condition 3.22 holds)

$$\Longleftrightarrow E[\|\boldsymbol{Y_0}-\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}}\|_2^2]\leq E[\|\boldsymbol{Y_0}-\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,E}}\|_2^2]$$

Proofs of Propositions 3.1 and 3.2 (Optimal Weights)

Least Squares. For least squares, we want to minimize

$$\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{GZ_0^TZ_0})+\sigma^2_\epsilon\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{R_k^{-1}R_0})$$

subject to the constraint $\sum_k w_k = 1$.

Using Lagrange multipliers, we get the system of equations

$$2w_k\left(tr(\boldsymbol{GZ_0^TZ_0})+\sigma^2_\epsilon tr(\boldsymbol{R_k^{-1}R_0})\right)=\alpha \qquad (\text{for } k=1,\ldots,K;\ \alpha \text{ is the Lagrange multiplier})$$

$$\sum_k w_k = 1$$

which is solved by

$$w_k=\frac{\left(tr(\boldsymbol{GZ_0^TZ_0})+\sigma^2_\epsilon tr(\boldsymbol{R_k^{-1}R_0})\right)^{-1}}{\sum_k\left(tr(\boldsymbol{GZ_0^TZ_0})+\sigma^2_\epsilon tr(\boldsymbol{R_k^{-1}R_0})\right)^{-1}}$$

Plugging the optimal weights into the MSPE of the ensemble learner, $tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}}))$, we get

$$tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}}))=\frac{1}{\sum_k\left(tr(\boldsymbol{GZ_0^TZ_0})+\sigma^2_\epsilon tr(\boldsymbol{R_k^{-1}R_0})\right)^{-1}}$$

In the equal variances setting,

$$tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}})) = \frac{1}{\sum_k \left(\sigma^2 tr(\boldsymbol{Z_0^T Z_0}) + \sigma_\epsilon^2 tr(\boldsymbol{R_k^{-1} R_0})\right)^{-1}}$$

so $tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,E}}))$ is approximately linear in σ^2 when $tr(\boldsymbol{R_k^{-1} R_0})$ is similar across studies or when σ^2 is large.

Ridge Regression. For ridge regression, let

$$v_k = tr(\boldsymbol{M_k^{-1} X_k^T Z_k G Z_k^T X_k M_k R_0}) + \sigma_\epsilon^2 tr(\boldsymbol{M_k^{-1} X_k^T X_k M_k^{-1}} \mathbf{R_0})$$

$$\boldsymbol{b_k} = \lambda_k \boldsymbol{X_0 M_k^{-1} I_p^- S_k^{-2}} \beta.$$

We want to minimize

$$\sum_k w_k^2 c_k + \left(\sum_k w_k \boldsymbol{b_k}\right)^T \left(\sum_k w_k \boldsymbol{b_k}\right)$$

subject to the constraint $\sum_k w_k = 1$.

Using Lagrange multipliers, we get the system of equations

$$2w_k(v_k + \boldsymbol{b_k^T b_k}) + 2\sum_{j \neq k} w_j \boldsymbol{b_j^T b_k} = \alpha \qquad \text{(for } k = 1, \ldots, K\text{)}$$

$$\sum_k w_k = 1.$$

The solution to the system is

$$\begin{bmatrix} \boldsymbol{w} \\ \alpha \end{bmatrix} = \begin{bmatrix} \boldsymbol{C} & -\boldsymbol{j} \\ \boldsymbol{j}^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix} \tag{27}$$

where $\boldsymbol{C}$ has entries $(\boldsymbol{C})_{kk} = (v_k + \boldsymbol{b_k^T b_k})$ and $(\boldsymbol{C})_{jk} = \boldsymbol{b_j^T b_k}$ for $j \neq k$ and $\boldsymbol{j} = (1, \ldots, 1)^T \in \mathbb{R}^K$.

$$\boldsymbol{w} = \boldsymbol{C^{-1} J} (\boldsymbol{j^T C^{-1} j})^{-1} \tag{28}$$

S.3.3 Additional Simulation Results

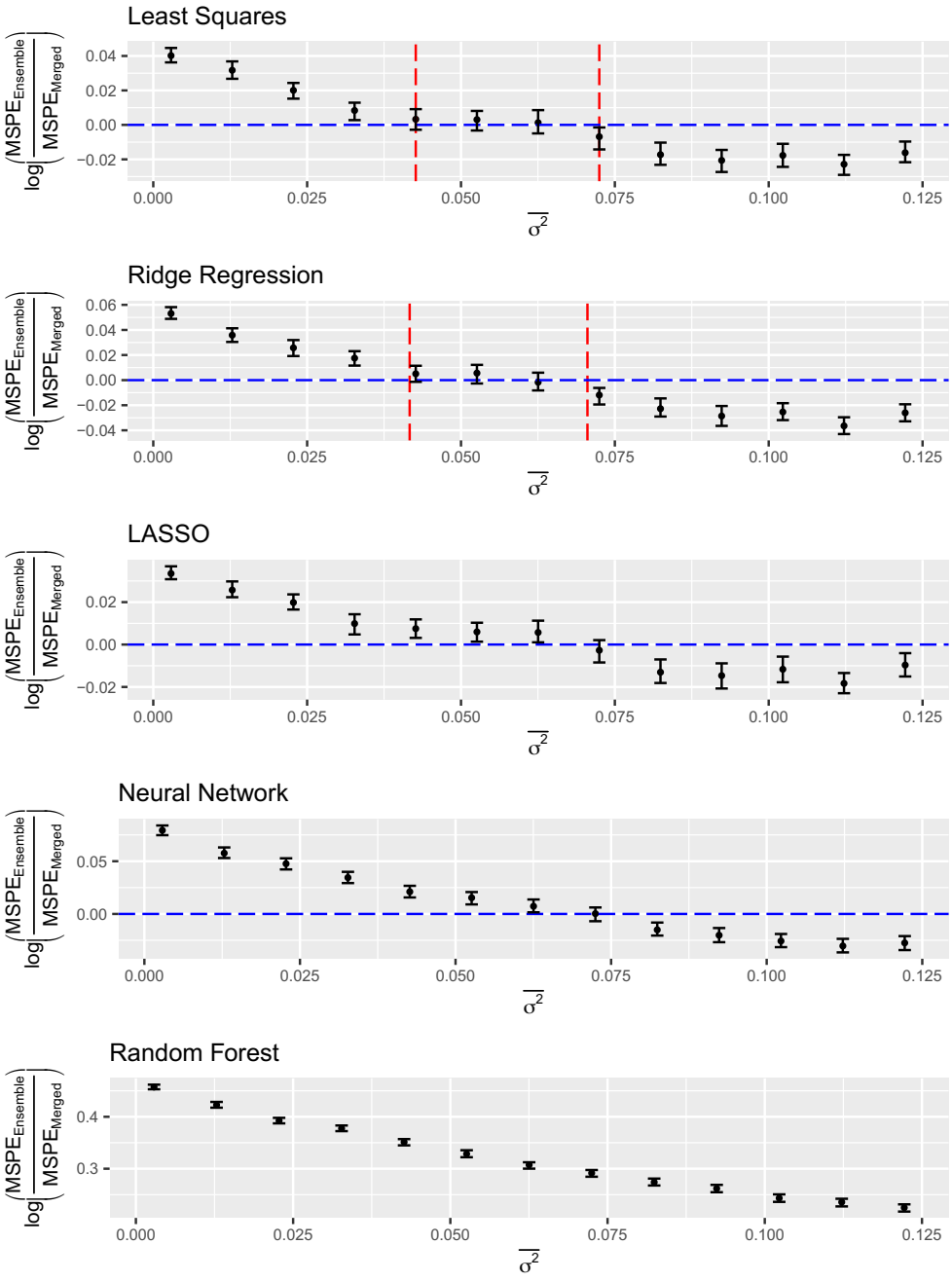

Figure S.3.1: Relative performance of merging and ensembling as a function of heterogeneity when $p = 10$, $q = 5$, and the random effects have unequal variances. MSPE: mean squared prediction error. The vertical dashed lines correspond to the bounds from Theorems 3.2 and 3.4.

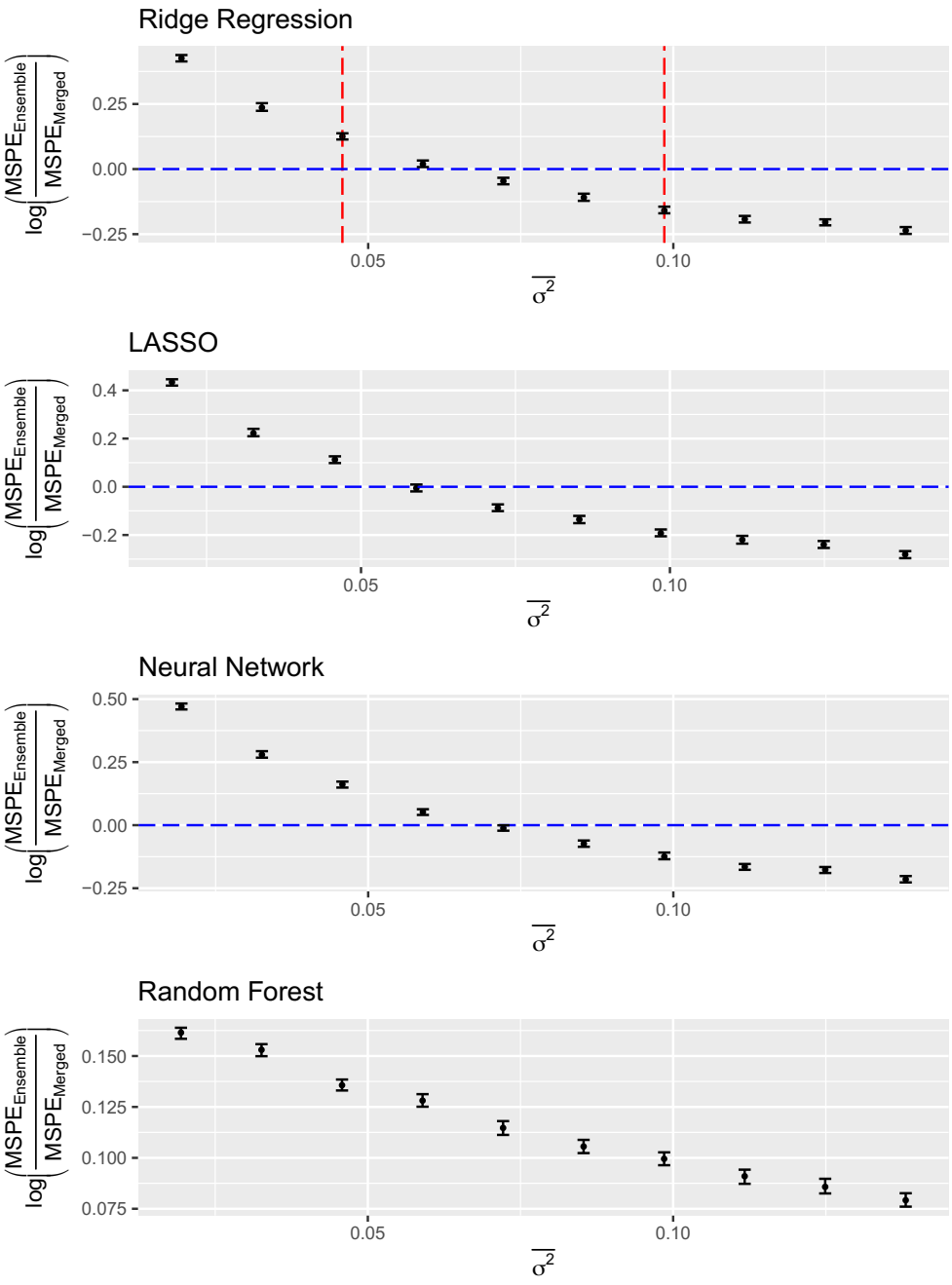

Figure S.3.2: Relative performance of merging and ensembling as a function of heterogeneity when $p = 100$, $q = 10$, and the random effects have unequal variances. MSPE: mean squared prediction error. The vertical dashed lines correspond to the bounds from Theorem 3.4.

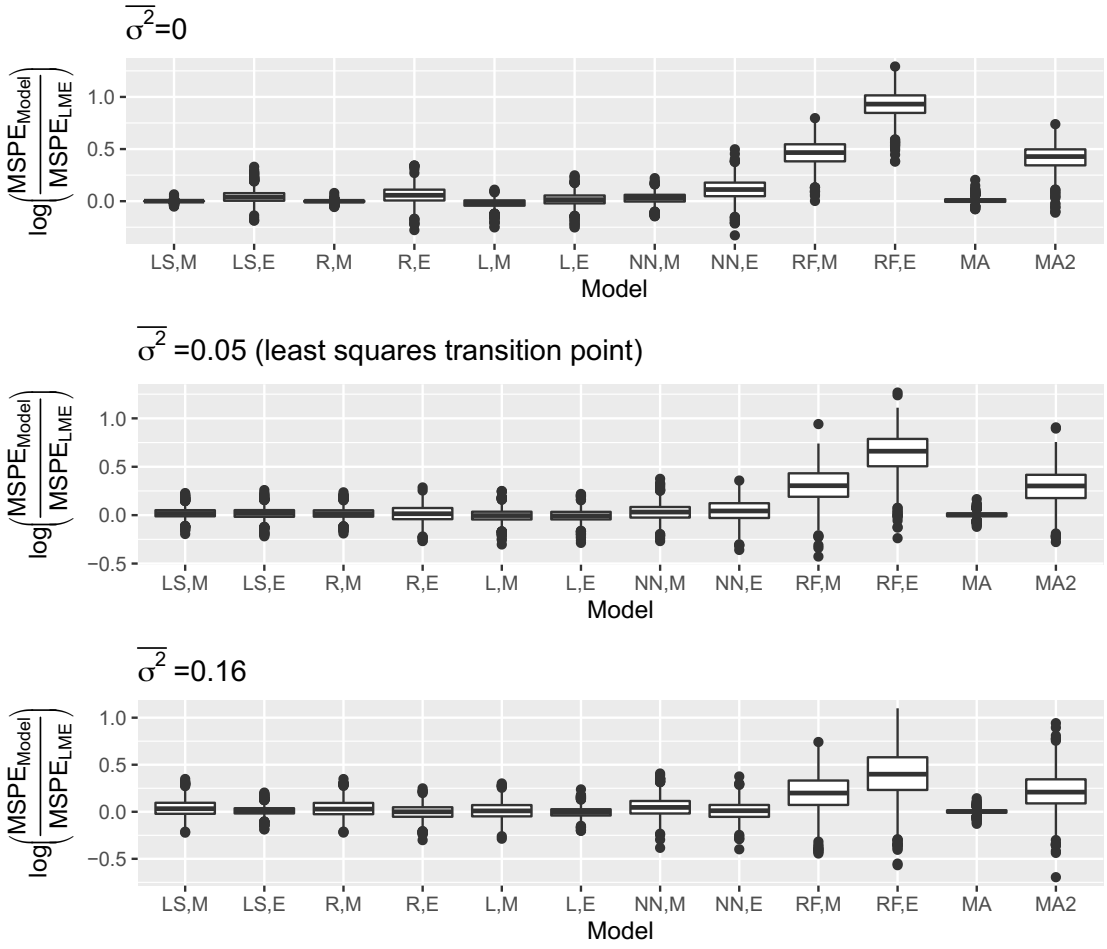

Figure S.3.3: Figure 3.3 with random forest and univariate meta-analysis results included. MSPE: mean squared prediction error; LME: linear mixed effects model; LS,M: merged least squares learner; LS,E: ensemble learner based on least squares; R,M: merged ridge regression learner; R,E: ensemble learner based on ridge regression; L,M: merged lasso learner; L,E: ensemble learner based on lasso; NN,M: merged neural network; NN,E: ensemble learner based on neural networks; RF,M: merged random forest; RF,E: ensemble learner based on random forests; MA: multivariate meta-analysis; MA2: multiple univariate meta-analyses.

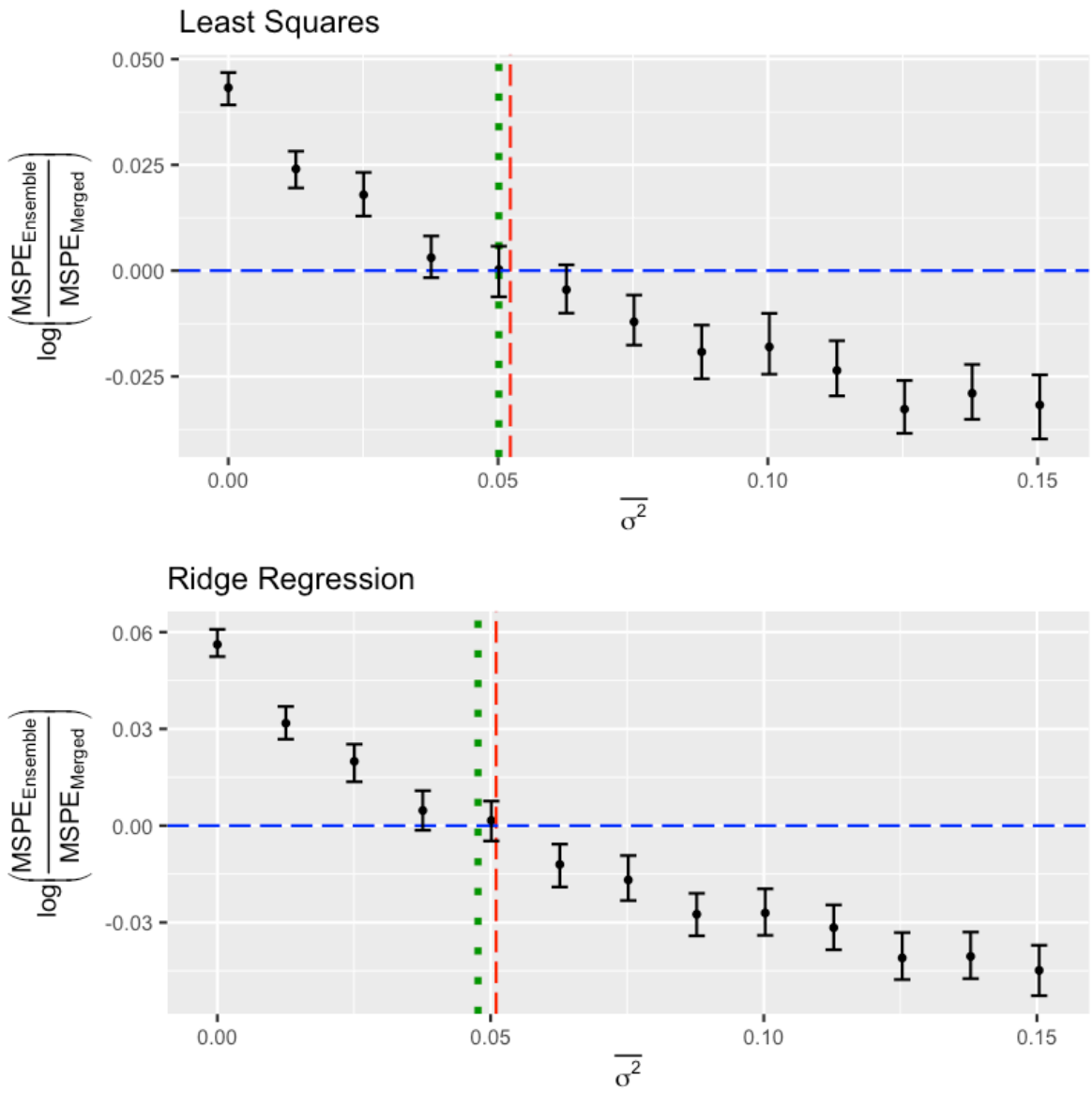

Figure S.3.4: Relative performance of merging and ensembling as a function of heterogeneity when $p = 10$, $q = 5$, and the random effects have equal variances. MSPE: mean squared prediction error. We used the same simulated data as described in the first simulation scenario in Section 4. However, instead of using equal weights in the ensembles, we estimated the optimal weights by plugging in estimates of $\boldsymbol{G}$, σ^2_ϵ, and (for ridge regression) $\boldsymbol{\beta}$ obtained from fitting the correctly specified linear mixed effects model. The vertical dotted lines correspond to the transition points based on the true optimal weights. The vertical dashed lines correspond to the transition points based on equal weights. The empirical transition point occurs at the value of $\overline{\sigma^2}$ where the log ratio of the prediction errors for ensembling and merging is equal to 0.