The Annals of Applied Statistics
2022, Vol. 16, No. 1, 495–520
https://doi.org/10.1214/21-AOAS1510

PREDICTION OF HEREDITARY CANCERS USING NEURAL NETWORKS

BY ZOE GUAN[1, a], GIOVANNI PARMIGIANI[2, b], DANIELLE BRAUN[3, d] AND LORENZO TRIPPA[2, c]

[1]*Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center,* [a]*guanz@mskcc.org*
[2]*Department of Data Sciences, Dana-Farber Cancer Institute,* [b]*gp@jimmy.harvard.edu,* [c]*ltrippa@ds.dfci.harvard.edu*
[3]*Department of Biostatistics, Harvard T.H. Chan School of Public Health,* [d]*dbraun@mail.harvard.edu*

Family history is a major risk factor for many types of cancer. Mendelian risk prediction models translate family histories into cancer risk predictions, based on knowledge of cancer susceptibility genes. These models are widely used in clinical practice to help identify high-risk individuals. Mendelian models leverage the entire family history, but they rely on many assumptions about cancer susceptibility genes that are either unrealistic or challenging to validate, due to low mutation prevalence. Training more flexible models, such as neural networks, on large databases of pedigrees can potentially lead to accuracy gains. In this paper we develop a framework to apply neural networks to family history data and investigate their ability to learn inherited susceptibility to cancer. 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. We propose adaptations of fully-connected neural networks and convolutional neural networks to pedigrees. In data simulated under Mendelian inheritance, we demonstrate that our proposed neural network models are able to achieve nearly optimal prediction performance. Moreover, when the observed family history includes misreported cancer diagnoses, neural networks are able to outperform the Mendelian BRCAPRO model embedding the correct inheritance laws. Using a large dataset of over 200,000 family histories, the Risk Service cohort, we train prediction models for future risk of breast cancer. We validate the models using data from the Cancer Genetics Network.

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. (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 (Antoniou et al. (2004), Berry et al. (1997), Tyrer, Duffy and Cuzick (2004), Wang et al. (2007, 2010)) and (2) using summaries of family history (e.g., the number of relatives with a previous cancer diagnosis) as covariates in regression models (Balmaña et al. (2006), Banegas et al. (2017), Choudhury et al. (2020a, 2020b), Gail et al. (1989, 2007), Matsuno et al. (2011), Tice et al. (2008)). Recently, deep learning models based on mammographic images have also been proposed Portnoi et al. (2019), Yala et al. (2019).

Mendelian models take as input a pedigree (Figure 1) that reflects family history of cancer (including relatives' cancer diagnoses, ages at cancer onset, and current ages). 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 (e.g., see Chen and Parmigiani (2007)). The individual risk of cancer is then calculated as a weighted

Received November 2020; revised June 2021.
Key words and phrases. Machine learning, Mendelian risk prediction, family history.

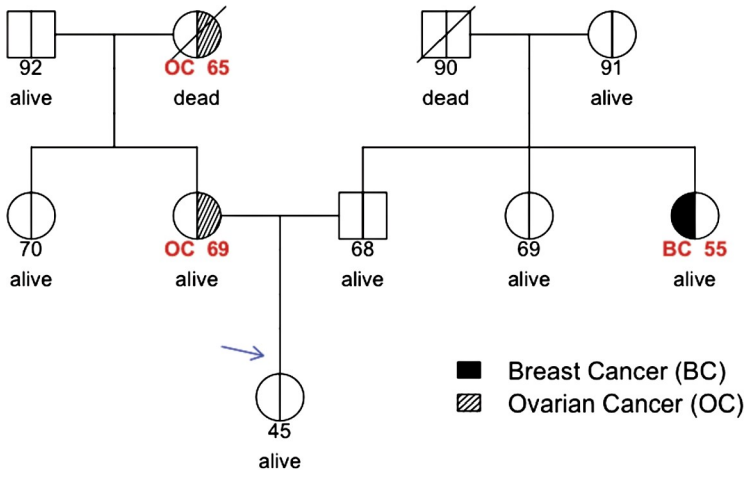

FIG. 1. *Example of a pedigree with family history of breast and ovarian cancers. Circles represent females and squares represent males. The arrow indicates the counselee, the individual undergoing risk assessment. Numbers below each family member represent the individual's current age* (*if alive and unaffected*), *age at death* (*if dead*), *and age of diagnosis* (*if affected by breast or ovarian cancer*).

average of mutation carrier and noncarrier 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)), and existing Mendelian models consider only a small subset of these genes. 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 and heterogeneity of prevalence across populations.

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 (Bishop (1995), Nielsen (2015)) is a model based on a directed graph that represents the relationship between a set of input features, typically provided in the form of a vector or matrix, and an outcome of interest. The graph consists of layers of nodes that apply a series of potentially nonlinear transformations to the input to produce a prediction or classification. In our setting the input to the NN will be a set of variables that describes the family history of an individual who presents for risk assessment. 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, Sutskever and Hinton (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, and investigate their ability to learn patterns of inherited susceptibility. 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. Our methodological contribution is adapting CNNs for pedigree data by defining local functions, similar to convolutional filters for image classification, that are applied repeatedly to sets of first-degree relatives within the pedigree (Section 2.4). We compare the performance of NNs to BRCAPRO (Parmigiani, Berry and Aguilar (1998)), a widely used Mendelian model, and logistic regression (LR). While there are many established risk factors for breast cancer (Gail et al. (1989), Brentnall et al. (2019)), in this paper we focus on prediction models based on family history. To allow for an interpretable comparison with BRCAPRO, which uses only family history information (along with race and ethnicity), the NN and LR models trained here do not include risk factors beyond family history (the counselee's age and personal history of cancer are considered to be part of the family history information). The inputs to the NN models are specified in Section 2.1. We note that it is straightforward to add new risk factors (e.g., breast density) to the NN models, and we discuss how this can be done in the Sections 2.3 and 2.4 (the methodology for FCNNs remains identical, while adding new features to CNNs potentially requires modifying the way in which nodes are connected). In our simulations we generate data, based on the Mendelian assumptions of BRCAPRO, and determine how large a sample size is needed for NNs to achieve competitive performance compared to the generating model. Moreover, we show that, when the observed family history includes misreported cancer diagnoses, NNs are able to outperform the Mendelian BRCAPRO model embedding the correct inheritance laws.

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. Methods.

2.1. *Notation*. Our notation is summarized in Table S1 of the Supplementary Material (Guan et al. (2022a)). Consider a counselee (someone who presents for risk assessment) who has not previously been diagnosed with a given type of cancer. Let t be a prespecified number of years. Let $Y_0 = 1$ if the counselee 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 (described below).

Family history can be visualized using a pedigree (Figure 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 H where each row corresponds to a family member, containing their features and the indices of their parents. Let R be the number of relatives in the pedigree besides the counselee. The family members are indexed by $r = 0, 1, \ldots, R$, where $r = 0$ corresponds to the counselee. 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 mother and A_{r2} the index of r's father (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. *Fully-connected neural networks.* A NN is a directed graph consisting of a sequence of layers (see Bishop (1995) or Nielsen (2015)) for examples and graphical representations of NNs). Each layer is a set of nodes that are linked to nodes in the previous layer through incoming edges and to nodes in the next layer through outgoing edges. A node receives a set of inputs via incoming edges, computes a function of its inputs, and propagates the result via outgoing edges. The first layer, which receives the input features (typically in the form of a vector), is called the input layer (in our setting the input features will correspond to the family history of the counselee). The final layer, which provides the output in the form of prediction or classification, is called the output layer. The layers in between, which are optional layers that apply transformations to the input data, are called hidden layers. A FCNN is a NN where every node in a given layer 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, 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 2.3.

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 $N_{L+1} = 1$. The outputs of the layers are

$$a^0 = X \in \mathbb{R}^{N_0},$$

$$a^l = \varphi^l\left(W^l a^{l-1} + b^l\right) \in \mathbb{R}^{N_l}, \quad (l = 1, \ldots, L),$$

where $W^l \in \mathbb{R}^{N_l * N_{l-1}}$ is the matrix of weights for layer l with row i containing the weights of node i, $b^l \in \mathbb{R}^{N_l}$ is the bias vector for layer l, and $\varphi^l : \mathbb{R}^{N_l} \to \mathbb{R}^{N_l}$ represents the componentwise application of an activation function $\varphi : \mathbb{R} \to \mathbb{R}$. Commonly used activation function include the logistic function $\sigma(z) = \frac{\exp(z)}{1+\exp(z)}$ and the rectifier function $\mathrm{ReLU}(z) = \max(0, z)$. The output layer ($l = L + 1$) consists of a single node that uses the logistic activation function, outputting the predicted probability

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

Given a cost function C and M training observations (X_m, Y_{0m}), $m = 1, \ldots, M$, the weight and bias parameters are randomly initialized and iteratively updated to minimize $\sum_{m=1}^{M} C(Y_{0m}, \hat{Y}_{0m})$, using methods such as stochastic gradient descent (Kiefer and Wolfowitz (1952)) and the Adam optimizer (Kingma and Ba (2014)). Examples of cost functions (Janocha and Czarnecki (2017)) include mean squared error, $C(y, z) = (y - z)^2$, and cross-entropy loss, $C(y, z) = -y \log(z) - (1 - y) \log(1 - z)$. When squared error loss or cross-entropy loss is used, it is appropriate to interpret the NN output as a probability (Hampshire II and Pearlmutter (1991)).

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, such as dropout (Srivastava et al. (2014)).

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 prespecified relatives (e.g., counselee, 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.

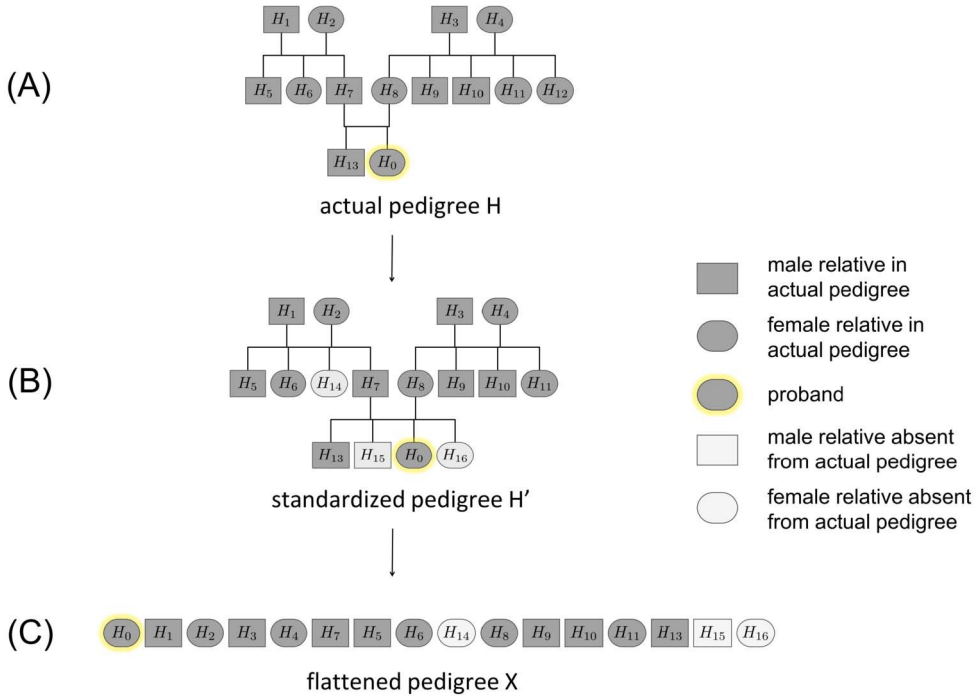

FIG. 2. *Consider a reference pedigree that includes the counselee's grandparents, parents, uncles, aunts, and siblings, with each couple having two children of each sex.* (*A*) *Actual pedigree* H. (*B*) *Standardized pedigree* H', *obtained by mapping* H *to the reference structure. The actual pedigree 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, sisters, and brothers than the reference pedigree, so in* H' *we use prespecified noninformative values for the paternal aunts, sisters, and brothers absent from* H. (*C*) *Flattened pedigree* X *which is used as input for a FCNN.*

We first describe the reference pedigree (see Figure 2(B) for an example of a reference structure). Let the reference pedigree contain the counselee and Q other types of relatives (mother, father, sister, brother, etc). Let $q = 0, 1, \ldots, Q$ index the relative types, with $q = 0$ corresponding to the counselee. 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 counselee, $r = 1, \ldots, R_1$ correspond to relatives of type $q = 1$, $r = R_1 + 1, \ldots, R_1 + R_2$ correspond to relatives of type $q = 2$, and so on.

The choice of the reference structure should depend on the family structures observed in the training data, and it is a compromise between model complexity/computational costs and potential loss of information. Since every counselee has two parents and four grandparents, the reference structure should, at least, include these relatives (assuming that most counselees provide information on these relatives). For other relatives, one approach is to calculate a summary measure, such as the median, for the number of relatives of each type (example: sister, brother, etc.) in the training data and define a reference structure where the number of relatives of a given type is equal to the value of the summary measure for the number of relatives of that type (example: if the median number of sisters is one in the training data, then include one sister in the reference structure). In order to reduce potential loss of information, the median can be replaced with a higher threshold, such as the third quartile. The amount of information lost can be quantified for each reference structure, using the mean proportion of family members dropped from the original pedigree. The reference structure can then be chosen based on the investigator's judgment of how much information loss is acceptable (this can be informed by prior knowledge or a sensitivity analysis looking at performance metrics

for models trained using different reference structures). Implementation details are provided in Section 2.8.

Now, we consider an actual pedigree matrix H and describe how to standardize and flatten it (Figure 2). For $q = 0, 1, \ldots, Q$, let R_q be the number of relatives of type q in H ($R_0 = 1$). To construct a standardized pedigree matrix, $\tilde{H}$, 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 $\tilde{H}$. If the actual number is smaller than the reference number ($R_q < R_q$), then we include the R_q actual relatives in $\tilde{H}$ and represent each of the $R_q - R_q$ absent relatives using a vector of prespecified 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 $\tilde{H}$. We also include a column in $\tilde{H}$ to indicate whether each row corresponds to a relative who is absent from the actual pedigree (0 if present, 1 if absent). Therefore, $\tilde{H}$ 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 $\tilde{H}_r$ be the vector for relative r in $\tilde{H}$. We flatten $\tilde{H}$ by concatenating its rows to get a vector, $X = (\tilde{H}_0, \tilde{H}_1, \ldots, \tilde{H}_R) \in \mathbb{R}^{(R+1)*(K+1)}$, which can be used as input to a FCNN. If there are additional features of interest beyond the family history features specified above (e.g., breast density), then they can simply be appended to the input vector X.

2.4. *Convolutional neural networks.* FCNNs are prone to overfitting since the number of parameters grows quickly with network size (Geman, Bienenstock and Doursat (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 (e.g., pixels that are spatially close often have highly correlated values), a convolutional layer applies the same functions (e.g., $x \to \max(0, wx)$) repeatedly to different fixed-size neighborhoods of the input (e.g., sets of neighboring pixels). These functions are called convolutional filters. The number of parameters in these local functions depends on the choice of reference pedigree and on K, the number of features considered for each family member. The reference pedigree and K can vary across different applications (e.g., the available family history information might be more detailed in some datasets than others), and, therefore, the corresponding local functions are tailored and applied to domains with distinct dimensionalities.

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 Material A.2 (Guan et al. (2022a)). 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 (Niepert, Ahmed and Kutzkov (2016), Wu et al. (2021)), such as molecular compounds. We follow two main steps: (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.

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). Prior to running the CNN, for each family member r in $\tilde{H}$, we define a fixed-size neighborhood, centered at r, consisting of r and r's first-degree relatives: self, mother, father, m_1 sisters, m_2 brothers, m_3 daughters, and m_4 sons. Similar to Figure 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 prespecified index representing an

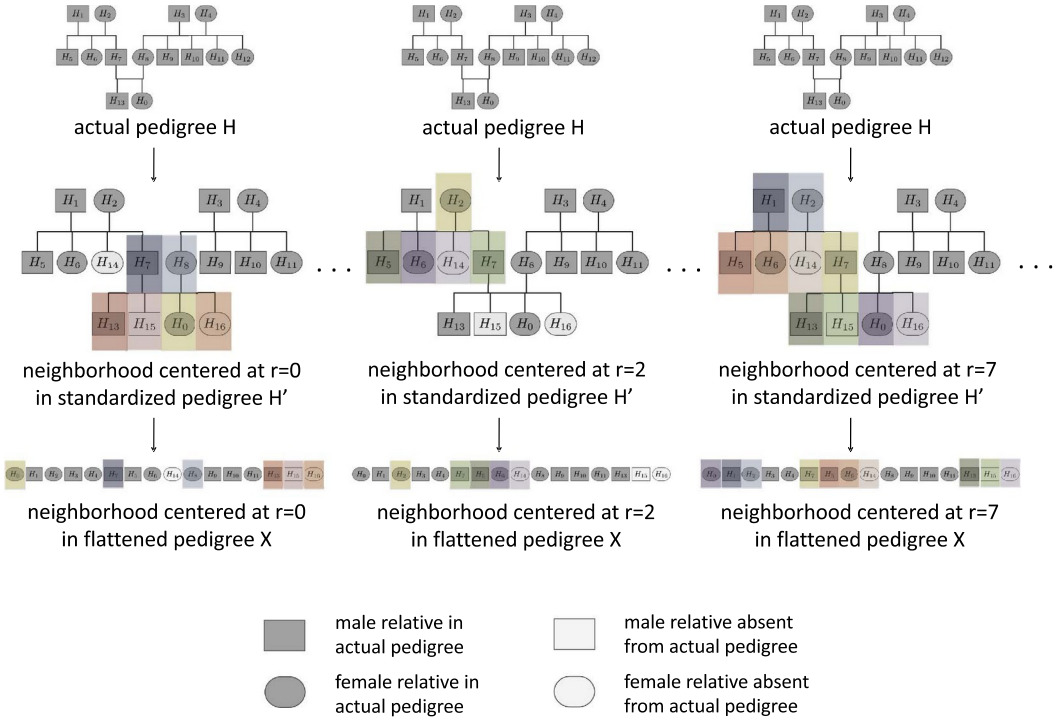

FIG. 3. *The neighborhoods centered at relatives* 0, 2, *and* 7 *are shown above using shaded boxes. The same convolutional filters are applied to all neighborhoods of the pedigree.*

absent relative whose features are set to zero (analogous to zero padding in standard CNNs, as described in Supplementary Material A.2, Guan et al. (2022a)). The same approach is used for brothers, daughters, and sons. The neighborhood is represented by a vector $N(r)$ of length $U = 3 + \sum_{i=1}^{4} m_i$. Within $N(r)$, 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 3). For $i = 1, \ldots, M_l$, let $f_i^l : \mathrm{R}^{U*M_{l-1}} \to \mathrm{R}$ (let $M_0 = K + 1$ since each relative has $K + 1$ features in H; see Section 2.3). Let $a_r^l \in \mathrm{R}^{M_l}$ be the output of layer l for neighborhood/family member r. Let $a_{N(r)}^{l-1} \in \mathrm{R}^{U*M_{l-1}}$ be the vector obtained by concatenating the layer inputs of the relatives in $N(r)$. The output from applying filter i to r's neighborhood is

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

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

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

$$a_r^0 = H_r \in \mathrm{R}^{K+1} \quad \text{for } l = 0, \quad \text{and}$$
$$a_r^l = f^l\left(a_{N(r)}^{l-1}\right) \in \mathrm{R}^{M_l}, \quad (l = 1, \ldots, L),$$

and the overall layer outputs are

$$a^l = \left(a_0^l, \ldots, a_R^l\right) \in \mathrm{R}^{M_l*(R+1)}, \quad (l = 0, 1, \ldots, L).$$

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

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

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 $\sum_{m=1}^{M} C(Y_{0m}, \hat{Y}_{0m})$, and the optimization can be carried out using stochastic gradient descent.

There are various ways to incorporate additional features beyond family history. Additional features that are applicable to all relatives (e.g., body mass index) can be appended to the input vector a_r^0 for each relative r. For additional features that are only applicable to the counselee, a modification to the appendment approach is necessary since the use of convolutional filters requires the input vector for each relative to have the same size. Two possible approaches are: (1) append the features to the input vector of each relative, but set their values to 0 for noncounselees, or (2) append the features for the counselee to the output vector of the Lth convolutional layer (i.e., the layer before the final output layer), thus expanding the input vector for the final output layer (a related approach is used in Li et al. (2017)).

2.4.1. *Model space*. Universal approximation theorems characterize the approximation capabilities of models and algorithms. The universal approximation theorem for FCNNs indicates that any continuous function over a given domain (e.g., the real line) can be approximated with arbitrary precision by a FCNN with a single hidden layer (Cybenko (1989), Hornik (1991), Leshno et al. (1993)). The theorem establishes the existence of a FCNN that satisfies the desired level of precision but does not provide a practical way to construct it. In our setting this is an attractive property because it means that any continuous relation between the family history (in the form of a fixed-length vector) and cancer risk can be approximated arbitrarily well by a FCNN. We show in this section that the CNNs we propose are just as powerful: they satisfy a universal approximation property similar to that of FCNNs.

Fix a reference pedigree H^* of size $R + 1$ containing relatives of up to degree d of the counselee. Let Q be the number of relative types in H^*, besides the counselee, and let $m = \max_{q=0,1,\ldots,Q} R_q$. Let $X^* \subset \mathbb{R}^{(R+1)*(K+1)}$ be the space of pedigrees with the same structure as H^*. We consider the CNN's ability to approximate functions from X^* to $[0, 1]$. We first state the universal approximation theorem for standard FCNNs (Leshno et al. (1993)) and then verify that the same property extends to CNNs (proof provided in Supplementary Material A.3, Guan et al. (2022a)).

2.4.1.1. *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 $\varphi : \mathbb{R} \to \mathbb{R}$ be a piecewise continuous, locally bounded, and nonpolynomial activation function. Then, given $\varepsilon > 0$, 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

$$F(X) = \sum_{i=1}^{N} \alpha_i \varphi(w_i \cdot X + b_i)$$

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

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* $g : X^* \to [0, 1]$ *be continuous. Let* $\varphi : \mathbb{R} \to \mathbb{R}$ *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* $\varepsilon > 0$, *there exists a pedigree CNN of the form described in Section* 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*

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

satisfies $|F(X) - g(X)| < \quad \forall X \in \quad X^*$.

2.5. *Missing data*. In practice, there is often missing information in family history data (e.g., an unreported relative or an unknown diagnosis age). Missing values in the training and/or test set can be handled using standard imputation methods or complete case analysis (Little and Rubin (2019)), though the latter may result in a substantial decrease in sample size. Missing value imputation can be implemented as a preprocessing step separate from training or prediction (García-Laencina, Sancho-Gómez and Figueiras-Vidal (2010)). In clinical practice some models do not allow missing values (e.g., the Claus model (Claus, Risch and Thompson (1994))), and clinicians impute missing information (e.g., ages of diagnosis for relatives) to compute predictions. Some popular clinical tools automatically impute missing information. For example, in the Risk Service tool, a missing diagnosis age for a relative is imputed, based on the relative's current age Chipman et al. (2013).

Another approach that can be implemented for NNs and prediction models in general is to include as predictors indicator functions denoting whether certain features are missing (Choi, Dekkers and le Cessie (2019)). In our analyses we used this approach to represent absent family members when mapping families to a reference pedigree (Figure 2) that potentially contains relative types absent from the actual family. Since missing values are distinct from nonexistent data, separate indicators could be used for missingness versus absence.

As described in Section 3.2.2, we performed a sensitivity analysis using simulated data to evaluate the impact of missing relatives and missing ages of diagnosis.

2.6. *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, Berry and Aguilar (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, using Bayes' rule, laws of Mendelian inheritance, mutation prevalence and penetrance, and family history of breast and ovarian cancer. The family history information includes the $K = 6$ features described in Section 2.1: breast/ovarian cancer status, age at onset of breast/ovarian cancer if applicable, and current age or age at death. In addition, BRCAPRO provides the option of modifying the default prevalences and penetrances, using the following covariates if they are available: race, ethnicity, genetic testing results for BRCA1/BRCA2, marker testing results (ER/CK14/CK5/CK6/PR/HER2), and prophylactic mastectomy/oophorectomy (these additional covariates were not included in our simulations).

Let γ_r be the genotype of relative r (noncarrier, carrier of a pathogenic BRCA1 mutation, carrier of a pathogenic BRCA2 mutation, or carrier of pathogenic mutations in both BRCA1 and BRCA2). Using Bayes' rule and the assumption of conditional independence of phenotypes given genotypes, the counselee's probability of having genotype γ_0 is

$$P(\gamma_0|H) = \frac{P(\gamma_0) \sum_{\gamma_1,\ldots,\gamma_R} \prod_{r=0}^{R} P(H_r|\gamma_r) P(\gamma_1, \ldots, \gamma_R|\gamma_0)}{\sum_{\gamma_0} P(\gamma_0) \sum_{\gamma_1,\ldots,\gamma_R} \prod_{r=0}^{R} P(H_r|\gamma_r) P(\gamma_1, \ldots, \gamma_R|\gamma_0)}. \tag{1}$$

The summation over genotypes is calculated using the Elston–Stewart peeling algorithm; Elston and Stewart (1971) and $P(\gamma_1, \ldots, \gamma_K|\gamma_0)$ is calculated based on Mendelian laws of inheritance. The prevalences $P(\gamma_r)$ are obtained from the literature and are ethnicity-specific (in particular, different prevalences are used for Ashkenazi Jewish and non-Ashkenazi Jewish families). $P(H_r|\gamma_r)$ is calculated using literature-based penetrances for breast and ovarian cancer. The penetrances are functions that represent the risk of cancer at different ages, and they are genotype-, cancer-, and sex-specific. The penetrance functions for noncarriers 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 cancer through a weighted average of the genotype-specific penetrance functions $P(Y_0 = 1|\gamma_0)$,

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

2.7. *Model evaluation*. We evaluated model performance using four metrics (Steyerberg et al. (2010)): (1) the ratio of observed (O) to expected (E) events (where E is the sum of the predictions in the test set), a measure of calibration, (2) the area under the receiver operating characteristic curve (AUC), a measure of discrimination, (3) the area under the precision recall curve (PR-AUC), another measure of discrimination that is more sensitive to class imbalance than the AUC, and (4) the Brier score which is the mean squared difference between the predicted probabilities and actual outcomes. We obtained 95% confidence intervals (CIs) for the metrics by bootstrapping the test set 1000 times.

2.8. *Implementation*. We ran BRCAPRO using the BayesMendel R package (version 2.1-6) (Chen et al. (2004)). 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 adapted code from Hechtlinger, Chakravarti and Qin (2017).

In the simulations, 887,353 randomly generated families were split into a training set of 800,000 and a test set of 87,353. In the data application the Risk Service dataset (279,460 families) was used for training, and the CGN dataset (7489 families) was used for testing. In both the simulations and data application, we used the Adam optimizer Kingma and Ba (2014) and the mean squared error loss function (while cross-entropy loss is more commonly used for binary outcomes, we chose to use mean squared error, because it corresponds to the minimization of the Brier score, which is a standard performance metric in risk prediction Steyerberg et al. (2010), and one of the metrics we used to compare models; more discussion on this choice and a sensitivity analysis are provided in Supplementary Material B.4, Guan et al. (2022a)). We used a typical 90/10 split of the training set to tune NN hyperparameters via a random search Bergstra and Bengio (2012): 10% of the training set was held out for evaluating the performance of different choices for the number of hidden layers (one to three), sizes of hidden layers (10 to 100), number of filters for the CNN (three to 10), learning rate (0.0001 to 0.01), weight decay parameter (0 to 0.01), activation function (ReLU, or elu), and dropout rate (0 to 0.5). The performance in the tuning set was highly sensitive to the hyperparameter values (in the simulations, AUCs in the held-out subset ranged from 0.38–0.65 for the FCNN and 0.56–0.65 for the CNN: https://github.com/zoeguan/nn_cancer_risk/tree/master/tuning_results), so it is important to explore different sets of hyperparameters.

In the simulations, the FCNNs had two hidden layers of sizes 30 and 10, and the CNNs had two convolutional layers with 10 and five filters. In the data application the FCNN had two hidden layers of size 30, and the CNN had two convolutional layers with five 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, Unterthiner and Hochreiter (2015)). For the NNs and LR, features were normalized to be between 0 and 1 using min–max normalization (Patro and Sahu (2015)). The code for the analyses is provided as a supplement (Guan et al. (2022b)) (it is also available at github.com/zoeguan/nn_cancer_risk) and contains additional details on hyperparameter values.

In the simulations we used a reference pedigree of size 26, containing the counselee's grandparents, parents, aunts (two maternal, three paternal), uncles (three maternal, two paternal), siblings (two sisters, three brothers), and children (two daughters, two sons). This was chosen based on the distribution of family structures in the CGN (see Supplementary Material B.2, Guan et al. (2022a)). 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 two relatives of each type and omitted sons and daughters, due to the smaller family sizes in the training dataset (see Supplementary Material B.2, Guan et al. (2022a)). We used $m_1 = m_2 = 2$ and $m_3 = m_4 = 1$ for the CNN neighborhoods.

Studies have shown that restricting family history to first- and second-degree relatives (Biswas et al. (2013), Terry et al. (2019)) has little impact on discriminative accuracy. Therefore, we considered only first- and second-degree relatives in the reference pedigree. As described in Section 3.2.2, we conducted a sensitivity analysis for various choices of reference pedigree structures and found little variation in performance.

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 the data are consistent with BRCAPRO and one where they are not.

3.1. *Simulation approach.* 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 4.1). For counselees we also sampled dates of birth and baseline dates for risk assessment from the CGN. For noncounselees, dates of birth were generated relative to the counselee'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 counselee'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 noncarriers 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.

3.2. *Results.* We excluded counselees who died or were diagnosed with breast cancer prior to baseline. For the remaining counselees ($n = 887{,}353$), we predicted 10-year risk of breast cancer, using the baseline family history. We used 800,000 families for training and the other 87,353 for testing. In the training set there were 23,606 cases (counselees who developed breast cancer within 10 years). In the test set there were 2570 cases.

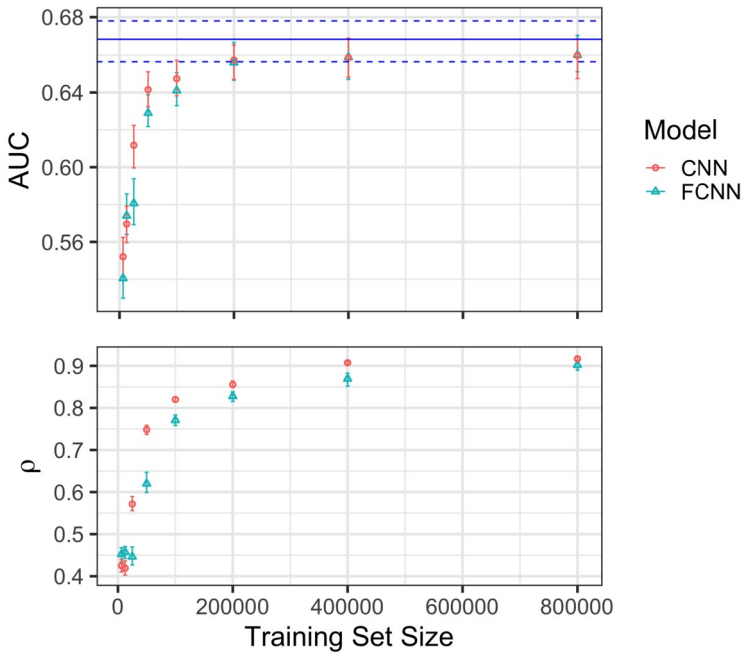

FIG. 4. *AUC and correlation (ρ) of NN predictions with BRCAPRO predictions for* 10*-year risk of developing breast cancer as a function of training sample size* (*ranging from to* 6250 *to* 800,000) *in simulations*.

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 6250 to 800,000 (Figure 4). As the sample size increased, the AUCs of the NNs approached that of BRCAPRO, the true data generating model, and the predictions from the NNs became highly correlated with those from BRCAPRO. 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 the true model (both NNs had an AUC of 0.660 while the true model had an AUC of 0.668).

The NNs provided a better approximation of the true model than LR. The FCNN and CNN trained on the entire training set achieved correlations of 0.9 and 0.92 with the true model, while the LR model trained on the same data had a correlation of 0.82 ("True Family History" section in Table 1). The NNs also outperformed LR with respect to AUC, PR-AUC, and Brier score: across 1000 bootstrap replicates of the test set, the NNs had a better AUC and Brier score than LR more than 99% of the time. The proportion of cases in our dataset is very small; therefore, all of the models have low PR-AUCs (the baseline PR-AUC, or PR-AUC of a model that does no better than random guessing, is the proportion of cases, 0.029). The CNN was more highly correlated with BRCAPRO than the FCNN across all 1000 bootstrap replicates. Also, the CNN had a better Brier score than the FCNN in more than 95% of the bootstrap replicates and a higher AUC in 58% of the replicates. The CNN and LR both had good overall calibration, with O/E = 0.99 (95% CI 0.95–1.03) for the CNN and O/E = 1.00 (95% CI 0.96–1.04) for LR (Table 1), while the FCNN slightly overestimated risk, with O/E = 0.93 (95% CI 0.89–0.96). Across the bootstrap replicates, the CNN and LR performed similarly with respect to calibration, with the CNN showing better calibration in about half of the replicates. The CNN and LR had better calibration than the FCNN in more than 97% of the replicates. Calibration plots by decile of estimated risk (Figure 5) show that

TABLE 1

Model performance in simulated families (training set of 800,000), based on true and misreported family history. AUC: % rel
PR-AUC: % relative improvement in precision-recall AUC compared to BRCAPRO. sqrt(BS): % relative improvement
correlation with BRCAPRO. The "Comparisons Across Bootstrap Replicates" section shows pairwise comparisons betwee
bootstrap replicates of the test set; the row for A > B shows the proportion of bootstrap replicates where model A outp

	O/E	AUC	PR-AUC
True Family History			
Performance Metrics			
FCNN	0·93 (0·89, 0·96)	−1·21 (−1·73, −0·63)	−10·16 (−13·81, −7·08)
CNN	0·99 (0·96, 1·03)	−1·24 (−1·80, −0·69)	−7·93 (−11·52, −4·35)
LR	1·00 (0·97, 1·04)	−2·07 (−2·68, −1·47)	−14·59 (−19·04, −10·25)
BRCAPRO	1.02 (0.98, 1.06)	AUC = 0.668	PR-AUC = 0.065
Comparisons Across Bootstrap Replicates			
FCNN > CNN	0.021	0.582	0.020
FCNN > LR	0.025	1.000	0.990
FCNN > BRCAPRO	0.083	0.000	0.000
CNN > LR	0.464	1.000	0.999
CNN > BRCAPRO	0.691	0.000	0.000
Misreported Family History			
Performance Metrics			
FCNN	1.06 (1.02, 1.10)	2.82 (1.72, 3.99)	9.31 (2.66, 16.43)
CNN	1.01 (0.97, 1.05)	2.70 (1.63, 3.72)	11.15 (5.41, 17.47)
LR	1.00 (0.96, 1.04)	2.35 (1.23, 3.49)	6.12 (0.41, 12.47)
BRCAPRO	0.81 (0.78, 0.84)	AUC = 0.627	PR-AUC = 0.050
Comparisons Across Bootstrap Replicates			
FCNN > CNN	0.033	0.666	0.232
FCNN > LR	0.061	0.968	0.869
FCNN > BRCAPRO	1.000	1.000	0.998
CNN > LR	0.406	0.900	0.991
CNN > BRCAPRO	1.000	1.000	1.000

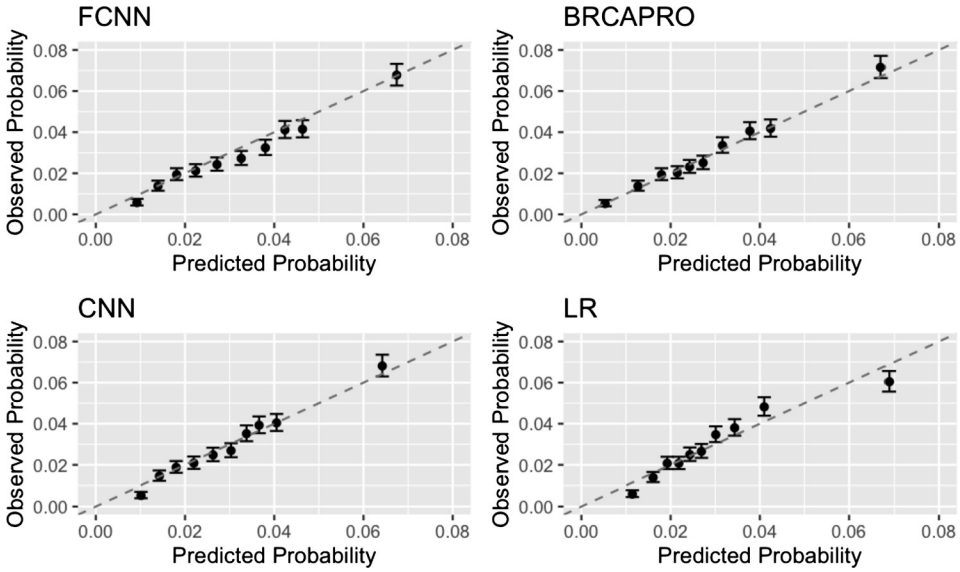

FIG. 5. *Calibration plots by decile of risk in simulated families* (*training set of* 800,000).

LR underestimated or overestimated risk in more deciles, compared to the other models. We also plotted the precision-recall curves for the models (Figure S2, Guan et al. (2022a)) which were not substantially different across models.

Differences between LR and CNN. Under the true model the counselee's risk of breast cancer increases with more affected relatives and earlier diagnosis ages. To assess whether NN and LR predictions captured these trends, we fixed a family structure and varied the phenotypes of the mother and maternal grandmother (Figure 6). We considered five scenarios ordered by increasing risk with respect to the true model: (A) no affected relatives,

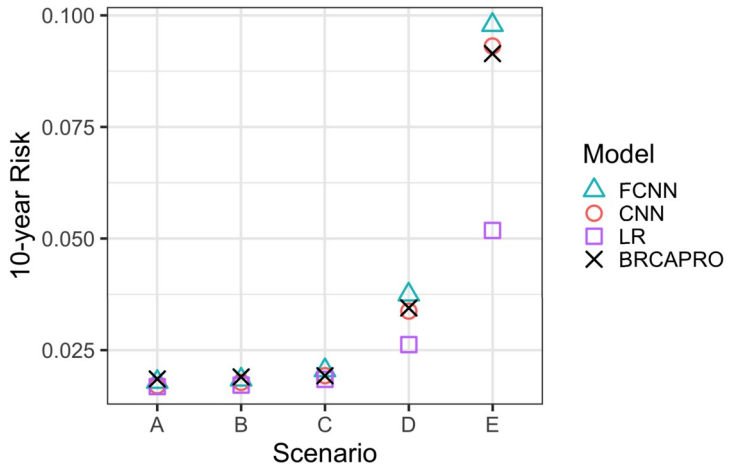

FIG. 6. *Under the first simulation setting, we fixed a simulated family structure* (*counselee, grandparents, parents, one paternal aunt, one maternal aunt, two maternal uncles*) *for a* 40*-year-old counselee and varied the level of family history across five scenarios* (*ordered by increasing risk 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.*

(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 6), LR slightly underestimated risk in Scenario (D) and severely underestimated 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, the data generating model, 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.

3.2.1. *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 noncounselees, using misreporting rates from Ziogas and Anton-Culver (2003) and Braun et al. (2018) (see Supplementary Material B.1 for details, Guan et al. (2022a)).

Under misreporting the NNs outperformed BRCAPRO with respect to calibration, AUC, and Brier score across almost all of the 1000 bootstrap replicates of the test set (Table 1), illustrating the advantage of NNs over BRCAPRO when the Mendelian assumptions are not fully satisfied. The NNs also outperformed LR with respect to AUC and PR-AUC in most of the bootstrap replicates. With respect to the Brier score, the CNN outperformed LR in more than 99% of the replicates, while the FCNN performed similarly to LR. The CNN had similar calibration to LR, while the FCNN had worse calibration.

3.2.2. *Sensitivity analyses.* Using simulated data, we performed sensitivity analyses to evaluate the impact of the choice of reference pedigree and missing data.

To evaluate the impact of the choice of reference pedigree, we quantified the amount of information lost (mean proportion of family members dropped from the original pedigree) for various reference structures based on different summary measures for the number of relatives of each type (Supplementary Material B.2, Guan et al. (2022a)). We considered "symmetric" reference structures, where the number of daughters is equal to the number of sons for each couple as well as reference structures without this constraint. We assessed the discriminatory accuracy of the models trained, using the various reference structures. The results show only small differences in performance for reference structures based on using the first, second, third, or fourth quartile as the summary measure, even though the mean proportion of family members dropped varies substantially across these choices (from 0 for the fourth quartile to approximately 0.4 for the first quartile). Therefore, in our application the performance of the NNs is not particularly sensitive to the choice of reference structure. In the main analyses we used a symmetric reference structure, based on the third quartile of relative counts. In other settings where performance may be more sensitive to the choice of reference structure, it can be chosen based on cross-validation AUCs or other performance metrics.

We also considered the impact of different proportions of missing data in the training and test sets. We evaluated the impact of: (1) missing relatives by removing relatives from the pedigree and (2) missing diagnosis ages for affected relatives. In the first scenario we considered removing relatives at random, which corresponds to noninformative missingness as well as removing only unaffected relatives, which corresponds to informative missingness. In prediction problems, missing data is likely to have a larger impact when the amount of missingness differs between the training and test datasets, so we considered scenarios where there were missing data in the training set but complete data in the test set as well as symmetric scenarios with complete data in the training set and missing data in the test set. We varied the

proportion of missing relatives and missing diagnosis ages from 0.05 to 0.3. We used single imputation to handle the missing ages, setting them to 50 for individuals over 50 and setting them to the individual's current age otherwise. The types of missingness considered did not have a substantial impact on any of the performance measures (Tables S5–S8, Guan et al. (2022a)).

3.2.3. *Computational costs.* Among the models trained, LR is the least computationally intensive and CNN the most computationally intensive. The training times, using a single CPU core for different training set sizes ranging from 6250 to 800,000, are provided in Figure S3 (Guan et al. (2022a)) (using the NN hyperparameters from the main simulation analysis). The relationship between sample size and training time is approximately linear for each model. With 800,000 training families, it took about one minute to train the LR model, five minutes to train the FCNN, and 20 minutes to train the CNN. The NNs also require additional computation to tune the hyperparameters prior to training the final model which can considerably increase the computational burden. However, hyperparameter tuning methods, such as grid search and random search, can be parallelized, and in some cases, using a GPU Oh and Jung (2004) can speed up the tuning/training process.

4. Data application. We trained NN and LR models to predict 5-year risk of breast cancer using data from the Risk Service and compared their performance to BRCAPRO using data from the CGN. We excluded male counselees, counselees who had breast cancer/bilateral mastectomy/bilateral oophorectomy before baseline, counselees under 18 years old, and counselees for whom we could not run BRCAPRO (counselees over 89 years old).

4.1. *Datasets.*

4.1.1. *Risk service.* The Risk Service (Chipman et al. (2013)) is a web service that provides risk predictions from various family history-based cancer risk models, including BRCAPRO. It has been used in primary care, breast imaging, and genetic counseling clinics. As of January 2018, the Risk Service database contained patient-reported family history inputs for over 450,000 counselees, with 285,161 counselees consenting to the use of their data for research.

Model training requires baseline and follow-up data, but the Risk Service does not follow counselees over time. We, therefore, defined each counselee's baseline date to be five 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 noncounselees (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 noncounselees for training, and we imputed baseline cancer status for noncounselees with missing diagnosis ages (see Supplementary Material B.7 for more details, Guan et al. (2022a)).

The training set consisted of 279,460 counselees (Table 2). The median age was 45, and the median family size was eight. Also, 36,783 counselees (13.2%) had at least one affected first-degree relative, and 13,307 (4.8%) developed breast cancer during the follow-up period.

4.1.2. *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 selfreferral.

TABLE 2
Characteristics of training (Risk Service) and test (CGN) datasets

Variable	Category	Risk Service	CGN
N (counselees)		279,460	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	242,677 (86.8)	4277 (57.1)
	1	35,241 (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 (%)		13,307 (4.8)	114 (1.5)

They provided information on personal and family history of cancer and sociodemographic factors through a baseline phone interview and annual follow-up updates.

The test cohort consisted of 7489 counselees. The median age was 47, and the median family size was 16. The majority (54.1%) of counselees were recruited from population-based cancer registries. Also, 42.9% of counselees had at least one female first-degree relative with breast cancer (a much higher proportion than in the Risk Service), 114 (1.5%) counselees developed breast cancer within five years of baseline, and 1017 counselees (13.6%) were lost to follow-up within five years without being diagnosed with breast cancer (Table 2). To adjust for censoring, we used inverse probability of censoring weights (Uno et al. (2007), Gerds and Schumacher (2006)) (see Supplementary Material B.8 for details, Guan et al. (2022a)).

4.2. *Training and test populations*. There are many differences between the Risk Service and CGN cohorts (Table 2). Since CGN participants were recruited based on family history of cancer, the CGN cohort represents a higher-risk population and has more counselees with a positive family history (Table 2). Due to different data collection and ascertainment procedures, the family history information available in the CGN 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 noncounselee relatives in the NN features and used only their breast and ovarian cancer affection statuses (we still used the counselee'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.

To account for the described differences between the CGN and Risk Service populations, we recalibrated the models trained on the Risk Service to general U.S. population incidence rates adjusted for family history. The approaches have been previously discussed for various regression calibration problems (Carroll et al. (2006)). We calculated age-specific five-year risks, based on 2012–2016 incidence rates 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 Collaborative Group on Hormonal Factors in Breast Cancer (2001) (the relative risks were 1.8 for one affected relative, 2.9 for two affected relatives, and 3.9 for two or more affected relatives). To recalibrate each model, we used the Risk Service data to fit a linear regression with the family history-adjusted five-year SEER risk as the outcome and the five-year risk from the model as the predictor. We also evaluated a recalibrated version of BRCAPRO, obtained via the SEER recalibration approach.

4.3. *Results*. Table 3 compares the performance of five models (FCNN, CNN, LR, BRCAPRO, and BRCAPRO^C, the SEER-recalibrated version of BRCAPRO) in the CGN dataset which were not used for training. All models underpredicted risk, with underprediction being most severe in the clinic-based subset of CGN. This may be because the CGN counselees were ascertained based on having a family history of cancer and, therefore, represented a higher-risk population than the sources of the data used for training and recalibration. Overall, the NNs and BRCAPRO had comparable PR-AUCs and Brier scores, performing better than LR with respect to these metrics. The CNN and BRCAPRO also performed better than LR with respect to the AUC. In the analyses stratified by ascertainment mode, the comparisons across 1000 bootstrap replicates show evidence of accuracy improvements achieved by the CNN over the other models. Both in the *population-based* test pedigrees (63 cases) and in the *clinic-based* test pedigrees (39 cases), the CNN achieved better PR-AUCs and Brier scores than LR and BRCAPRO in the majority of the bootstrap replicates. The CNN also achieved a higher AUC than LR in the majority of the bootstrap replicates in each stratum. In the *population-based* pedigrees, the CNN achieved a higher AUC than BRCAPRO in 94% of the bootstrap replicates, while, in the *clinic-based* pedigrees, BRCAPRO achieved a higher AUC than the CNN in 58% of the replicates.

We performed an additional analysis where we trained the NN and LR models, using only 40,000 Risk Service families, instead of all 279,460 families. The models trained, using the smaller sample size, all performed worse (Table S10, Guan et al. (2022a)) than the versions trained using all Risk Service families (Table 3). In particular, the models trained using 40,000 families had worse calibration. The FCNN had considerably lower discrimination in the overall cohort and population-based subset compared to before, indicating that large training sets are needed to develop accurate empirical models. However, the CNN trained using 40,000 families still performed reasonably well, compared to BRCAPRO.

5. Discussion. The main contributions of our paper are: (1) adapting FCNNs and CNNs to family history data and (2) investigating their potential for learning genetic susceptibility to cancer. To the best of our knowledge, we are the first to develop cancer risk prediction models using a dataset of more than 200,000 pedigrees. Our simulations and data application show that NNs are a promising approach for developing new 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. In the setting where family history was subject to misreporting, the NNs outperformed BRCAPRO. The simulations also showed that NNs can learn feature interactions that are not prespecified (such as rare but strongly predictive patterns involving multiple affected individuals on the same side of the family or multiple cancers in the same individual).

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 performance, compared to BRCAPRO in the overall cohort. They had slightly higher AUCs than BRCAPRO in population-based counselees but performed worse than BRCAPRO in clinic-based counselees with a stronger family history. These results are promising because BRCAPRO is based on domain knowledge accumulated over two decades of epidemiological studies (including Antoniou et al. (2002), Chen and Parmigiani (2007), Easton, Ford and Bishop (1995), Miki et al. (1994), Wooster et al. (1995)) while the NNs were trained on a single dataset. The poorer performance of the NNs in clinic-based counselees may partly

TABLE 3

Performance in CGN cohort, overall and stratified by ascertainment mode. The NN and LR models were trained using a randomly selected subset of 40,000 *Risk Service counselees. BRCAPRO^C: Recalibrated version of BRCAPRO. AUC: % relative improvement in AUC compared to BRCAPRO. AUC: % relative improvement in PR-AUC compared to BRCAPRO. sqrt(BS): % relative improvement in root Brier Score compared to BRCAPRO. ρ: correlation with BRCAPRO. In the table the "Comparisons Across Bootstrap Replicates" component 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	PR-AUC	sqrt(BS)
Overall (114 cases)				
Performance Metrics				
FCNN	1.16 (0.95, 1.37)	− 4.22 (− 12.37, 4.76)	− 6.70 (− 33.06, 27.75)	− 0.02 (− 0.27, 0.24)
CNN	1.10 (0.90, 1.30)	− 2.53 (− 10.69, 5.92)	− 4.79 (− 31.35, 33.38)	0.03 (− 0.22, 0.31)
LR	1.07 (0.89, 1.27)	− 4.56 (− 12.87, 4.25)	− 11.34 (− 34.51, 20.35)	− 0.09 (− 0.36, 0.19)
BRCAPRO	1.34 (1.11, 1.59)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	− 0.03 (− 0.05, − 0.00)
BRCAPRO^C	1.20 (0.99, 1.42)	AUC = 0.654	PR− AUC = 0.029	sqrt(BS) = 0.130
Comparisons Across Bootstrap Replicates				
FCNN > CNN	0.090	0.251	0.334	0.153
FCNN > LR	0.109	0.597	0.784	0.887
FCNN > BRCAPRO	0.956	0.195	0.314	0.396
CNN > LR	0.179	0.803	0.852	0.972
CNN > BRCAPRO	0.923	0.279	0.386	0.566
Population− Based (63 cases)				
Performance Metrics				
FCNN	1.12 (0.87, 1.39)	6.05 (− 5.26, 16.80)	25.25 (− 14.99, 63.46)	0.07 (− 0.18, 0.28)
CNN	1.05 (0.81, 1.31)	6.12 (− 1.55, 14.09)	22.35 (− 12.58, 55.70)	0.07 (− 0.14, 0.26)
LR	1.07 (0.83, 1.34)	4.23 (− 7.14, 15.67)	13.35 (− 21.87, 53.71)	− 0.04 (− 0.33, 0.20)
BRCAPRO	1.41 (1.09, 1.76)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	− 0.03 (− 0.05, 0.00)
BRCAPRO^C	1.25 (0.97, 1.56)	AUC = 0.648	PR− AUC = 0.024	sqrt(BS) = 0.128
Comparisons Across Bootstrap Replicates				
FCNN > CNN	0.269	0.499	0.618	0.450
FCNN > LR	0.244	0.839	0.923	0.955
FCNN > BRCAPRO	0.907	0.872	0.891	0.703
CNN > LR	0.668	0.697	0.773	0.893
CNN > BRCAPRO	0.864	0.943	0.891	0.758
Clinic− Based (39 cases)				
Performance Metrics				
FCNN	1.49 (1.08, 1.97)	− 7.02 (− 23.59, 14.35)	3.08 (− 44.82, 107.03)	0.07 (− 0.46, 0.67)
CNN	1.40 (1.01, 1.84)	− 1.65 (− 17.46, 18.21)	17.62 (− 37.69, 164.32)	0.24 (− 0.35, 0.90)
LR	1.29 (0.94, 1.71)	− 5.49 (− 23.12, 14.61)	− 5.46 (− 46.78, 62.30)	0.05 (− 0.58, 0.64)
BRCAPRO	1.38 (1.00, 1.84)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	− 0.04 (− 0.07, − 0.00)
BRCAPRO^C	1.27 (0.92, 1.69)	AUC = 0.619	PR− AUC = 0.033	sqrt(BS) = 0.146
Comparisons Across Bootstrap Replicates				
FCNN > CNN	0.016	0.084	0.024	0.168
FCNN > LR	0.023	0.399	0.584	0.753
FCNN > BRCAPRO	0.024	0.233	0.607	0.552
CNN > LR	0.038	0.845	0.964	0.955
CNN > BRCAPRO	0.046	0.420	0.803	0.693

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 noncounselees 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.

Model performance can be highly dependent on how similar the test population is to the training population (Castaldi, Dahabreh and Ioannidis (2011), Bernau et al. (2014)). In practice, the training and test datasets are often representative of distinct populations with different characteristics. Some methodologies are more robust to these differences than others (Yu (2013), Trippa et al. (2015)). Our application is an example of training and testing using data from different populations: the Risk Service represents a lower-risk population than the test data from the CGN which specifically recruited participants with a family history of cancer. An advantage of training and testing in populations with different characteristics is that it allows us to evaluate how robust the model is to heterogeneity across populations. Despite the differences between the Risk Service and the CGN, the NNs trained in the Risk Service achieved comparable discriminatory accuracy to BRCAPRO which uses parameter estimates based on higher-risk populations. Also, various methods have been developed to adjust for differences between the training and test populations (Janssen et al. (2008), Sugiyama, Krauledat and Müller (2007), Zhang et al. (2013)) which can help improve predictions.

One challenging problem we have not investigated in this paper is ascertainment, or the sampling mechanism. Pedigree-based studies of cancer risk typically use inclusion criteria that enrich for the genotypes and/or phenotypes of interest (e.g., including only families with affected members). This can lead to ascertainment bias, that is, risk estimates that are not generalizable to the population of interest. In particular, when developing pedigree-based risk prediction models, there can be differences in ascertainment between training and test datasets, and not adjusting for these differences can affect performance (especially calibration) in the test dataset. There is an extensive literature on methods for adjusting for ascertainment (Carayol and Bonaïti-Pellié (2004), Choi, Kopciuk and Briollais (2008), Kraft and Thomas (2000), Le Bihan et al. (1995), Iversen and Chen (2005)). One approach for obtaining general population estimates from an ascertained population is to weight families by the inverse probability of being ascertained (Choi, Kopciuk and Briollais (2008)). This approach has similarities to weighting approaches that adjust for differences in covariate distributions between training and tests sets (Sugiyama, Krauledat and Müller (2007), Zhang et al. (2013)) and can be applied during training by using weights in the calculation of the loss function. However, the approach requires a model for the ascertainment mechanism, which is generally unknown or difficult to quantify, and is not directly applicable to existing models, such as BRCAPRO. In our data application, ascertainment differed for the training and test datasets. The Risk Service counselees mostly came from mammography screening populations while the CGN counselees were ascertained based on having a family history of cancer. Moreover, there was heterogeneous ascertainment in both cohorts, since the Risk Service includes some counselees from genetic counseling clinics and the CGN used both population-based and clinic-based recruitment. We took some steps to address the ascertainment differences between the Risk Service and the CGN by recalibrating the models trained in the Risk Service before applying them to the CGN. However, this did not perfectly calibrate the models, especially for the clinic-based subset of the CGN, highlighting the challenge of quantifying ascertainment.

While NNs allow for greater flexibility than Mendelian models and traditional regression models and do not require prior biological understanding, one disadvantage of NNs is that

their black box nature makes it challenging to interpret the relationship between the predictors and risk predictions (Fan, Xiong and Wang (2020)). In contrast, traditional regression methods, such as logistic regression, explicitly describe monotone relationships between the predictors and risk predictions. Various post hoc methods have been developed to determine feature importance in black box models (Ribeiro, Singh and Guestrin (2016), Shrikumar, Greenside and Kundaje (2017)). Methods have also been proposed for developing NN models that are intrinsically interpretable (Dong et al. (2017), Zhang, Nian Wu and Zhu (2018), Li et al. (2018)), but further investigation is needed in the context of family history-based cancer risk prediction.

Other disadvantages of NNs include computational burden (especially in the case of CNNs) and sample size requirements. Our simulations and data application suggest that NNs need large sample sizes (~ 100,000 or more) to achieve good accuracy in family history-based cancer risk prediction. In the data application the FCNN performed particularly poorly when the training set was restricted from over 200,000 families to 40,000 families, though the CNN was still able to achieve reasonable performance. The potential benefits of using NNs are currently limited to a small number of diseases for which many pedigrees are available. In healthcare, datasets with over 100,000 pedigrees exist yet are still uncommon since collecting detailed and accurate family history is a time-consuming process. Examples besides the Risk Service include the Breakthrough Generations breast cancer study, which includes over 113,000 women (Swerdlow et al. (2011)), the Swedish Family-Cancer Database, which includes over two million families (Dong and Hemminki (2001)), cancer studies based on the Utah Population Database, which includes over 1.3 million probands (Cannon-Albright, Carr and Akerley (2019), Teerlink et al. (2012)), and a cancer study based on an Icelandic genealogical database with over 600,000 individuals (Amundadottir et al. (2004)). Though sample sizes are currently limited for most diseases, in recent years, extensive progress has been made to improve and expand family health history collection, including growing efforts in systematic data collection by research consortia (John et al. (2004), Petersen et al. (2006), Newcomb et al. (2007)) and genetic testing companies (Ginsburg, Wu and Orlando (2019)), the development of a wide array of electronic patient-facing family history tools Welch et al. (2018), which allow patients to gather family history information outside the clinic and, therefore, overcome the time constraints of traditional approaches where practitioners record family history during clinical visits, and the implementation of technology allowing for communication between family history tools and electronic health records (Mandel et al. (2016)). Also, electronic genealogical databases are rapidly expanding, and there are continuing efforts to link them with clinical data to generate pedigrees (Amundadottir et al. (2004), Stefansdottir et al. (2013, 2019), Teerlink et al. (2012)). These developments will lead to increased opportunities to refine NN models for hereditary cancer and to train NN models for other hereditary diseases.

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

Acknowledgments. Danielle Braun and Lorenzo Trippa contributed equally. Giovanni Parmigiani and Lorenzo Trippa are also affiliated with the Department of Biostatistics at the Harvard T.H. Chan School of Public Health and Danielle Braun is also affiliated with the Department of Data Sciences at Dana-Farber Cancer Institute. The authors thank the Editor, Associate Editor, and reviewers for their constructive comments and suggestions. The authors also thank Matthew Ploenzke for helpful suggestions.

Funding. Work supported by the Friends of Dana-Farber Fund, NSF Award 1810829, and NSERC PGSD35023622017.

SUPPLEMENTARY MATERIAL

Supplement to "Prediction of hereditary cancers using neural networks" (DOI: 10.1214/21-AOAS1510SUPPA; .pdf). The supplementary material includes a notation table, an overview of standard CNNs, the proof of Theorem 2.1, and additional details on the simulations and data application.

Code for "Prediction of hereditary cancers using neural networks" (DOI: 10.1214/21-AOAS1510SUPPB; .zip). The simulation code is included as a zip file. It is also available at github.com/zoeguan/nn_cancer_risk.

REFERENCES

AMUNDADOTTIR, L. T., THORVALDSSON, S., GUDBJARTSSON, D. F., SULEM, P., KRISTJANSSON, K., ARNASON, S., GULCHER, J. R., BJORNSSON, J., KONG, A. et al. (2004). Cancer as a complex phenotype: Pattern of cancer distribution within and beyond the nuclear family. *PLoS Med.* **1** e65.

ANTON-CULVER, H., ZIOGAS, A., BOWEN, D., FINKELSTEIN, D., GRIFFIN, C., HANSON, J., ISAACS, C., KASTEN-SPORTES, C., MINEAU, G. et al. (2003). The cancer genetics network: Recruitment results and pilot studies. *Publ. Health Genomics* **6** 171–177.

ANTONIOU, A. C., PHAROAH, P. D. P., MCMULLAN, G., DAY, N. E., STRATTON, M. R., PETO, J., PONDER, B. J. and EASTON, D. F. (2002). A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *Br. J. Cancer* **86** 76–83. https://doi.org/10.1038/sj.bjc.6600008

ANTONIOU, A. C., PHAROAH, P. P. D., SMITH, P. and EASTON, D. F. (2004). The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *Br. J. 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. 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. et al. (2017). Projecting individualized absolute invasive breast cancer risk in US Hispanic women. *J. Natl. Cancer Inst.* **109** djw215.

BERGSTRA, J. and BENGIO, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13** 281–305. MR2913701

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.

BERRY, D. A., PARMIGIANI, G., SANCHEZ, J., SCHILDKRAUT, J. and WINER, E. (1997). Probability of carrying a mutation of breast-ovarian cancer gene BRCA1 based on family history. *J. Natl. Cancer Inst.* **89** 227–237.

BERRY, D. A., IVERSEN JR., E. S., GUDBJARTSSON, D. F., HILLER, E. H., GARBER, J. E., PESHKIN, B. N., LERMAN, C., WATSON, P., LYNCH, H. T. et al. (2002). BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. *J. Clin. Oncol.* **20** 2701–2712.

BISHOP, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford Univ. Press, New York. MR1385195

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 Res. Treat.* **139** 571–579.

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. (2018). Nonparametric adjustment for measurement error in time-to-event data: Application to risk prediction models. *J. Amer. Statist. Assoc.* **113** 14–25. MR3803436 https://doi.org/10.1080/01621459.2017.1311261

BRENTNALL, A. R., COHN, W. F., KNAUS, W. A., YAFFE, M. J., CUZICK, J. and HARVEY, J. A. (2019). A case-control study to add volumetric or clinical mammographic density into the Tyrer–Cuzick breast cancer risk model. *J. Breast Imaging* **1** 99–106.

CANNON-ALBRIGHT, L. A., CARR, S. R. and AKERLEY, W. (2019). Population-based relative risks for lung cancer based on complete family history of lung cancer. *J. Thorac. Oncol.* **14** 1184–1191. https://doi.org/10.1016/j.jtho.2019.04.019

CARAYOL, J. and BONAÏTI-PELLIÉ, C. (2004). Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. *Genet. Epidemiol.* **27** 109–117.

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press/CRC, Boca Raton, FL. MR2243417 https://doi.org/10.1201/9781420010138

CASTALDI, P. J., DAHABREH, I. J. and IOANNIDIS, J. P. A. (2011). An empirical assessment of validation practices for molecular classifiers. *Brief. Bioinform.* **12** 189–202. https://doi.org/10.1093/bib/bbq073

CHEN, S. and PARMIGIANI, G. (2007). Meta-analysis of BRCA1 and BRCA2 penetrance. *J. Clin. Oncol.* **25** 1329.

CHEN, S., WANG, W., BROMAN, K. W., KATKI, H. A. and PARMIGIANI, G. (2004). BayesMendel: An R environment for Mendelian risk prediction. *Stat. Appl. Genet. Mol. Biol.* **3** Art. 21, 21. MR2101490 https://doi.org/10.2202/1544-6115.1063

CHEN, S., WANG, W., LEE, S., NAFA, K., LEE, J., ROMANS, K., WATSON, P., GRUBER, S. B., EUHUS, D. et al. (2006). Prediction of germline mutations and cancer risk in the Lynch syndrome. *JAMA J. Am. Med. Assoc.* **296** 1479.

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 Spectr.* **4** pkaa029. https://doi.org/10.1093/jncics/pkaa029

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 Res. Treat.* **140** 187–193.

CHOI, J., DEKKERS, O. M. and LE CESSIE, S. (2019). A comparison of different methods to handle missing data in the context of propensity score analysis. *Eur. J. Epidemiol.* **34** 23–36. https://doi.org/10.1007/s10654-018-0447-z

CHOI, Y.-H., KOPCIUK, K. A. and BRIOLLAIS, L. (2008). Estimating disease risk associated with mutated genes in family-based designs. *Hum. Hered.* **66** 238–251.

CHOUDHURY, P. P., MAAS, P., WILCOX, A., WHEELER, W., BROOK, M., CHECK, D., GARCIA-CLOSAS, M. and CHATTERJEE, N. (2020a). iCARE: An R package to build, validate and apply absolute risk models. *PLoS ONE* **15** e0228198.

CHOUDHURY, P. P., WILCOX, A. N., BROOK, M. N., ZHANG, Y., AHEARN, T., ORR, N., COULSON, P., SCHOEMAKER, M. J., JONES, M. E. et al. (2020b). Comparative validation of breast cancer risk prediction models and projections for future risk stratification. *J. Natl. Cancer Inst.* **112** 278–285.

CLAUS, E. B., RISCH, N. and THOMPSON, W. D. (1994). Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer* **73** 643–651.

CLEVERT, D.-A., UNTERTHINER, T. and HOCHREITER, S. (2015). Fast and accurate deep network learning by exponential linear units (ELUs). Preprint. Available at arXiv:1511.07289.

CYBENKO, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** 303–314. MR1015670 https://doi.org/10.1007/BF02551274

DONG, C. and HEMMINKI, K. (2001). Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families. *Int. J. Cancer* **92** 144–150.

DONG, Y., SU, H., ZHU, J. and BAO, F. (2017). Towards interpretable deep neural networks by leveraging adversarial examples. Preprint. Available at arXiv:1708.05493.

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

EASTON, D. F., FORD, D. and BISHOP, D. T. (1995). Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* **56** 265–271.

ELSTON, R. C. and STEWART, J. (1971). A general model for the genetic analysis of pedigree data. *Hum. Hered.* **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. et al. (2002). Pretest prediction of BRCA1 or BRCA2 mutation by risk counselors and the computer model BRCAPRO. *J. Natl. Cancer Inst.* **94** 844–851.

FAN, F., XIONG, J. and WANG, G. (2020). On interpretability of artificial neural networks. Available at https://arxiv.org/abs/2001.02522.

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. *J. Natl. Cancer Inst.* **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. et al. (2007). Projecting individualized absolute invasive breast cancer risk in African American women. *J. Natl. Cancer Inst.* **99** 1782–1792.

GARCÍA-LAENCINA, P. J., SANCHO-GÓMEZ, J. and FIGUEIRAS-VIDAL, A. R. (2010). Pattern classification with missing data: A review. *Neural Comput. Appl.* **19** 263–282.

GEMAN, S., BIENENSTOCK, E. and DOURSAT, R. (1992). Neural networks and the bias/variance dilemma. *Neural Comput.* **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. *Biom. J.* **48** 1029–1040. MR2312613 https://doi.org/10.1002/bimj.200610301

GINSBURG, G. S., WU, R. R. and ORLANDO, L. A. (2019). Family health history: Underused for actionable risk assessment. *Lancet* **394** 596–603.

GUAN, Z., PARMIGIANI, G., BRAUN, D. and TRIPPA, L. (2022a). Supplement to "Prediction of hereditary cancers using neural networks." https://doi.org/10.1214/21-AOAS1510SUPPA

GUAN, Z., PARMIGIANI, G., BRAUN, D. and TRIPPA, L. (2022b). Supplement to "Prediction of hereditary cancers using neural networks." https://doi.org/10.1214/21-AOAS1510SUPPB

HAMPSHIRE II, J. B. and PEARLMUTTER, B. (1991). Equivalence proofs for multi-layer perceptron classifiers and the Bayesian discriminant function. In *Connectionist Models* 159–172. Elsevier, Amsterdam.

HECHTLINGER, Y., CHAKRAVARTI, P. and QIN, J. (2017). A generalization of convolutional neural networks to graph-structured data. Preprint. Available at arXiv:1704.08165.

HINTON, G., DENG, L., YU, D., DAHL, G. E., MOHAMED, A., JAITLY, N., SENIOR, A., VANHOUCKE, V., NGUYEN, P. et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **29** 82–97.

HORNER, M. J., RIES, L. A. G., KRAPCHO, M., NEYMAN, N., AMINOU, R., HOWLADER, N., ALTEKRUSE, S. F., FEUER, E. J., HUANG, L. et al. (2009). SEER cancer statistics review, 1975–2006. National Cancer Institute, Bethesda, MD.

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

IVERSEN, E. S. JR. and CHEN, S. (2005). Population-calibrated gene characterization: Estimating age at onset distributions associated with cancer genes. *J. Amer. Statist. Assoc.* **100** 399–409. MR2170463 https://doi.org/10.1198/016214505000000196

JANOCHA, K. and CZARNECKI, W. M. (2017). On loss functions for deep neural networks in classification. Preprint. Available at arXiv:1702.05659.

JANSSEN, K. J. M., MOONS, K. G. M., KALKMAN, C. J., GROBBEE, D. E. and VERGOUWE, Y. (2008). Updating methods improved the performance of a clinical prediction model in new patients. *J. Clin. Epidemiol.* **61** 76–86. https://doi.org/10.1016/j.jclinepi.2007.04.018

JOHN, E. M., HOPPER, J. L., BECK, J. C., KNIGHT, J. A., NEUHAUSEN, S. L., SENIE, R. T., ZIOGAS, A., ANDRULIS, I. L., ANTON-CULVER, H. et al. (2004). The Breast Cancer Family Registry: An infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res.* **6** R375.

KATKI, H. A. (2006). Effect of misreported family history on Mendelian mutation prediction models. *Biometrics* **62** 478–487. MR2236830 https://doi.org/10.1111/j.1541-0420.2005.00488.x

KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.* **23** 462–466. MR0050243 https://doi.org/10.1214/aoms/1177729392

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at arXiv:1412.6980.

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* 27*th Annual Conference* 2417–2420. IEEE.

KRAFT, P. and THOMAS, D. C. (2000). Bias and efficiency in family-based gene-characterization studies: Conditional, prospective, retrospective, and joint likelihoods. *Am. J. Hum. Genet.* **66** 1119–1131.

KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 1097–1105.

LE BIHAN, C., MOUTOU, C., BRUGIÈRES, L., FEUNTEUN, J. and BONAÏTI-PELLIÉ, C. (1995). ARCAD: A method for estimating age-dependent disease risk associated with mutation carrier status from family data. *Genet. Epidemiol.* **12** 13–25.

LECUN, Y., BOTTOU, L., BENGIO, Y., HAFFNER, P. et al. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* **86** 2278–2324.

LESHNO, M., LIN, V. Y., PINKUS, A. and SCHOCKEN, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.* **6** 861–867.

LI, Y., ZHANG, T., LIU, Z. and HU, H. (2017). A concatenating framework of shortcut convolutional neural networks. Preprint. Available at arXiv:1710.00974.

LI, O., LIU, H., CHEN, C. and RUDIN, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence* **32**. 1.

LITTLE, R. J. and RUBIN, D. B. (2019). *Statistical Analysis with Missing Data* **793**. Wiley, New York.

MANDEL, J. C., KREDA, D. A., MANDL, K. D., KOHANE, I. S. and RAMONI, R. B. (2016). SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. *J. Am. Med. Inform. Assoc.* **23** 899–908. https://doi.org/10.1093/jamia/ocv189

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. *J. Natl. Cancer Inst.* **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. et al. (2019). Performance of breast cancer risk assessment models in a large mammography cohort. *J. Natl. Cancer Inst.*

MIKI, Y., SWENSEN, J., SHATTUCK-EIDENS, D., FUTREAL, P. A., HARSHMAN, K., TAVTIGIAN, S., LIU, Q., COCHRAN, C., BENNETT, L. M. et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 66–71.

NEWCOMB, P. A., BARON, J., COTTERCHIO, M., GALLINGER, S., GROVE, J., HAILE, R., HALL, D., HOPPER, J. L., JASS, J. et al. (2007). Colon Cancer Family Registry: An international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol. Biomark. Prev.* **16** 2331–2343.

NIELSEN, M. A. (2015). *Neural Networks and Deep Learning* **25**. Determination Press, San Francisco, CA.

NIEPERT, M., AHMED, M. and KUTZKOV, K. (2016). Learning convolutional neural networks for graphs. In *International Conference on Machine Learning* 2014–2023.

OH, K.-S. and JUNG, K. (2004). GPU implementation of neural networks. *Pattern Recognit.* **37** 1311–1314.

COLLABORATIVE GROUP ON HORMONAL FACTORS IN BREAST CANCER (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. *Lancet* **358** 1389–1399.

PARMIGIANI, G., BERRY, D. A. and AGUILAR, O. (1998). Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *Am. J. Hum. Genet.* **62** 145–158.

PATRO, S. and SAHU, K. K. (2015). Normalization: A preprocessing stage. Preprint. Available at arXiv:1503.06462.

PETERSEN, G. M., DE ANDRADE, M., GOGGINS, M., HRUBAN, R. H., BONDY, M., KORCZAK, J. F., GALLINGER, S., LYNCH, H. T., SYNGAL, S. et al. (2006). Pancreatic cancer genetic epidemiology consortium. *Cancer Epidemiol. Biomark. Prev.* **15** 704–710.

PICHERT, G., BOLLIGER, B., BUSER, K., PAGANI, O. and SWISS INSTITUTE FOR APPLIED CANCER RESEARCH NETWORK FOR CANCER PREDISPOSITION TESTING (2003). Evidence-based management options for women at increased breast/ovarian cancer risk. *Ann. Oncol.* **14** 9–19. https://doi.org/10.1093/annonc/mdg030

PORTNOI, T., YALA, A., SCHUSTER, T., BARZILAY, R., DONTCHOS, B., LAMB, L. and LEHMAN, C. (2019). Deep learning model to assess cancer risk on the basis of a breast MR image alone. *Am. J. Roentgenol.* **213** 227–233. https://doi.org/10.2214/AJR.18.20813

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 Res.* **14** R144. https://doi.org/10.1186/bcr3352

RIBEIRO, M. T., SINGH, S. and GUESTRIN, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144.

SHRIKUMAR, A., GREENSIDE, P. and KUNDAJE, A. (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning* 3145–3153. PMLR.

SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I. and SALAKHUTDINOV, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15** 1929–1958. MR3231592

STEFANSDOTTIR, V., JOHANNSSON, O. T., SKIRTON, H., TRYGGVADOTTIR, L., TULINIUS, H. and JONSSON, J. J. (2013). The use of genealogy databases for risk assessment in genetic health service: A systematic review. *J. Community Genet.* **4** 1–7. https://doi.org/10.1007/s12687-012-0103-3

STEFANSDOTTIR, V., SKIRTON, H., JOHANNSSON, O. T., OLAFSDOTTIR, H., OLAFSDOTTIR, G. H., TRYGGVADOTTIR, L. and JONSSON, J. J. (2019). Electronically ascertained extended pedigrees in breast cancer genetic counseling. *Fam. Cancer* **18** 153–160. https://doi.org/10.1007/s10689-018-0105-3

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* **21** 128.

SUGIYAMA, M., KRAULEDAT, M. and MÜLLER, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* **8**.

SWERDLOW, A. J., JONES, M. E., SCHOEMAKER, M. J., HEMMING, J., THOMAS, D., WILLIAMSON, J. and ASHWORTH, A. (2011). The breakthrough generations study: Design of a long-term UK cohort study to investigate breast cancer aetiology. *Br. J. Cancer* **105** 911–917.

TEAM, T. T. D., AL-RFOU, R., ALAIN, G., ALMAHAIRI, A., ANGERMUELLER, C., BAHDANAU, D., BALLAS, N., BASTIEN, F., BAYER, J. et al. (2016). Theano: A Python framework for fast computation of mathematical expressions. Preprint. Available at arXiv:1605.02688.

TEERLINK, C. C., ALBRIGHT, F. S., LINS, L. and CANNON-ALBRIGHT, L. A. (2012). A comprehensive survey of cancer risks in extended families. *Genet. Med.* **14** 107–114.

TERRY, M. B., LIAO, Y., WHITTEMORE, A. S., LEOCE, N., BUCHSBAUM, R., ZEINOMAR, N., DITE, G. S., CHUNG, W. K., KNIGHT, J. A. et al. (2019). 10-year performance of four models of breast cancer risk: A validation study. *Lancet Oncol.* **20** 504–517.

TICE, J. A., CUMMINGS, S. R., SMITH-BINDMAN, R., ICHIKAWA, L., BARLOW, W. E. and KERLIKOWSKE, K. (2008). Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model. *Ann. Intern. Med.* **148** 337–347.

TRIPPA, L., WALDRON, L., HUTTENHOWER, C. and PARMIGIANI, G. (2015). Bayesian nonparametric cross-study validation of prediction methods. *Ann. Appl. Stat.* **9** 402–428. MR3341121 https://doi.org/10.1214/14-AOAS798

TYRER, J., DUFFY, S. W. and CUZICK, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Stat. Med.* **23** 1111–1130.

UNO, H., CAI, T., TIAN, L. and WEI, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *J. Amer. Statist. Assoc.* **102** 527–537. MR2370850 https://doi.org/10.1198/016214507000000149

WANG, W., CHEN, S., BRUNE, K. A., HRUBAN, R. H., PARMIGIANI, G. and KLEIN, A. P. (2007). PancPRO: Risk assessment for individuals with a family history of pancreatic cancer. *J. Clin. Oncol.* **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 Res.* **70** 552–559.

WELCH, B. M., WILEY, K., PFLIEGER, L., ACHIANGIA, R., BAKER, K., HUGHES-HALBERT, C., MORRISON, H., SCHIFFMAN, J. and DOERR, M. (2018). Review and comparison of electronic patient-facing family health history tools. *J. Genet. Couns.* **27** 381–391. https://doi.org/10.1007/s10897-018-0235-7

WOOSTER, R., BIGNELL, G., LANCASTER, J., SWIFT, S., SEAL, S., MANGION, J., COLLINS, N., GREGORY, S., GUMBS, C. 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. (2021). A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **32** 4–24. MR4205495

YALA, A., LEHMAN, C., SCHUSTER, T., PORTNOI, T. and BARZILAY, R. (2019). A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292** 60–66. https://doi.org/10.1148/radiol.2019182716

YU, B. (2013). Stability. *Bernoulli* **19** 1484–1500. MR3102560 https://doi.org/10.3150/13-BEJSP14

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* 8827–8836.

ZHANG, K., SCHÖLKOPF, B., MUANDET, K. and WANG, Z. (2013). Domain adaptation under target and conditional shift. In *International Conference on Machine Learning* 819–827. PMLR.

ZIOGAS, A. and ANTON-CULVER, H. (2003). Validation of family history data in cancer family registries. *Am. J. Prev. Med.* **24** 190–198.