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 [10, 22, 95] 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 [10, 93, 4, 95, 96] and 2) using summaries of family history (for example, the number of relatives with a previous cancer diagnosis) as covariates in regression models [38, 5, 39, 66, 6, 91, 26, 27]. Recently, deep learning models based on mammographic images have also been proposed [78, 100].

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 (for example, see [20]). The individual risk of cancer is then calculated as a weighted average of mutation carrier and non-carrier risks of developing cancer. Mendelian models are typically recommended over regression-based models for individuals with a strong family history of cancer,

**Keywords and phrases:** Machine learning, Mendelian risk prediction, family history.
since Mendelian models use more detailed family history information [79, 77]. 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 [34], and existing Mendelian models consider only a small subset of these genes. Furthermore, Mendelian models are sensitive to misreporting of family history [13, 53] 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 [11, 70] 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 non-linear 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 [30, 48, 61], and in practice they have achieved state-of-the-art performance in many tasks, such as image recognition [58] and natural language processing [46]. 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. [56] 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 [74], 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.

2.1. Notation. Our notation is summarized in Table A.1 of the supplementary material (SM). 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 pre-specified 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 [11] or [70] 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 family history $H$). The final layer, which provides the output in the form of prediction, is called the output layer. The layers in between, which are optional layers 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, 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

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

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

where $W^l \in \mathbb{R}^{N_l \times N_{l-1}}$ 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 $\phi^l : \mathbb{R}^{N_l} \rightarrow \mathbb{R}^{N_l}$ represents the component-wise application of an activation function $\phi : \mathbb{R} \rightarrow \mathbb{R}$. Commonly used activation functions include the logistic function $\sigma(z) = \frac{\exp(z)}{1 + \exp(z)}$ and the rectifier function $ReLU(z) = \max(0, z)$.

The output layer ($l = L + 1$) consists of a single node that uses the logistic activation function, outputting the predicted probability

$$\hat{Y}_0 = a^{L+1} = \sigma(W^{L+1}a^L + b^{L+1}).$$
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 [54] and the Adam optimizer [55]. Examples of cost functions [50] 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, then it is appropriate to interpret the NN output as a probability [44].

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 [82].

2.3. Standardizing and Flattening Pedigrees. Since FCNNs require a fixed-size input, they cannot be directly applied to pedigrees, which vary in size and structure. It is possible to generate a fixed-size input based on simple summaries of family history, but this can result in substantial loss of information. Therefore, we propose the following approach: define a reference pedigree with pre-specified relatives (for example: 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.

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
correlated values), a convolutional layer applies the same functions (e.g., correlation structure of the input (for example, pixels that are spatially close often have highly
weights are set to 0 and certain weights are constrained to have the same value. To exploit the
number of parameters by using convolutional layers that enforce selective connections and
problems where the input has a spatial structure, such as image classification, reduce the
of parameters grows quickly with network size [41]. CNNs [60], which are widely used in
pixels). These functions are called convolutional filters. The number of parameters in these
repeatedly to different fixed-size neighborhoods of the input (for example, sets of neighboring
which can be used as input to a FCNN. If
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, $H'$, with the same structure as the reference
pedigree matrix, we compare the number of relatives of type $q$ in the actual pedigree to the
number in the reference pedigree for each $q \in \{0, 1, \ldots, Q'\}$. If the two numbers are the
same ($R_q = R_q'$), then we include all of the $R_q'$ actual relatives in $H'$. If the actual number
is smaller than the reference number ($R_q < R_q'$), then we include the $R_q$ actual relatives in
$H'$ and represent each of the $R_q' - R_q$ absent relatives using a vector of pre-specified null
values (zeros). If the actual number is larger than the reference number ($R_q > R_q'$), then we
randomly select $R_q'$ of the actual relatives to include in $H'$. We also include a column in $H'$ to
indicate whether each row corresponds to a relative who is absent from the actual pedigree (0
if present, 1 if absent). Therefore, $H'$ is an $R' + 1$ by $K + 1$ matrix where each row consists of
a family member’s $K$ cancer history features, along with the presence/absence indicator.
Let $H'_r$ be the vector for relative $r$ in $H'$. We flatten $H'$ by concatenating its rows to get a
vector, $X = (H'_0, H'_1, \ldots, H'_{R'}) \in \mathbb{R}^{(R' + 1) \times (K + 1)}$, 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 [41]. CNNs [60], which are widely used in
problems where the input has a spatial structure, such as image classification, reduce the
number of parameters by using convolutional layers that enforce selective connections and
weight sharing. A convolutional layer can be viewed as a fully-connected layer where certain
weights are set to 0 and certain weights are constrained to have the same value. To exploit the
correlation structure of the input (for example, pixels that are spatially close often have highly
correlated values), a convolutional layer applies the same functions (e.g., $x \rightarrow \max(0, wx)$)
repeatedly to different fixed-size neighborhoods of the input (for example, sets of neighboring
pixels). These functions are called convolutional filters. The number of parameters in these
Fig 2: Consider a reference pedigree that includes the counselee’s grandparents, parents, uncles, aunts, and siblings, with each couple having 2 children of each sex. (A) Actual pedigree \( H \). (B) Standardized pedigree \( H' \), obtained by mapping \( H \) to the reference structure. The actual pedigree has more maternal aunts than the reference pedigree, so we randomly select the desired number of maternal aunts to include in \( H' \). The actual pedigree has fewer paternal aunts, sisters, and brothers than the reference pedigree, so in \( H' \), we use pre-specified non-informative values for the paternal aunts, sisters, and brothers absent from \( H \). (C) Flattened pedigree \( X \), which is used as input for a FCNN.

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 (for example, 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 SM A.2. 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 [99, 71], 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 \( H' \), 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 pre-specified index representing an
absent relative whose features are set to zero (analogous to zero padding in standard CNNs, as described in SM A.2). 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 \( f^l_1, \ldots, f^l_{M_l} \) to each of the \( R' + 1 \) neighborhoods of the pedigree (Figure 3). For \( i = 1, \ldots, M_l \), let \( f^l_i : \mathbb{R}^{U*M_{l-1}} \rightarrow \mathbb{R} \) (let \( M_0 = K + 1 \) since each relative has \( K + 1 \) features in \( H' \) - see Section 2.3). Let \( a^l_r \in \mathbb{R}^{M_l} \) be the output of layer \( l \) for neighborhood/family member \( r \). Let \( a^{l-1}_{N(r)} \in \mathbb{R}^{U*M_{l-1}} \) 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

\[
f^l_i(a^{l-1}_{N(r)}) = \phi \left( w^l_i \cdot a^{l-1}_{N(r)} + b^l_i \right),
\]

where \( \cdot \) is the dot product, \( w^l_i \in \mathbb{R}^{U*M_{l-1}} \) is the vector of weights for filter \( i \) and \( b^l_i \in \mathbb{R} \) is the bias for filter \( i \).

Let \( f^l = (f^l_1, \ldots, f^l_{M_l}) : \mathbb{R}^{U*M_{l-1}} \rightarrow \mathbb{R}^{M_l} \). The layer outputs for relative \( r \) are

\[
a^l_r = H^l_r \in \mathbb{R}^{K+1}, \quad (l = 0, \text{and})
\]

\[
a^l_r = f^l(a^{l-1}_{N(r)}) \in \mathbb{R}^{M_l}, \quad (l = 1, \ldots, L),
\]

and the overall layer outputs are

\[
a^l = \left( a^l_0, \ldots, a^l_R \right) \in \mathbb{R}^{M_l*(R'+1)}, \quad (l = 0, 1, \ldots, L).
\]

The final output is a transformation of \( a^L_0 \) using a logistic activation function:

\[
\hat{Y}_0 = \sigma(w^{L+1} \cdot a^L_0 + b^{L+1})
\]

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

As in FCNNs, the weight and bias parameters are optimized with respect to \( \sum_{m=1}^{M} C(Y_{0m}, \hat{Y}_{0m}) \) 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 \( a^0_0 \) 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 non-counselees or 2) append the features for the counselee to the output vector of the \( L \)-th 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 [63]).

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 [30, 48, 61]. 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 [61] and then verify that the same property extends to CNNs (proof provided in SM A.3).

2.4.1.1. Universal Approximation Theorem for FCNN.\footnote{[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 $g : I \to \mathbb{R}$ be continuous. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a piecewise continuous, locally bounded, and non-polynomial activation function. Then given $\epsilon > 0$, there exists a positive integer $N$, and for $i = 1, \ldots, N$, constants $\alpha_i, b_i \in \mathbb{R}$ and vectors $w_i \in \mathbb{R}^k$ such that

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

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

**Theorem 2.1.** 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 $\phi : \mathbb{R} \to \mathbb{R}$ be a continuous and invertible activation function. Let the fixed-size neighborhood about each relative contain $m_1 = \cdots = m_4 = m$ sisters/brothers/daughters/sons. Then
given $\epsilon > 0$, there exists a pedigree CNN of the form described in subsection 2.4 with $d$ hidden layers with activation function $\phi$, $M_l$ 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 $w^i_l \in \mathbb{R}^{U \times M_{l-1}}$ ($i = 1, \ldots, M_l$; $l = 1, \ldots, L + 1$), such that the final output

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

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

2.5. Missing Data. In practice, there is often missing information in family history data (for example, 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 [64], 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 [40]. In clinical practice, some models do not allow missing values (e.g. the Claus model [28]), 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 [23].

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 [24]. 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 2.9, 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 [10, 74] is widely used in clinical practice and has been validated in various populations [9, 36, 90, 67]. 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 $\gamma_r$ be the genotype of relative $r$ (non-carrier, 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 $\gamma_0$ is

$$P(\gamma_0|H) = \frac{P(\gamma_0) \sum_{\gamma_1, \ldots, \gamma_R} \prod_{r=0}^R P(H_r|\gamma_r) P(\gamma_1, \ldots, \gamma_R|\gamma_0)}{\sum_{\gamma_0} P(\gamma_0) \sum_{\gamma_1, \ldots, \gamma_R} \prod_{r=0}^R P(H_r|\gamma_r) P(\gamma_1, \ldots, \gamma_R|\gamma_0)}.$$
The summation over genotypes is calculated using the Elston-Stewart peeling algorithm [35] and $P(\gamma_1, \ldots, \gamma_R | \gamma_0)$ 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- and sex-specific. The penetrance functions for non-carriers are based on rates from the Surveillance, Epidemiology, and End Results (SEER) program and are race-specific, while the penetrance functions for carriers are from a meta-analysis of published studies [19].

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

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

2.7. Model Evaluation. We evaluated model performance using four metrics [85]: 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) [21]. The NNs were implemented in Python using Keras (https://github.com/keras-team/keras) with the Theano backend [88]. For the CNNs, we adapted code from [45].

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 [55] 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 [85] and one of the metrics we used to compare models; more discussion on this choice and a sensitivity analysis are provided in SM B.4). We used a typical 90/10 split of the training set to tune NN hyperparameters via a random search [7]: 10% of the training set was held out for evaluating the performance of different choices for the number of hidden layers (1 to 3), sizes of hidden layers (10 to 100), number of filters for the CNN (3 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 2 hidden layers of sizes 30 and 10 and the CNNs had 2 convolutional layers with 10 and 5 filters. In the data application, the FCNN had 2 hidden layers of size 30 and the CNN had 2 convolutional layers with 5 filters each. We also used a dropout layer following the first hidden layer in each NN, with a dropout rate of 20%. We used the Exponential Linear Unit (ELU) activation function [29]. For the NNs and LR,
features were normalized to be between 0 and 1 using min-max normalization [75]. The code for the analyses is 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 (2 maternal, 3 paternal), uncles (3 maternal, 2 paternal), siblings (2 sisters, 3 brothers), and children (2 daughters, 2 sons). This was chosen based on the distribution of family structures in the CGN (see SM B.2). We used \( m_1 = m_2 = 3 \) and \( m_3 = m_4 = 2 \) for the CNN neighborhoods. In the data application, we used a reference pedigree of size 19 with the same relative types as in the simulations, but restricted to 2 relatives of each type and omitted sons and daughters due to the smaller family sizes in the training dataset (see SM B.2). 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 [12, 90] has little impact on discriminative accuracy. Therefore we considered only first- and second-degree relatives in the reference pedigree. As described in Section 2.9, we conducted a sensitivity analysis for various choices of reference pedigree structures and found little variation in performance.

2.9. 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 (SM B.2). 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 non-informative 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 (SM B.3).

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 non-counselees, 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 non-carriers to 0.79 for carriers of mutations in both BRCA1 and BRCA2). We also generated a death age for each individual from a distribution with mean 80 and standard deviation 15. If an individual’s age of onset was greater than their baseline age or death age, then their cancer status at baseline was set to 0.

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.

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 6,250 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 LR underestimated or overestimated risk in more deciles compared to the other models. We also plotted the precision-recall curves for the models (Figure A.2), 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, (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.

![Fig 4: AUC and correlation (\( \rho \)) of NN predictions with BRCAPRO predictions for 10-year risk of developing breast cancer as a function of training sample size (ranging from to 6,250 to 800,000) in simulations.](image)

3.2.1. **Perturbations of Mendelian Models.** Misreported cancer diagnoses can considerably distort predictions from Mendelian models [53, 13]. In the second simulation setting, we introduced noise to the simulated family histories through incorrectly reported diagnoses, diagnosis ages, and current ages for non-counselees using misreporting rates from [104] and [14] (see SM B.1 for details).
Table 1

Model performance in simulated families (training set of 800,000) based on true and misreported family history. 

<table>
<thead>
<tr>
<th></th>
<th>O/E</th>
<th>ΔAUC</th>
<th>ΔPR-AUC</th>
<th>Δsqrt(BS)</th>
<th>ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True Family History</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FCNN</td>
<td>0.93 (0.89, 0.96)</td>
<td>-1.21 (-1.73, -0.63)</td>
<td>-10.16 (-13.81, -7.08)</td>
<td>-0.19 (-0.26, -0.12)</td>
<td>0.90 (0.89, 0.91)</td>
</tr>
<tr>
<td>CNN</td>
<td>0.99 (0.96, 1.03)</td>
<td>-1.24 (-1.80, -0.69)</td>
<td>-7.93 (-11.52, -4.35)</td>
<td>-0.14 (-0.19, -0.10)</td>
<td>0.92 (0.91, 0.92)</td>
</tr>
<tr>
<td>LR</td>
<td>1.00 (0.97, 1.04)</td>
<td>-2.07 (-2.68, -1.47)</td>
<td>-14.59 (-19.04, -10.25)</td>
<td>-0.28 (-0.36, -0.21)</td>
<td>0.82 (0.81, 0.83)</td>
</tr>
<tr>
<td>BRCAPRO</td>
<td>1.02 (0.98, 1.06)</td>
<td>0.668</td>
<td>0.065</td>
<td>0.168</td>
<td>1.00 (1.00, 1.00)</td>
</tr>
</tbody>
</table>

Comparisons Across Bootstrap Replicates

<table>
<thead>
<tr>
<th></th>
<th>AUC=0.668</th>
<th>PR-AUC=0.065</th>
<th>sqrt(BS)=0.168</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCNN&gt;CNN</td>
<td>0.021</td>
<td>0.582</td>
<td>0.038</td>
</tr>
<tr>
<td>FCNN&gt;LR</td>
<td>0.025</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>FCNN&gt;BRCAPRO</td>
<td>0.083</td>
<td>0.000</td>
<td>0.999</td>
</tr>
<tr>
<td>CNN&gt;LR</td>
<td>0.464</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>CNN&gt;BRCAPRO</td>
<td>0.691</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Misreported Family History**

<table>
<thead>
<tr>
<th></th>
<th>AUC=0.627</th>
<th>PR-AUC=0.050</th>
<th>sqrt(BS)=0.169</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCNN&gt;CNN</td>
<td>0.033</td>
<td>0.666</td>
<td>0.232</td>
</tr>
<tr>
<td>FCNN&gt;LR</td>
<td>0.061</td>
<td>0.968</td>
<td>0.869</td>
</tr>
<tr>
<td>FCNN&gt;BRCAPRO</td>
<td>1.000</td>
<td>1.000</td>
<td>0.998</td>
</tr>
<tr>
<td>CNN&gt;LR</td>
<td>0.406</td>
<td>0.900</td>
<td>0.991</td>
</tr>
<tr>
<td>CNN&gt;BRCAPRO</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Comparisons Across Bootstrap Replicates

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FCNN&gt;CNN</td>
<td>0.038</td>
<td>0.666</td>
<td>0.232</td>
</tr>
<tr>
<td>FCNN&gt;LR</td>
<td>0.487</td>
<td>0.968</td>
<td>0.869</td>
</tr>
<tr>
<td>FCNN&gt;BRCAPRO</td>
<td>1.000</td>
<td>1.000</td>
<td>0.998</td>
</tr>
<tr>
<td>CNN&gt;LR</td>
<td>0.495</td>
<td>0.900</td>
<td>0.991</td>
</tr>
<tr>
<td>CNN&gt;BRCAPRO</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

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

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),
Fig 6: Under the first simulation setting, we fixed a simulated family structure (counsellee, grandparents, parents, 1 paternal aunt, 1 maternal aunt, 2 maternal uncles) for a 40-year-old counsellee and varied the level of family history across 5 scenarios (ordered by increasing risk with respect to BRCAPRO, the true model):

1. no affected relatives
2. maternal grandmother diagnosed with breast cancer at age 80
3. maternal grandmother diagnosed with breast cancer at age 60
4. maternal grandmother diagnosed with breast cancer at age 60, mother diagnosed with breast cancer at age 50
5. maternal grandmother diagnosed with breast cancer at age 60, mother diagnosed with breast cancer at age 50 and ovarian cancer at age 60. We calculated 10-year risk predictions for each scenario using each model 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. 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 6,250 to 800,000 are provided in Figure A.3 (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 1 minute to train the LR model, 5 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 [72] 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 [23] 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 5 years prior to the date at which they used the Risk Service and the follow-up date to be the date at which they used the Risk Service. We retrospectively reconstructed the family history at the baseline date based on the ages and diagnosis ages of the family members. However due to a considerable amount of missing age information for non-counselees (74% of first- and second-degree relatives were missing age and 34% of affected first- and second-degree relatives were missing age at diagnosis), we decided not to use ages or diagnosis ages of non-counselees for training and we imputed baseline cancer status for non-counselees with missing diagnosis ages (see SM B.7 for more details).

The training set consisted of 279,460 counselees (Table 2). The median age was 45 and the median family size was 8. 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 [2]. Between 1999 and 2010, 26,941 participants with cancer or a family history of cancer were recruited through population-based registries, high-risk clinics, and self-referral. They provided information on personal and family history of cancer and sociodemographic factors through a baseline phone interview and annual follow-up updates.

The test cohort consisted of 7,489 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 5 years of baseline, and 1017 counselees (13.6%) were lost to follow-up within 5 years without being diagnosed with breast cancer (Table 2). To adjust for censoring, we used inverse probability of censoring weights [94, 42] (see SM B.8 for details).

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

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

Risk Service data, we did not use current or diagnosis ages of non-counselee 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 re-calibrated 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 [17]. We calculated age-specific 5-year risks based on 2012-2016 incidence rates from the Surveillance, Epidemiology, and End Results (SEER) program [47]. We then modified the risk based on the number of affected first-degree relatives using relative risk estimates from [73] (the relative risks were 1.8 for 1 affected relative, 2.9 for 2 affected relatives, and 3.9 for 2 or more affected relatives). To re-calibrate each model, we used the Risk Service data to fit a linear regression with the family history-adjusted 5-year SEER risk as the outcome and the 5-year risk from the model as the predictor. We also evaluated a re-calibrated version of BRCAPRO obtained via the SEER re-calibration approach.

4.3. Results. Table 3 compares the performance of five models (FCNN, CNN, LR, BRCAPRO, and BRCAPRO\textsuperscript{C}, the SEER-recalibrated version of BRCAPRO) in the CGN dataset, which was 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 re-calibration. 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 A.3) 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 pre-specified (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 [68, 98, 33, 3, 20]) while the NNs were trained on a single dataset. The poorer performance of the NNs in clinic-based counselees may partly be explained by the fact that the NNs used less detailed family history information than BRCAPRO. Due to missing data in the training set, we did not include age information on non-counselees 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 [18, 8]. 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 [101, 92]. 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 \([51, 86, 102]\), 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 (for example, including only fam-
ilies with affected members). This can lead to ascertainment bias, i.e. 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 [25, 57, 59, 16, 49]. One approach for obtaining general population estimates from an ascertainment population is to weight families by the inverse probability of being ascertained [25]. This approach has similarities to weighting approaches that adjust for differences in covariate distributions between training and test datasets [86, 102] 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 re-calibrating 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 [37]. 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 [80, 81]. Methods have also been proposed for developing NN models that are intrinsically interpretable [32, 103, 62], 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 [87], the Swedish Family-Cancer Database, which includes over 2 million families [31], cancer studies based on the Utah Population Database, which includes over 1.3 million probands [89, 15], and a cancer study based on an Icelandic genealogical database with over 600,000 individuals [1]. 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 [52, 76, 69] and genetic testing companies [43], the development of a wide array of electronic patient-facing family history tools [97], 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 [65]. Also, electronic genealogical databases are rapidly expanding and there are continuing efforts to link them with clinical data to generate pedigrees [83, 84, 1, 89]. 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.

Acknowledgements. Work supported by the Friends of Dana-Farber Fund and NSERC PGSD35023622017. The authors thank Matthew Ploenzke for helpful suggestions.

SUPPLEMENTARY MATERIAL

Supplementary Material. 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. The simulation code is available at github.com/zoeguan/nn_cancer_risk.

REFERENCES


