Risk evaluation to identify individuals who are at greater risk of cancer as a result of heritable pathogenic variants is a valuable component of individualized clinical management. Using principles of Mendelian genetics, Bayesian probability theory, and variant-specific knowledge, Mendelian models derive the probability of carrying a pathogenic variant and developing cancer in the future, based on family history. Existing Mendelian models are widely employed, but are generally limited to specific genes and syndromes. However, the upsurge of multi-gene panel germline testing has spurred the discovery of many new gene-cancer associations that are not presently accounted for in these models. We have developed PanelPRO, a flexible, efficient Mendelian risk prediction framework that can incorporate an arbitrary number of genes and cancers, overcoming the computational challenges that arise because of the increased model complexity. We implement an eleven-gene, eleven-cancer model, the largest Mendelian model created thus far, based on this framework. Using simulations and a clinical cohort with germline panel testing data, we evaluate model performance, validate the reverse-compatibility of our approach with existing Mendelian models, and illustrate its usage. Our implementation is freely available for research use in the PanelPRO R package.

1. Introduction. Certain genetic variants are known to greatly increase one’s risk for developing cancer (Foulkes, 2008). Estimating the probability of carrying cancer-causing variants is important for personalized clinical management (Win et al., 2013). Mendelian models use principles of Mendelian genetics, probability theory, and gene-specific knowledge to estimate the probability that an individual receiving counseling (the counselee) is a pathogenic variant carrier (Murphy and Mutalik, 1969). As input, they take a potentially detailed pedigree of family history, which contains information on the relationships between individuals in the counselee’s family, as well as their ages, sex, cancer diagnosis statuses, and ages at diagnosis. As training, they rely on external estimates of the allele frequencies of the cancer-causing variants by gene, as well as the sex-specific and age-dependent probabilities of developing cancer conditional on presence of a pathogenic variant (cancer penetrances). These are population-level quantities and can be estimated based on published literature. The model outputs the counselee’s carrier probabilities conditional on family history and the future risk of developing cancer.

Mendelian models have been widely adopted (Euhus, 2001; Riley et al., 2012), but existing models are generally limited to a small number of genes and cancers within an individual syndrome. For example, Mendelian models implemented in the free, open-source BayesMendel R package (Chen et al., 2004) include BRCAPRO, which computes carrier probabilities for

**Keywords and phrases:** Mendelian models, risk prediction, germline panel gene testing, precision prevention, pathogenic variants.
pathogenic variants on BRCA1 and BRCA2 using family history of breast and ovarian cancers (Berry et al., 1997; Parmigiani, Berry and Aguilar, 1998); MMRpro, which considers Lynch syndrome genes MLH1, MSH2, and MSH6 using family history of colorectal and endometrial cancers (Chen et al., 2006); Pancpro, which considers an unspecified susceptibility gene for pancreatic cancer (Wang et al., 2007); and Melapro, which computes the carrier probability for a variant on CDKN2A using family history of melanoma (Wang et al., 2010). These models have been validated in the literature (Berry et al., 2002; Chen et al., 2006; Wang et al., 2007, 2010). Other Mendelian risk prediction models for breast and ovarian cancer susceptibility include Claus (Claus, Risch and Thompson, 1994), IBIS (Tyrer-Cuzick) (Tyrer, Duffy and Cuzick, 2004), and BOADICEA (Antoniou et al., 2008). In the context of future risk, non-Mendelian models are also prevalent; many summarize family history into binary or categorical predictors, followed by fitting logistic regression or another statistical learning model (Couch et al., 1997; Frank et al., 2002; Gail et al., 1989; Shattuck-Eidens et al., 1997; Vahteristo et al., 2001).

In recent years, many deleterious gene variants have been linked with the increased risk of various cancers. For example, ATM, CHEK2, and PALB2 (Antoniou et al., 2014; Marabelli, Cheng and Parmigiani, 2016; Schmidt et al., 2016) are among those associated with higher risk of breast cancer, in addition to the previously established BRCA1 and BRCA2. Furthermore, there is rising evidence that syndromes once thought to be distinct are determined by variants that increase the risk of multiple cancers (Kastrinos et al., 2009; Kastrinos and Syn-gal, 2011; Hruban et al., 2010; Moran et al., 2012; Kote-Jarai et al., 2011). These discoveries point to the need for flexible, scalable models that account for these relationships, many of which may be known but not yet accounted for by existing models. Simultaneously, the decreasing cost of DNA sequencing (Plichta et al., 2016) positions us to leverage data from multi-gene panel germline testing. As more associations are discovered using panel studies, these associations will also need to be effectively incorporated into models for quantitative risk assessment.

The development of a multi-syndrome model necessitates extending the Mendelian model framework from syndrome-specific to a fully generalizable framework that can account for associations between any number of genes and cancers. The increased model complexity implied by the inclusion of many genes and cancers also requires overcoming significant computational challenges. In particular, the Elston-Stewart peeling algorithm (Elston and Stewart, 1971) used to compute the carrier probabilities involves summing over all possible genotype configurations of a pedigree, a burden that is exponential with respect to the number of genes in the model. This computational cost must be appropriately addressed in order for larger, complicated models to return results in a reasonable runtime.

We present PanelPRO, a flexible, computationally efficient Mendelian risk model that incorporates an arbitrary number of genes and cancers. Generalizing beyond syndrome-specific approaches, we model carrier probabilities and cancer future risk probabilities for multi-syndrome associations for a large number of genes. For example, BRCAPO identifies individuals at high risk for breast or ovarian cancers due to pathogenic BRCA1 and BRCA2 variants. PanelPRO can also incorporate penetrances of breast cancer specific to pathogenic variants of ATM, CHEK2, and PALB2, as well as return carrier probabilities for these genes. We note that the BOADICEA model (Lee et al., 2019) has recently been extended to account for information from all five of these genes, as well as the effects of polygenic risk scores and other risk factors. Additionally, BRCA1 and BRCA2 have been linked to other syndromes, including pancreatic cancer (Greer and Whitcomb, 2007). Therefore, it may be beneficial to incorporate the penetrance of pancreatic cancer for BRCA1 and BRCA2 pathogenic variant carriers, allowing us to borrow strength across additional cancers and heighten our understanding of cross-syndrome effects. The framework introduced is highly flexible and can be used to obtain the future risk of cancer for all cancers in the model.
PanelPRO generalizes beyond syndrome-specific approaches while resolving computational challenges that arise from including many associations. We provide comprehensive notation for our statistical model, including the flexible incorporation of additional risk modifiers, and derive conditions for model collapsibility. By making some reasonable simplifying assumptions (Madsen et al., 2018), we reduce the computational expense of the peeling algorithm to a feasible level for clinical use. Using simulation studies, we rigorously verify the back-compatibility of PanelPRO against existing models.

Although the proposed model is general, allowing users to enter an arbitrary number of genes and cancers, for this study, we illustrate the model using a proposed PanelPRO-5BC, a five-gene (ATM, BRCA1, BRCA2, CHEK2, and PALB2) breast and ovarian cancer model; and a proposed PanelPRO-11, an eleven-gene (ATM, BRCA1, BRCA2, CDKN2A, CHEK2, EPCAM, MSH1, MSH2, MLH6, PALB2, and PMS2), eleven-cancer (brain, breast, colorectal, endometrial, gastric, kidney, melanoma, ovarian, pancreatic, prostate, and small intestine) model. These PanelPRO models are applied to simulated data to illustrate the properties and usage of our approach. For validation, we apply the PanelPRO-5BC and PanelPRO-11 models to a high-risk multi-gene panel testing dataset from the USC-Stanford Hereditary Cancer Panel (HCP) Testing Study (Idos et al., 2019). The PanelPRO R package is freely available for research purposes at https://projects.iq.harvard.edu/bayesmendel/panelpro (Lee et al., 2021).

In Section 2, we detail the statistical framework for PanelPRO. In Section 3, we describe the data used for specifying the input parameters and for validating our models. The results from the simulation studies and data application are reported in Sections 4 and 5, respectively. We conclude with a discussion in Section 6.

2. Statistical Methods.

2.1. Proposed Model. Suppose that a given counselee has a family of size I, and let \( \mathbf{G}_i = (G_{i1}, \ldots, G_{iK_i}) \) be the genotype vector for the \( i \)th family member, \( i = 1, \ldots, I \), where the counselee is indexed by 1. Each element \( G_{ki} \) stores individual \( i \)’s carrier status for a pathogenic variant in the \( k \)th gene. Let \( \mathcal{G} \) denote the set of all \( 2^K \) possible genotypes. In the simple case where only wild-type and heterozygous carriers need to be considered, \( G_{ki} \) is binary. Then, in a two-gene model, \( \mathbf{G}_i \) has length \( K = 2 \) and there are \( 2^2 = 4 \) possible genotypes (noncarrier for both genes, carrier of only the first pathogenic variant, carrier of only the second pathogenic variant, and carrier of pathogenic variants for both genes). Generally, more than two states can be assumed: for instance, we can differentiate between carriers of variants on one allele vs. both alleles. Then there are three possible states encoded as 0, 1, or 2, and \( 3^K \) possible genotypes. One can also assume the presence of additional states implied by having multiple variant types for a given gene.

Each individual in the family also has a vector of phenotypes \( \mathbf{H}_i = (H_{1i}, \ldots, H_{Ri}) \), where \( R \) is the arbitrary but fixed number of cancers or other phenotypes in the model. For individual \( i \), let \( T_{ri} \) be the age of diagnosis for cancer \( r \), \( C_i \) be their censoring age (current age or age of death), and \( \delta_{ri} = \mathbb{1}(T_{ri} \leq C_i) \) be the binary indicator for the occurrence of cancer \( r \) before the censoring age. Then the observed history of cancer \( r \) is \( \mathbf{H}_{ri} = (C_i, \delta_{ri}, T_{ri}) \) when \( \delta_{ri} = 0 \) and \( \mathbf{H}_{ri} = (C_i, \delta_{ri}, T_{ri}, \delta_{ri}) \) when \( \delta_{ri} = 1 \).

Denote the binary indicator of the \( i \)th individual being male as \( U_i \). Finally, let \( \mathbf{H} = (\mathbf{H}_1, \ldots, \mathbf{H}_I) \) and \( \mathbf{U} = (U_1, \ldots, U_I) \) indicate the phenotypes and sexes of all family members, respectively.

Of interest is the estimation of the counselee’s conditional probabilities \( \Pr(\mathbf{G}_1 = g_1 | \mathbf{H}, \mathbf{U}) \) for all genotypes \( g_1 \in \mathcal{G} \). To obtain the posterior distribution, we apply Bayes’ rule as presented by Murphy and Mutalik (1969) and described by Lange (2003):

\[
\Pr(\mathbf{G}_1 = g_1 | \mathbf{H}, \mathbf{U}) \propto \Pr(\mathbf{G}_1 = g_1) \Pr(\mathbf{H} | \mathbf{G}_1 = g_1, \mathbf{U}).
\]
After computing the right-hand side, the posterior probabilities must be normalized so that they sum to 1. \( \mathbb{P}(G_1 = g_1) \) is the population-based frequency of the counselee’s observed genotype \( g_1 \), which we generally estimate from published literature. \( \mathbb{P}(H|G_1 = g_1, U) \) is the conditional probability of the observed cancers for the entire pedigree given that the counselee has genotype \( g_1 \). It can be computed by integrating over the entire set of possible family genotypes, using the law of total probability:

\[
\mathbb{P}(H|G_1 = g_1, U) = \sum_{g_2, \ldots, g_I \in \mathcal{G}} \mathbb{P}(H|G_1 = g_1, \ldots, G_I = g_I, U) \mathbb{P}(G_2 = g_2, \ldots, G_I = g_I|G_1 = g_1)
\]

(2) \[
= \sum_{g_2, \ldots, g_I \in \mathcal{G}} \left[ \prod_{i=1}^{I} \prod_{r=1}^{R} \mathbb{P}(H_{ri}|G_i = g_i, U_i) \right] \mathbb{P}(G_2 = g_2, \ldots, G_I = g_I|G_1 = g_1).
\]

In Eq. 2, the joint genotype distribution conditional on the counselee’s genotype \( \mathbb{P}(G_2, \ldots, G_I|G_1) \) is derived using properties of Mendelian inheritance. One must have estimates of the population allele frequency of each variant and know the exact relationship between the counselee and each relative in the pedigree. We assume genotype independence across the founding family members, followed by conditionally independent genotypes in subsequent generations (Berry et al., 1997). In order to model \( \mathbb{P}(H|G_1, \ldots, G_I, U) = \prod_{i=1}^{I} \prod_{r=1}^{R} \mathbb{P}(H_{ri}|G_i, U_i) \) in the integration step, we assumed that the relatives’ phenotypes are independent conditional on genotype and sex. \( \mathbb{P}(H_{ri}|G_i, U_i) \) is computed from the sex-specific penetrances.

Two types of penetrances can be used. The net penetrance is the genotype-specific probability of developing a phenotype by a given age, in the absence of death and other competing risks. The crude penetrance is the genotype-specific probability of developing a phenotype by a given age prior to dying or developing other phenotypes. We typically use estimates reported in peer-reviewed literature, which are most often net. In such cases, we convert the estimates to crude when needed. Our implementation uses net penetrances \( \mathbb{P}(T_{ri} = t|G_i, U_i) \) to calculate the carrier probabilities. Explicitly,

\[
\mathbb{P}(H_{ri}|G_i, U_i) = \begin{cases} 
1 - \sum_{s=1}^{C_i} \mathbb{P}(T_{ri} = s|G_i, U_i) & \text{if } \delta_{ri} = 0 \\
\mathbb{P}(T_{ri} = T_{ri}^{obs}|G_i, U_i) & \text{if } \delta_{ri} = 1,
\end{cases}
\]

(3) where \( T_{ri} \) is the random variable and \( T_{ri}^{obs} \) is the observed cancer age. We assume that the censoring process and deaths by causes unrelated to a given cancer are independent of the time to cancer diagnosis. We further assume that censoring and death by other causes are non-informative with respect to carrier status. An in-depth discussion of the trade-offs involved in these modeling choices is provided by Katki et al. (2008).

Once these carrier probabilities have been obtained, they can be used to calculate the future risk of developing any of the cancers in the model. The future risk that the counselee develops the cancer indexed by \( r \) can be computed using either the net or crude penetrances, and has a slightly different interpretation depending on which penetrance estimates are used.

The counselee’s \( t_0 \)-year net future risk for a given cancer indexed by \( r \) is

\[
\mathbb{P}(T_{r1} \leq C_1 + t_0|T_{r1} > C_1, H, U)
\]

\[
= \sum_{g_1} \mathbb{P}(G_1 = g_1|H, U) \mathbb{P}(T_{r1} \leq C_1 + t_0|T_{r1} > C_1, G_1 = g_1, U_1)
\]

(4) \[
= \sum_{g_1} \mathbb{P}(G_1 = g_1|H, U) \cdot \frac{\mathbb{P}(C_1 \leq T_{r1} \leq C_1 + t_0|G_1 = g_1, U_1)}{\mathbb{P}(T_{r1} > C_1|G_1 = g_1, U_1)}.
\]
The numerator and denominator of the fractional term in the final expression can be computed from the net genotype-specific penetrances. This $t_0$-year risk prediction can be interpreted as the probability of developing cancer $r$ within $t_0$ years, conditional on being disease-free at the counselee’s current age and assuming that the counselee survives all competing risks and death. For individual $i$, let $T_{d,ri}$ be their age of death from causes other than cancer $r$, $T_{r1}^* = \min(T_{r1}, T_{d,ri})$ be their age of first outcome (either cancer $r$ or death from causes other than cancer $r$), and $J_{ri} = \mathbb{1}(T_{r1}^* = T_{d,ri})$ be the binary indicator for developing the $r$th cancer. The counselee’s crude risk can be calculated as

$$\mathbb{P}(T_{r1}^* \leq C_1 + t_0, J_{r1} = 1 | C_1, H, U)$$

$$= \sum_{g_1} \mathbb{P}(G_1 = g_1 | H, U) \mathbb{P}(T_{r1}^* \leq C_1 + t_0, J_{r1} = 1 | T_{r1}^* > C_1, G_1 = g_1, U_1)$$

$$= \sum_{g_1} \mathbb{P}(G_1 = g_1 | H, U)$$

$$\cdot \frac{\mathbb{P}(C_1 \leq T_{r1}^* \leq C_1 + t_0, J_{r1} = 1 | G_1 = g_1, U_1) - \mathbb{P}(T_{r1}^* > C_1, J_{r1} = 1 | G_1 = g_1, U_1)}{\mathbb{P}(T_{r1}^* > C_1, J_{r1} = 1 | G_1 = g_1, U_1)}.$$ (5)

The fractional term in the final expression is computed directly from the crude penetrances for cancer $r$ and the probability distribution of dying by other causes. The distribution of death by other causes is an additional parameter that can be estimated based on peer-reviewed literature. PanelPRO outputs crude future risk estimates by default.

2.2. Computation. The integration step over the counselee genotypes in computing the probabilities in a Mendelian model is usually obtained by applying the Elston-Stewart peeling algorithm (Elston and Stewart, 1971). The peeling algorithm is broadly used and generalizable to any number of heritable sites specified in the model. It is linear with respect to the number of relatives $I$ in the counselee’s family, but sums over all possible genotypes in the pedigree. Therefore, computation time is exponential with respect to the number of genes $K$, making integration challenging for even a moderate number of genes. The peeling and paring algorithm (Madsen et al., 2018) restricts the set of possible genotypes $\mathcal{G}$ to the set of genotypes $\mathcal{G}_M$ that contain up to $M$ deleterious variants. Then, one applies local integration over the “pared” set of genotypes $\mathcal{G}_M$ to approximate $\mathbb{P}(H | G_1 = g_1, U)$ as

$$\sum_{g_2, \ldots, g_l \in \mathcal{G}_M} \mathbb{P}(H | G_1 = g_1, \ldots, G_l = g_l, U) \mathbb{P}(G_2 = g_2, \ldots, G_l = g_l | G_1 = g_1).$$ (6)

When $M \geq K$, the peeling-paring algorithm is equivalent to the peeling algorithm.

For example, if the model has $K = 11$ genes, each of which is fairly uncommon in the population, it would be very rare to observe an individual with all 11 variants, or even 9 or 10 variants. It is also possible that many genotypes with several such pathogenic variants may not be viable. Illustrating again with the scenario of two states for each gene, if $M = 2$ and $K = 11$, there are only $\sum_{m=0}^{M} \binom{K}{m} = 78$ admissible genotypes in $\mathcal{G}_M$ that need to be summed over. This is a substantial reduction compared to the $M^K = 2048$ in the original space $\mathcal{G}$.

There are trade-offs in selecting the value for $M$. Higher values of $M$ lead to more precise approximations of the posterior genotypic distribution, but as $M$ approaches $K$, the computational complexity approaches $2^K$ once again. A smaller value of $M$ leads to higher
efficiency, but this simplifying assumption has the cost of lower precision. These trade-offs are discussed in greater detail by Madsen et al. (2018). Our PanelPRO implementation allows users to set the value for \( M \), but in both our simulations and data application, we used \( M = 2 \). Even as the number of genes in PanelPRO increases to 5, 7, 11, and potentially more, the computational expense of the integration step is kept feasible by imposing this restriction. The Lander-Green algorithm (Lander and Green, 1987) is an alternative algorithm that is expected to be more efficient when the number of genes \( K \) exceeds the number of family members \( I \) in the pedigree. Further discussion of genetic linkage algorithms and their computational considerations can be found in Lee et al. (2021).

2.3. Additional Modifiers.

2.3.1. Secondary Cancers. Separate from the general approach of including primary cancers in a model, secondary cancers such as contralateral breast cancer modify risk estimation based on years since the primary cancer diagnosis. Suppose that primary breast cancer is indexed by \( r = b \) and contralateral breast cancer by \( r = c \). Then \( P(H_{ci}|G_i, U_i) \) can be expressed in terms of the contralateral breast cancer net penetrances \( P(T_{ci} = t|G_i, U_i) \) as

\[
\begin{align*}
P(H_{ci}|G_i, U_i) &= \begin{cases} 
1 - \frac{1}{\prod_{s=T_{bi}^{obs}}^{T_{ci}^{obs}} P(T_{ci} = s|G_i, U_i)} & \text{if } \delta_{bi} = 1, \delta_{ci} = 0 \\
\prod_{s=T_{ci}^{obs}}^{T_{bi}^{obs}} P(T_{ci} = t_{bi}^{obs} - T_{bi}^{obs} | G_i, U_i) & \text{if } \delta_{bi} = 1, \delta_{ci} = 1 \\
1 & \text{if } \delta_{bi} = 0.
\end{cases}
\end{align*}
\]

\( T_{ci} \) is a random variable, whereas \( T_{bi}^{obs} \) and \( T_{ci}^{obs} \) are the observed ages of diagnosis for breast and contralateral breast cancer, respectively.

2.3.2. Risk Modifiers. Our implementation supports modifying risk at the cancer penetrance level using risk or hazard ratios for interventions such as prophylactic mastectomies, oophorectomies, and hysterectomies. When calculating carrier probabilities, simply use the modified penetrances \( P^{mod}(T_{ri}|G_i, U_i) \) in place of the raw penetrances \( P(T_{ri}|G_i, U_i) \). Suppose family member \( i \) received a preventive intervention at age \( T_{int,i} \), and let \( RR_{r,int}(T_{int,i}|G_i, U_i) \) be the relative risk of developing cancer \( r \) for those that receive the intervention at age \( T_{int,i} \), computed based on probabilities that are conditional on genotype and sex. Their modified penetrance for cancer \( r \) is

\[
P^{mod}(T_{ri} = t|G_i, U_i) = P(T_{ri} = t|G_i, U_i) \\
\cdot [T_{ri} < T_{int,i}] + RR_{r,int}(T_{int,i}|G_i, U_i) [T_{ri} \geq T_{int,i}].
\]

Suppose that instead of a relative risk, the hazard ratio for the intervention at age \( T_{int,i} \), \( HR_{r,int}(T_{int,i}|G_i, U_i) \), is available. Define \( S_r(t|G_i, U_i) = 1 - \sum_{s=1}^{t} P(T_{ri} = s|G_i, U_i) \) as the survival function for cancer \( r \) (conditional on sex and genotype), and let \( \lambda_r(t|G_i, U_i) = P(T_{ri} = t|G_i, U_i) / S_r(t|G_i, U_i) \) be the corresponding hazard. Then the modified hazard, survival, and penetrance functions at age \( t \) for this intervention are:

\[
\lambda_r^{mod}(t|G_i, U_i) = \lambda_r(t|G_i, U_i) [T_{ri} < T_{int,i}] + HR_r(T_{int,i}) [T_{ri} \geq T_{int,i}]
\]

\[
S_r^{mod}(t|G_i, U_i) = \prod_{s=1}^{t} (1 - \lambda_r^{mod}(s|G_i, U_i))
\]

\[
P^{mod}(T_{ri} = t|G_i, U_i) = S_r^{mod}(t + 1|G_i, U_i) - S_r^{mod}(t|G_i, U_i).
\]
2.3.3. Germline and tumor biomarker testing. When germline or tumor marker testing results are available for some family members, this information can also be incorporated into the probability estimates. For each relative \( i \), one substitutes \( \mathbb{P}(H_i|G_i, U_i) = \prod_{r=1}^{R} \mathbb{P}(H_{ri}|G_i, U_i) \) with

\[
\mathbb{P}^\text{mod}(H_i|G_i, U_i) = \mathbb{P}(H_i|G_i, U_i) \cdot \mathbb{P}(V_i|G_i) \cdot \mathbb{P}(W_i|G_i),
\]

where \( V_i \) is the observed germline testing results and \( W_i \) is the observed tumor marker testing results for relative \( i \). The probabilities \( \mathbb{P}(W_i|G_i) \) can be interpreted as test sensitivities and specificities. In practice, we often assume conditional independence of the different germline and marker testing results when robust joint estimates are not available. Currently, our implementation allows for modifying risk based on germline testing for any of the genes included in the model. It also supports modifying the risk of breast cancer conditional on BRCA1 or BRCA2 carrier status based on breast cancer tumor markers, ER, CK5/6, CK14, HER, and PR; and the risk of colorectal cancer conditional on MLH1, MSH2, MSH6, or PMS2 carrier status based on the colorectal cancer tumor marker, MSI. The implementation is designed to make it straightforward to extend risk modification to additional tumor markers or to modify the values according to user inputs.

2.4. Conditions for Collapsibility. It is possible to collapse a larger “full” model into a smaller submodel that contains a subset of the specified genes and/or cancers. In this section we derive the following important facts:

1. If the submodel contains a subset of the genes in the full model, both models will return the same posterior probabilities if the allele frequency for being a noncarrier is 1 for all genes not present in the submodel. In other words, the probability of carrying a pathogenic variant is zero for any of the genes that are absent in the submodel.
2. If the submodel contains a subset of the cancers in the full model, the submodel can be collapsed within the full model if the penetrances of the cancers not included in the submodel do not change depending on genotype.

Suppose that one has fit a model with \( K \) genes and wishes to collapse it into a submodel that only includes the genes with indices in the set \( \mathcal{K}^(*) \). Let \( \mathcal{K}^{(-*)} \) be the index set for genes not included in the submodel, such that \( \mathcal{K}^(*) \cup \mathcal{K}^{(-*)} \) is the full set of \( K \) gene indices. Similarly, \( G_i \) can be partitioned as as \( G_i = (G_i^(*), G_i^{(-*)}) \), where \( G_i^(*) \) is the genotype vector for the genes included in the submodel and \( G_i^{(-*)} \) is the genotype vector for the genes not included in the submodel. For the \( i \)th individual and \( k \)th gene, let \( G_{ki} = 0 \) represent a noncarrier and allow carriers of pathogenic variants to be represented using numbers other than zero. In order to obtain the same results from both the full model and submodel, a sufficient condition for collapsibility is that \( \mathbb{P} \left( G_{ki}^(*) = 0 \right) = 1 \) for all genes \( k \in \mathcal{K}^{(-*)} \) and all individuals \( i \) in the family.

Under this condition, the counselee’s genotype frequency is

\[
\mathbb{P}(G_1 = g_1) = \prod_{i=1}^{K} \mathbb{P}(G_{ki} = g_{ki}) = \begin{cases} 0 & \text{if } g_{k1} \neq 0 \text{ for at least one } k \in \mathcal{K}^{(-*)} \\ \mathbb{P}(G_{k1}^(*) = g_{k1}^(*)) & \text{if } g_{k1} = 0 \text{ for all } k \in \mathcal{K}^{(-*)}. \end{cases}
\]

Under the Mendelian model’s conditional independence assumptions for the genotypes in the pedigree, we can also obtain

\[
\mathbb{P}(G_2 = g_2, \ldots, G_l = g_l | G_1 = g_1) = \mathbb{P} \left( G_{2}^{(*)} = g_{2}^{(*)}, \ldots, G_{l}^{(*)} = g_{l}^{(*)} | G_{1}^{(*)} = g_{1}^{(*)} \right)
\]
when \( g_{ki} = 0 \) for all \( k \in K^{(-s)} \) and \( i = 1, \ldots, I \); and \( \mathbb{P}(G_2 = g_2, \ldots, G_I = g_I | G_1 = g_1) = 0 \) otherwise.

Let \( K_{i,1} \) be the set of gene indices that satisfy \( g_{ki} = 1 \) for a given individual \( i \), and let \( K_{i,1}^{(s)} \) be the analogous index set for only the genes that are included in the submodel. In the case where \( g_{ki} = 0 \) for all \( k \in K^{(-s)} \), it follows that \( K_{i,1} = K_{i,1}^{(s)} \) and

\[
\mathbb{P}(H_{ri} | G_i = g_i, U_i) = \prod_{k \in K_{i,1}} \mathbb{P}(H_{ri} | G_{ki} = g_{ki}, U_i) \\
= \prod_{k \in K_{i,1}^{(s)}} \mathbb{P}(H_{ri} | G_{ki} = g_{ki}, U_i) \\
= \mathbb{P}(H_{ri} | G_i^{(s)} = g_i^{(s)}, U_i) .
\]

Using these findings to compute Eq. 2, the expression for the posterior probabilities (Eq. 1) can be shown to be equivalent for the full model with \( K \) genes and the submodel with \( |K^{(s)}| \) genes. Note that the equality below holds because the terms in the summation where \( g_{ki} \neq 0 \) for at least one \( k \in K^{(-s)} \) are zero, and therefore do not contribute.

\[
\mathbb{P}(G_1 = g_1 | H, U) \\
\propto \mathbb{P}(G_1 = g_1) \sum_{g_2, \ldots, g_I \in G} \left[ \prod_{i=1}^I \prod_{r=1}^R \mathbb{P}(H_{ri} | G_i = g_i, U_i) \right] \\
\times \mathbb{P}(G_2 = g_2, \ldots, G_I = g_I | G_1 = g_1) \\
= \mathbb{P}(G_1^{(s)} = g_1^{(s)}) \sum_{g_2^{(s)}, \ldots, g_I^{(s)} \in G^{(s)}} \left[ \prod_{i=1}^I \prod_{r=1}^R \mathbb{P}(H_{ri} | G_i^{(s)} = g_i^{(s)}, U_i) \right] \\
\times \mathbb{P}(G_2^{(s)} = g_2^{(s)}, \ldots, G_I^{(s)} = g_I^{(s)} | G_1^{(s)} = g_1^{(s)}) \\
\propto \mathbb{P}(G_1^{(s)} = g_1^{(s)} | H, U) .
\]

Now suppose that the full model has \( R \) cancers or other phenotypes and that the submodel of interest contains only the phenotypes with indices in the set \( R^{(s)} \). Denote the complementary set as \( R^{(-s)} \). Let \( H_i^{(s)} \) be the phenotype vector for individual \( i \) corresponding to the phenotypes included in the submodel. In order to collapse the submodel within the full model, we require \( \mathbb{P}(H_{ri} | G_i = g_i, U_i) \) to be constant for all \( g_i \in G \) with respect to each individual \( i \) and cancer \( r \in R^{(-s)} \). In words, each relative’s penetrances for the cancers/phenotypes not included in the submodel must not depend on genotype. When this value is constant across genotypes, the additional cancers do not contribute to the posterior probabilities (Eq. 1) after normalization:

\[
\mathbb{P}(G_1 = g_1 | H, U) \\
\propto \mathbb{P}(G_1 = g_1) \sum_{g_2, \ldots, g_I \in G} \left[ \prod_{i=1}^I \prod_{r=1}^R \mathbb{P}(H_{ri} | G_i = g_i, U_i) \right] \\
\times \mathbb{P}(G_2 = g_2, \ldots, G_I = g_I | G_1 = g_1) \\
= \mathbb{P}(G_1^{(s)} = g_1^{(s)}) \sum_{g_2^{(s)}, \ldots, g_I^{(s)} \in G^{(s)}} \left[ \prod_{i=1}^I \prod_{r \in R^{(-s)}} \mathbb{P}(H_{ri} | G_i^{(s)} = g_i^{(s)}, U_i) \prod_{r \in R^{(s)}} \mathbb{P}(H_{ri} | G_i = g_i, U_i) \right] \\
\propto \mathbb{P}(G_1^{(s)} = g_1^{(s)} | H, U) .
\]
When the peeling-paring algorithm is used to restrict the possible genotypes to those in $G_M$, one can simply substitute $G_M$ for $G$ in the derivations.

3. Data.

3.1. Model Parameters. In summary, the PanelPRO framework can be used to train a Mendelian model with any $K$ genes and $R$ cancers. The training involves meta-analysis or literature reviews to identify allele frequency inputs for each of the genes and age-dependent penetrance inputs for each of the gene-cancer combinations. PanelPRO converts these into clinically relevant carrier probabilities and absolute risk evaluations.

To illustrate, we present three different examples of PanelPRO models that collectively span 11 genes, plus a hypothetical PANC gene, and 11 cancers. PANC is an unspecified pancreatic cancer susceptibility gene used in the PancPRO (Wang et al., 2007) Mendelian model for identifying individuals at high risk for pancreatic cancer. The current PanelPRO package includes all the input estimates discussed here, as well as others, with the exception of PANC, which is only discussed here in the context of back-compatibility with PancPRO. Allele frequencies and penetrances used to construct the models are estimated based on existing literature. For BRCA1 and BRCA2, we used the non-Ashkenazi, Ashkenazi Jewish and Italian allele frequency estimates from BRCAPRO (Chen et al., 2004; Antoniou et al., 2002); for MLH1, MSH2, and MSH6, we used the allele frequency estimates from MMRpro (Chen et al., 2004, 2006); for the hypothetical PANC gene, we used the allele frequency estimate from Pancpro (Chen et al., 2004; Wang et al., 2007); and for CDKN2A, we used the allele frequency estimate from Melapro (Chen et al., 2004; Berwick et al., 2006). Allele frequency estimates for ATM, CHEK2, and PALB2 were taken from Lee et al. (2016). Finally, we estimated allele frequencies for EPCAM and PMS2 from a 25-gene panel study of 252,223 individuals (Rosenthal et al., 2017).

Racially-informed (All Races, American Indian and Alaska Native, Asian, Black, White, Hispanic, White Hispanic, and White Non-Hispanic) probabilities of developing cancer (unconditional on genotype) were taken from the DevCan database (Statistical Research and Applications Branch, National Cancer Institute, 2020); penetrances for non-carriers are estimated by PanelPRO based on these values as well as the carrier penetrances and allele frequencies. When available, we took cancer penetrances for carriers from data included in
the BayesMendel package: the BRCA1 and BRCA2 estimates for the probability of developing breast or ovarian cancer (Chen et al., 2020); the MLH1, MSH2, and MSH6 estimates for the probability of developing colorectal or endometrial cancer (Wang et al., 2020; Felton, Gilchrist and Andrew, 2007); the PANC estimates for the probability of developing pancreatic cancer (Wang et al., 2007; Klein et al., 2002); and the CDKN2A estimates for the probability of developing melanoma (Wang et al., 2010; Begg et al., 2005; Bishop et al., 2002). These gene × cancer combinations are colored orange in Figure 1. Combinations colored in blue were obtained based on a literature review and are also available in the All Syndromes Known to Man Evaluator (ASK2ME) (Braun et al., 2018) tool. Combinations in white are not included, as insufficient evidence of association has been accrued.

The sensitivities and specificities for tumor marker testing of ER, CK5/6, CK14, HER, and PR are from Lakhani et al. (2002, 2005); the sensitivity and specificity for MSI testing are from Chu, Chen and Louis (2007). While the probability distribution of death from other causes does not come into play for the analysis in this paper, the PanelPRO package takes estimates for these values from DevCan (Statistical Research and Applications Branch, National Cancer Institute, 2020). The hazard ratios used for incorporating prophylactic mastectomies and oophorectomies into risk assessment are based on Katki (2007).

3.2. USC-Stanford HCP Cohort. In Section 5, we validate our method using data from the USC-Stanford Hereditary Cancer Panel (HCP) Testing Study, a prospective, multi-center study on multiplex gene panel testing for cancer susceptibility (Idos et al., 2019). A diverse group of 2000 patients was recruited from three medical centers: 1) USC Norris Comprehensive Cancer Center; 2) Los Angeles County + USC Medical Center; 3) Stanford University Cancer Institute. Eligible patients met clinical guideline criteria for genetic testing or had a ≥2.5% probability of mutation carriage calculated by the following validated models or algorithms: BayesMendel (Chen et al., 2004), BOADICEA (Antoniou et al., 2008), IBIS (Tyrer-Cuzick) (Tyrer, Duffy and Cuzick, 2004), PREMM 1,2,6 (Euhus, 2001; Kastrinos et al., 2009; Kote-Jarai et al., 2011), National Comprehensive Cancer Network (NCCN) Guidelines, or a personal history of ≥10 cumulative lifetime tubular adenomas (Idos et al., 2019). Diagnostic yield and off-target mutation detection was evaluated for 25- or 28-gene multi-gene panels (Idos et al., 2019) and performed by Myriad Genetic Laboratories (Salt Lake City, UT).

We validate two models, PanelPRO-5BC and PanelPRO-11, using this data. We excluded families if: the counselee is a carrier of one or more variants only in genes not included in the model; the counselee has a variant of uncertain significance (VUS), but not a pathogenic variant for any of the genes in the model; or the models cannot be applied to the pedigree (typically due to “loops”/inter-marriages or the reporting of cancer affection ages that are greater than the individual’s censoring age). We thus consider 1612 families for validating PanelPRO-5BC (average family size of 34.15) and 1468 families for PanelPRO-11 (average family size of 33.78). Tables 1 and 2 summarize the number of pathogenic variant carriers and cancer cases among the counselees in the study for PanelPRO-5BC and PanelPRO-11, respectively.

Additionally, breast tumor marker and MSI testing results are reported for some of families. In the PanelPRO-5BC validation subset, there are 204 families with at least one relative tested for ER, 167 for PR, and 55 for HER (a total of 226 or 14% families reported any results). In the PanelPRO-11 validation subset, there are 185 families with at least one relative tested for ER, 153 for PR, 45 for HER, and 3 for MSI (a total of 208 or 14.2% families reported any results). Information on preventative interventions like mastectomies, oophorectomies, and hysterectomies was not available.

We used racial or ancestry-informed allele frequencies and penetrances when available to match those reported in the data. Missing ages of cancer diagnosis and censoring ages were imputed by PanelPRO (Lee et al., 2021).
Brain, Breast, Colorectal, Endometrial, Gastric, Kidney, Melanoma, Ovarian, Pancreatic, Prostate, Small Intestine

ATM (0.0019) BRCA1 (0.0005829) BRCA2 (0.000676) CDKN2A (0.00015) CHEK2 (0.0026) EPCAM (0.000011) MLH1 (0.00045) MSH2 (0.00050) MSH6 (0.00018) PALB2 (0.00057) PANC (0.0034) PMS2 (0.00084)

Fig 1: Associations between the 11 genes, plus a hypothetical PANC gene, (rows) and 11 cancers (columns) collectively incorporated in the PanelPRO models illustrated in this paper. PANC is only listed in order to discuss back-compatibility with PancPRO and is not a gene that can currently be specified in the PanelPRO package. Population-level allele frequencies are given in parentheses (the CHEK2 allele frequency is for the 1100delC only; the allele frequencies for all other genes are for any pathogenic variant). Orange cells correspond to penetrances that are available in the BayesMendel package (Chen et al., 2004), and blue cells correspond to penetrances based on a literature review that are also available in ASK2ME (Braun et al., 2018).

Table 1

Number of counselees in the HCP multi-gene panel testing study who are pathogenic variant carriers (Table 1a) and cancer cases (Table 1b) for genes and cancers modeled by PanelPRO-5BC. These counts are based on the 1612 families left after processing the data for validating PanelPRO-5BC.

(a) Counselees who are pathogenic variant carriers.

| Gene | ATM | BRCA1 | BRCA2 | CHEK2 | PALB2 | Any Carrier | Non-Carrier |
|------|-----|-------|-------|-------|-------|-------------|-------------|
|      | 23  | 41    | 35    | 22    | 15    | 113         | 1499        |

(b) Counselees who are cancer cases.

| Cancer Type | Breast | Ovarian | Other Cancer | Any Cancer | No Cancer |
|-------------|--------|---------|--------------|------------|----------|
|             | 485    | 106     | 693          | 1161       | 451      |

4. Simulation Studies.

4.1. Data Generation. For each simulation study, we generated 1,000,000 families with structures sampled from families in the HCP study. Each simulated family includes the coun-
Table 2

Number of counselees in the HCP multi-gene panel testing study who are pathogenic variant carriers (Table 2a) and cancer cases (Table 2b) for genes and cancers modeled by PanelPRO-11. These counts are based on the 1468 families left after processing the data for validating PanelPRO-11.

(a) Counselees who are pathogenic variant carriers.

| Gene(s) | ATM | BRCA1 | BRCA2 | CDKN2A | CHEK2 | EPCAM | MLH1 |
|---------|-----|-------|-------|--------|--------|-------|------|
|         | 23  | 41    | 35    | 5      | 27     | 1     | 9    |
| MSH2    | 11  | 9     | 14    | 15     | 150    | 1318  |

(b) Counselees who are cancer cases.

| Cancer(s) | Brain | Breast | Colorectal | Endometrial | Gastric | Kidney | Melanoma | Ovarian | Pancreatic | Prostate | Small Intestine | Other Cancer | Any Cancer | No Cancer |
|-----------|-------|--------|------------|-------------|---------|--------|----------|---------|------------|----------|-----------------|--------------|------------|----------|
|           | 9     | 444    | 218        | 50          | 36      | 15     | 92       | 93      | 36         | 26       | 1               | 254          | 1055       | 413      |

4.2. Performance Metrics. To evaluate the carrier probabilities from our models, we consider several performance metrics (Steyerberg et al., 2010). We use the area under the curve (AUC) as a measure for discrimination, the expected divided by the observed number of events (E/O) as a measure of calibration, and mean squared error (MSE) as a measure of overall accuracy. When computing these metrics, non-carriers are defined as individuals who are not variant carriers of any gene in the model. Percentile confidence intervals are obtained from 1000 bootstrap samples.

4.3. Back-Compatibility. Four Mendelian risk prediction models are currently implemented in the BayesMendel R package (Table 3), each incorporating information from between 1-3 genes and cancers. Altogether, these models span seven genes (including a hypothetical PANC gene) and six cancers.

Table 3

Genes and cancers incorporated in the four existing BayseMendel Mendelian risk prediction models.

| BayesMendel model | Gene(s)       | Cancer(s)           |
|-------------------|---------------|---------------------|
| BRCAPRO           | BRCA1, BRCA2  | Breast, Ovarian     |
| MMRpro            | MLH1, MSH2, MSH6 | Colorectal, Endometrial |
| Pancpro           | hypothetical PANC | Pancreatic          |
| Melapro           | CDKN2A        | Melanoma            |

To validate the reverse-compatibility of our approach, we fit these existing models using the PanelPRO and BayesMendel packages and evaluated their performance on 1,000,000
simulated families, using the same model parameters and settings. These families were simulated assuming only the presence of the gene-cancer associations incorporated in the four existing models (orange in Figure 1), and based on family structures sampled from the HCP cohort. Because PanelPRO and BRCAPRO from the BayesMendel package handle contralateral breast cancer using different approaches, we eliminated it from consideration when fitting the models, to obtain a more direct comparison. The resulting metrics and carrier probabilities are virtually identical between the two sets of models, as expected (Supplemental Figure S1 and Supplemental Table S1 (Liang et al., 2021b)).

4.4. PanelPRO-5BC. We then simulated and evaluated 1,000,000 families under PanelPRO-5BC, a model that obtains carrier probabilities for ATM, BRCA1, BRCA2, CHEK2, and PALB2 based on family history of breast and ovarian cancer. As a point of reference, we also used the PanelPRO package to evaluate the simulated families with the BRCAPRO submodel, which only considers BRCA1 and BRCA2. Figure 2 plots metrics with bootstrap confidence intervals for PanelPRO-5BC; these metrics are also reported in Supplemental Table S2 (Liang et al., 2021b).

The results for the PanelPRO-5BC model and its BRCAPRO submodel are highly similar. The metrics for PanelPRO-5BC and BRCAPRO were also similar to each other across bootstrap replicates, as described in Section S6 of the supplementary material (Liang et al., 2021b). Incorporating simulated tumor marker testing for ER, PR, and HER in the model led to slightly better discrimination and precision of BRCA1 and BRCA2, as well as differences in calibration.

![Fig 2: AUC, calibration, and MSE for the full PanelPRO-5BC model (black) and the BRCAPRO submodel (blue) evaluated on 1,000,000 families simulated based on PanelPRO-5BC, with (open points) and without (closed points) risk modifiers. “BRCAPRO genes” indicates any of the genes in BRCAPRO (BRCA1, BRCA2), and “Any” indicates any of the five genes in PanelPRO-5BC. 95% bootstrap percentile confidence intervals are also shown.](image)

4.5. PanelPRO-11. As another illustration of the flexibility of our approach, we considered PanelPRO-11, an 11-gene (ATM, BRCA1, BRCA2, CDKN2A, CHEK2, EPCAM, MLH1, MSH2, MSH6, PALB2, and PMS2), 11-cancer (brain, breast, colorectal, endometrial, gastric, kidney, melanoma, ovarian, pancreatic, prostate, and small intestine) model. The performance metrics resulting from evaluating PanelPRO-11 on 1,000,000 simulated
families are plotted in Figure 3 and are reported in detail in Supplemental Table S3 (Liang et al., 2021b). We also evaluated the three BayesMendel submodels (BRCAPRO, MMRpro, Melapro) that are nested within PanelPRO-11. Including biomarker testing results can improve the metrics, particularly discrimination, for related genes such as MLH1, MSH2, MSH6, and PMS2. The very large calibration confidence intervals for EPCAM can be attributed to its low allele frequency, and the therefore low number of cases among the 1,000,000 simulated counselees.

Throughout, the full PanelPRO-11 model’s metrics are closely aligned with those of its submodels, with slightly better discrimination in some cases. In nearly all of the bootstrap samples, PanelPRO-11 has a better AUC and MSE. PanelPRO-11’s improvement in terms of calibration is less consistent, but the differences in model calibration are typically small (less than 2%) within bootstrap samples. Furthermore, whether or not PanelPRO improves over the submodel seems to be largely dependent on how well calibrated both models are; when both models are well-calibrated, PanelPRO’s calibration will be better about half the time. See Section S6 of the supplementary material for additional details (Liang et al., 2021b). PanelPRO-11 is thus able to produce the results of the existing BRCAPRO, MMRpro, and Melapro models, while also computing probabilities for being a carrier of ATM, CHEK2, EPCAM, PALB2, and PMS2.

The AUC for predicting pathogenic variant carriers of “Any” gene is 0.69. This is lower than the discrimination for most of the individual genes, many of which have AUCs above 0.8 or even 0.9. Groupings of genes that have individually high AUCs, like the BRCAPRO genes (BRCA1 and BRCA2) or the MMRpro genes (MLH1, MSH2, and MSH6), tend to retain good discrimination. However, including poorly-discriminating genes, as expected, pulls down the group AUC. This trend is observable for both PanelPRO-11 and the PanelPRO-5BC AUC results in Figure 2(a).

Figure 4 plots the AUCs from this simulation study against the relative risks of developing each of the 11 cancers in PanelPRO-11 by age 70, for pathogenic variant carriers compared to noncarriers. Both male and female relative risks are shown; see Supplemental Figure S2 for plots that separate between sexes (Liang et al., 2021b). Genes with lower AUCs correspond to those that have low relative risks for each of the eleven cancers. It appears that good discrimination for a given gene requires the gene to have a high relative risk for at least one cancer in the model. This trend is not expected for allele frequency (Supplemental Figure S3 (Liang et al., 2021b)). Note that ATM and CHECK2 have both low penetrance and high prevalence, but their low AUC is likely attributable to the former. The AUCs, allele frequencies, and relative risks plotted in Figure 4 and Supplemental Figures S2-S3 are summarized in Supplemental Table S4 (Liang et al., 2021b).

5. Data Application.

5.1. Model Validation. For model validation, we applied PanelPRO-5BC and PanelPRO-11 to the data from the HCP study (Section 3.2). We also ran the BayesMendel models nested within these models (BRCAPRO for PanelPRO-5BC and BRCAPRO, MMRpro, and Melapro for PanelPRO-11). The code to reproduce the validation results is available in Liang et al. (2021a). Some of the families in the HCP study report tumor marker testing results, so we ran the models with and without incorporating this extra risk-modifying information. To evaluate the models, we used the performance metrics described in Section 4.2 and obtained percentile confidence intervals from 1000 bootstrap samples.

The metrics and confidence intervals for PanelPRO-5BC are reported in Figure 5 and Supplemental Table S5; those for PanelPRO-11 are reported in Figure 6 and Supplemental Table S6 (Liang et al., 2021b). Metrics use outcome labels based on carrier status of individual genes as well as groups of genes. Both types remain fairly consistent between the full
PanelPRO models and their submodels. The results for models run with risk modifiers are often similar to those without risk modifiers. The MMRpro submodel appears to have better discrimination than PanelPRO-11 for Lynch syndrome genes MLH1, MSH2, and MSH6. The HCP cohort contains a high proportion of ovarian cancer, as well as cases of brain, pancreatic, and small intestine cancers, all of which are also associated with MLH1, MSH2, and MSH6 (Figure 1). Only PanelPRO-11, which appears to be slightly over-predicting the carrier probabilities for these genes, incorporates these additional associations, suggesting that the ascertainment process of the cohort makes this a difficult discrimination problem.

We acknowledge that the small sample size of the data-processed cohort makes it difficult to detect meaningful differences. The wide bootstrap confidence intervals are likely also attributable to the small number of cases. Similarly, the observation that incorporating risk modifiers into the models makes a limited impact on the results may be due to the relatively low number of families with tumor marker information. Section S6 of the supplementary material (Liang et al., 2021b) contains discussion on how often PanelPRO-5BC and PanelPRO-11 improve over the BayesMendel submodels in the 1000 bootstrap replicates; the full model does not consistently have better metrics across bootstraps. These results are again likely due to limitations in the sample size and ascertainment process for the HCP data that cause it
Fig 4: Area under the curve (AUC) from evaluating PanelPRO-11 on 1,000,000 simulated families plotted against the relative risk of developing each cancer by age 70, for pathogenic variant carriers compared to noncarriers.

to deviate from the makeup of the general population, for which the model parameters were estimated.

Fig 5: AUC, calibration, and MSE for the full PanelPRO-5BC model (black) and the BR-CAPRO submodel (blue) evaluated on the HCP cohort, with (open points) and without (closed points) risk modifiers. “BRCAPRO genes” indicates any of the genes in BRCAPRO (BRCA1, BRCA2), and “Any” indicates any of the five genes in PanelPRO-5BC. 95% bootstrap percentile confidence intervals are also shown.

The AUC confidence intervals for some genes, especially low-penetrant genes, overlap with 0.5. However, we note that the lower bound for the “Any” AUC is consistently above 0.5
A key contribution of the multi-syndrome predictions enabled by PanelPRO is the ability to acknowledge syndrome overlap in counseling individuals at risk. To illustrate, Figure 7 shows the probabilities of carrying a pathogenic variant of any PanelPRO-11 gene against the probabilities for carrying a pathogenic variant of any BRCAPO gene (BRCA1 or BRCA2, left) and any MMRpro gene (MLH1, MSH2, or MSH6, right) in the HCP cohort. Of the 150 HCP counselees who are carriers of any pathogenic variant in PanelPRO-11, 61 had carrier probabilities above a 2.5% threshold for PanelPRO but not BRCAPO and 0 had carrier probabilities above 2.5% for BRCAPO but not PanelPRO. Similarly, 89 counselees had carrier probabilities above above 2.5% for PanelPRO but not MMRpro, and 0 had carrier probabilities above 2.5% for MMRpro but not PanelPRO. This demonstrates that the potential refinement in counseling practice by jointly considering syndromes whose phenotypes overlap can affect a large proportion of families.

5.2. Sensitivity to Family Size. To examine the performance that can be achieved by PanelPRO-5BC and PanelPRO-11 when very extensive family history is available, we simulated additional families with a large number of relatives. Each simulated counselee has four...
maternal aunts, maternal uncles, paternal aunts, paternal uncles, sisters, and brothers. The counselee and their siblings each have four daughters and four sons, for a total of 112 simulated relatives in each family. These are unrealistically large for most clinical applications, and serve here to provide a best case scenario.

Performance metrics and bootstrap confidence intervals obtained from evaluating PanelPRO-5BC and PanelPRO-11 on the HCP cohort, simulated families with structures sampled from the HCP cohort, and large simulated families are reported in Supplemental Figures S8-S9 and Supplemental Tables S11-S12 (Liang et al., 2021b). To make the simulation results more comparable to the results from the HCP data, we computed these metrics based on a subset of the simulated families that has the same number of families as the HCP cohort, with the same number of pathogenic variant carriers for each gene. While the metrics usually look best when PanelPRO can take advantage of a detailed family history, the results from the real data and a simulation with fewer relatives often do not lag too far behind. We also note that the results from the HCP cohort are typically better calibrated than those from either set of simulated families, albeit with wide confidence intervals. The process of subsetting the simulated families to have the same number of carriers as the HCP cohort does not match the ascertainment process for the cohort, which is likely the driving force behind this phenomenon.

**6. Discussion.** We have developed PanelPRO, a general, computationally efficient framework for Mendelian risk models that incorporates an arbitrary number of genes and cancers or syndromes (groups of cancers that frequently occur together). Our model can leverage more family history information than its syndrome-specific counterparts by including multiple cancer types. It also allows for the development of models that address syndromes caused
by potentially many hereditary genetic factors. Extending beyond syndrome-specific models, PanelPRO integrates multiple gene-cancer associations and is flexible enough to incorporate any number of additional associations as they may be discovered in the future. This more comprehensive approach for quantitative risk assessment can be used to make a better determination of which individuals should undergo genetic testing and targeted preventative interventions.

By limiting the maximum number of pathogenic variants to consider in an individual’s genotype, we keep the problem computationally feasible (Madsen et al., 2018) even as the number of genes in the model grows, trading off between speed and accuracy. Lee et al. (2021) describes the package in detail, provides practical examples for users, and contains further discussion on computational performance. For moderately-sized families, models with as many as 10 genes can be run within a few seconds, and models with as many as 20 genes can be run in under a minute, making this package suitable for clinical applications. As our simulation studies demonstrate, PanelPRO is fully reverse-compatible with existing models. These results suggests that the more general PanelPRO could easily be adopted by current users of the BayesMendel models. For simplicity and ease of use, some of the results from PanelPRO could be masked from viewers who are only interested in specific genes and cancers.

Validation of models jointly predicting the carrier status of a large number of genes remains a challenge. The rich multi-class nature of the labels, the relative rarity of pathogenic variants in certain genes, and the broad variety of gene panels that are emerging in clinical practice all contribute to the difficulty of assembling cohorts that can definitively establish performance. The HCP study considered here is among the best available, but nonetheless, the small sample size is not sufficient for addressing high resolution questions about the more rare genes. We do succeed in establishing the important conclusion that the five- and eleven-gene models can be implemented clinically without detriment of performance compared to current practice. While PanelPRO’s ability to discriminate pathogenic variant carriers of any gene in the model is reduced as one incorporates genes with low penetrances, doing so can still be beneficial for identifying additional at-risk individuals. As more multi-gene panel testing data becomes available, we hope to perform additional validation.

The PanelPRO package (Lee et al., 2021) includes some functionalities that were not illustrated in our analysis. Besides tumor markers, our models currently consider prophylactic surgeries and germline testing as risk modifiers, and additional modifiers can be flexibly incorporated. When the cancer penetrance information is available, PanelPRO can return carrier probabilities for genotypes with a given pathogenic variant on both alleles as well as one allele. A future package release will handle multiple gene variants for a given gene.

We made several modeling assumptions, such as assuming that genotypes are independent across first-generation family members, followed by conditionally independent genotypes for second-generation family members and so on for subsequent generations (Berry et al., 1997). More importantly, we also assumed that the phenotypes are independent conditional on the individual’s genotype. Censoring and deaths by causes unrelated to a given cancer are taken to be independent of the time to cancer diagnosis and non-informative with respect to carrier status. We assumed an absence of competing risks, but as the number of cancers included in the model increases, many of the competing risks will naturally be accounted for. PanelPRO is further limited by uncertainties in its inputs. The model’s ability to estimate carrier probabilities based on family history depends in part on the input pedigree, which may contain errors or missing information (Braun et al., 2014). The model also relies on robust estimation of the model parameters—namely, the allele frequencies and penetrances—based on existing literature.

Despite the numerous practical challenges remaining in this area, we hope that the availability of our comprehensive framework and associated open source software will contribute
to a more systematic consideration of the relations between inherited susceptibility and cancer phenotypes, support more systematic prevention efforts, decrease barriers to interdisciplinary management of families at high risk for cancer, and encourage the adoption of general-purpose early detection strategies (Fiala, Kulasingam and Diamandis, 2019; Cristiano et al., 2019) among relatives of carriers.

SUPPLEMENTARY MATERIAL

Supplement A: Supplemental figures and tables
(doi: 10.1214/00-AOASXXXXSUPP; .pdf). Additional figures, tables, and discussion.

Supplement B: Code to reproduce paper analysis
(doi: 10.1214/00-AOASXXXXSUPP; .zip). Directory with code to perform analysis and generate figures in paper, also available at https://github.com/janewliang/PanelRePROducible

REFERENCES

Antoniou, A., Pharoah, P., McMullan, G., Day, N., Stratton, M., Petö, J., Ponder, B. and Easton, D. (2002). A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. British journal of cancer 86 76.

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

Antoniou, A. C., Casadei, S., Heikkinen, T., Barrowdale, D., Pylkas, K., Roberts, J., Lee, A., Subramanian, D., De Leeneer, K., Fostira, F. et al. (2014). Breast-cancer risk in families with mutations in PALB2. New England Journal of Medicine 371 497–506.

Begg, C. B., Orlow, I., Hummer, A. J., Armstrong, B. K., Kricker, A., Marrett, L. D., Millikan, R. C., Gruber, S. B., Anton-Culver, H., Zanetti, R. et al. (2005). Lifetime risk of melanoma in CDKN2A mutation carriers in a population-based sample. Journal of the National Cancer Institute 97 1507–1515.

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. Journal of the National Cancer Institute 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., Hilsenbeck, S. G. et al. (2002). BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. Journal of Clinical Oncology 20 2701–2712.

Berwick, M., Orlow, I., Hummer, A. J., Armstrong, B. K., Kricker, A., Marrett, L. D., Millikan, R. C., Gruber, S. B., Anton-Culver, H., Zanetti, R. et al. (2006). The prevalence of CDKN2A germ-line mutations and relative risk for cutaneous malignant melanoma: an international population-based study. Cancer Epidemiology and Prevention Biomarkers 15 1520–1525.

Bishop, D. T., Demenais, F., Goldstein, A. M., Bergman, W., Bishop, J. N., Pailletter, B., B.-d., Chompret, A., Ghiorzo, P., Gruis, N., Hansson, J. et al. (2002). Geographical variation in the penetrance of CDKN2A mutations for melanoma. Journal of the National Cancer Institute 94 894–903.

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

Braun, D., Yang, J., Griffin, M., Parmigiani, G. and Hughes, K. S. (2018). A clinical decision support tool to predict cancer risk for commonly tested cancer-related germline mutations. Journal of genetic counseling 27 1187–1199.

Chen, S., Wang, W., Bromman, K. W., Katki, H. A. and Parmigiani, G. (2004). BayesMendel: an R environment for Mendelian risk prediction. Statistical applications in genetics and molecular biology 3 1–19.

Chen, S., Wang, W., Lee, S., Nafa, K., Lee, J., Romans, K., Watson, P., Gruber, S. B., Euhus, D., Kinzler, K. W. et al. (2006). Prediction of germline mutations and cancer risk in the Lynch syndrome. Jama 296 1479–1487.

Chen, J., Bae, E., Zhang, L., Hughes, K., Parmigiani, G., Braun, D. and Rebeck, T. R. (2020). Penetrance of Breast and Ovarian Cancer in Women Who Carry a BRCA1/2 Mutation and Do Not Use Risk-Reducing Salpingo-Oophorectomy: An Updated Meta-Analysis. JNCI cancer spectrum 4 pkaa029.
ARMIGIANI, A. J., CHERT, M. K., HATTUCK, P., E. AHTERISTO, E. A. and MURPHY, M. W. (2010). Assessing the performance of prediction models: a framework for applications. *British journal of cancer* 1111–1130.

BUCHOWSKI, M., N., RIDE, C., G. OLNICK, S. B., P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences* 84 2363–2367.

LIERS, A. C. (2016). Incorporating truncating variants in PALB2, CHEK2, and ATM into the BOADICEA probability of multiple inherited risk alleles from pedigree data. *Genetics in Medicine* 1708–1718.

LEE, A. J., CUNNINGHAM, A. P., TISCHKOWITZ, M., SIMARD, J., PHAROAHL, P. D., EASTON, D. F. and ANTONIOU, A. C. (2016). Incorporating truncating variants in PALB2, CHEK2, and ATM into the BOADICEA breast cancer risk model. *Genetics in Medicine* 18 1190.

LEE, A., MAVADDAT, N., WILCOX, A. N., CUNNINGHAM, A. P., CARVER, T., HARTLEY, S., DE VILLERS, C. B., IZQUIERDO, A., SIMARD, J., SCHMIDT, M. K. et al. (2019). BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genetics in Medicine* 21 1708–1718.

LEE, G., LIANG, J. W., ZHANG, Q., HUANG, T., CHOIRAT, C., PARMIGIANI, G. and BRAUN, D. (2021). Multi-syndrome, multi-gene risk modeling for individuals with a family history of cancer with the novel R package PanelPRO. *eLife*.

LIANG, J. W., IDOS, G. E. “PanelPRO: a general framework for multi-gene, multi-cancer Mendelian risk prediction models.”

LIANG, J. W., IDOS, G. E., HONG, C., GRUBER, S. B., PARMIGIANI, G. and BRAUN, D. (2021a). Code to reproduce analysis for “PanelPRO: a general framework for multi-gene, multi-cancer Mendelian risk prediction models.”

LIANG, J. W., IDOS, G. E., HONG, C., GRUBER, S. B., PARMIGIANI, G. and BRAUN, D. (2021b). Additional figures, tables, and discussion for “PanelPRO: a general framework for multi-gene, multi-cancer Mendelian risk prediction models.”

MADESEN, T., BRAUN, D., PENG, G., PARMIGIANI, G. and TRIPPA, L. (2018). Efficient computation of the joint probability of multiple inherited risk alleles from pedigree data. *Genetic Epidemiology* 42 528–538.

MARABELLI, M., CHENG, S.-C. and PARMIGIANI, G. (2016). Penetrance of ATM gene mutations in breast cancer: A meta-analysis of different measures of risk. *Genetic epidemiology* 40 425–431.

MORAN, A., O’HARA, C., KHAN, S., SHACK, L., WOODWARD, E., MAHER, E., LALLOO, F. and EVANS, D. (2012). Risk of cancer other than breast or ovarian in individuals with BRCA1 and BRCA2 mutations. *Familial cancer* 11 235–242.

MURPHY, E. A. and MUTALIK, G. S. (1969). The application of Bayesian methods in genetic counselling. *Human Heredity* 19 126–151.

PARMIGIANI, G., BERRY, D. A. and AGUILAR, O. (1998). Determining carrier probabilities for breast cancer–susceptibility genes BRCA1 and BRCA2. *The American Journal of Human Genetics* 62 145–158.

PLICHTA, J. K., GRIFFIN, M., THAKURIA, J. and HUGHES, K. S. (2016). What’s New in Genetic Testing for Cancer Susceptibility? *Oncology* 30.

STATISTICAL RESEARCH AND APPLICATIONS BRANCH, NATIONAL CANCER INSTITUTE (2020). DevCan: Probability of Developing or Dying of Cancer Software. Version 6.7.8.

RILEY, B. D., CULVER, J. O., SKRZYNIA, C., SENTER, L. A., PETERS, J. A., COSTALAS, J. W., CALLFDLEY, F., GRUMET, S. C., HUNT, K. S., NAGY, R. S. et al. (2012). Essential elements of genetic cancer risk assessment, counseling, and testing: updated recommendations of the National Society of Genetic Counselors. *Journal of genetic counseling* 21 151–161.

ROSENTHAL, E. T., BERNHISEL, R., BROWN, K., KIDD, J. and MANLEY, S. (2017). Clinical testing with a panel of 25 genes associated with increased cancer risk results in a significant increase in clinically significant findings across a broad range of cancer histories. *Cancer genetics* 218 58–68.

SCHMIDT, M. K., HOGERVORST, F., VAN HIERN, R., CORNELISSEN, S., BROEKS, A., ADANK, M. A., MEIJERS, H., WAISFISZ, Q., HOLLESTELLE, A., SCHUTTE, M. et al. (2016). Age-and tumor subtype–specific breast cancer risk estimates for CHEK2*1100delC carriers. *Journal of clinical oncology* 34 2750.

SHATTUCK-EIDENS, D., OLIPHANT, A., MCCLOURE, M., MCBRIDE, C., GUPTA, J., RUBANO, T., PRUSS, D., TAYTIGIAN, S. V., TENG, D. H.-F., ADEY, N. et al. (1997). BRCA1 sequence analysis in women at high risk for susceptibility mutations: risk factor analysis and implications for genetic testing. *Jama* 278 1242–1250.

STEYERBERG, E. W., VICKERS, A. J., COOK, N. R., GERDS, T., GONE, M., OBUCHOWSKI, N., PENCINA, M. J. and KATTAN, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)* 21 128.

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

VAHTERISTO, P., EEROLA, H., TAMMINEN, A., BLOMQVIST, C. and NEVANLINNA, H. (2001). A probability model for predicting BRCA1 and BRCA2 mutations in breast and breast-ovarian cancer families. *British journal of cancer* 84 704–708.
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. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* **25** 1417.

WANG, W., NIENDORF, K. B., PATEL, D., BLACKFORD, A., MARRONI, F., SOBER, A. J., PARMIGIANI, G. and TSAO, H. (2010). Estimating CDKN2A carrier probability and personalizing cancer risk assessments in hereditary melanoma using MelaPRO. *Cancer research* **70** 552–559.

WANG, C., WANG, Y., HUGHES, K. S., PARMIGIANI, G. and BRAUN, D. (2020). Penetrance of Colorectal Cancer Among Mismatch Repair Gene Mutation Carriers: A Meta-Analysis. *JNCI Cancer Spectrum.*

WIN, A. K., MACINNIS, R. J., DOWTY, J. G. and JENKINS, M. A. (2013). Criteria and prediction models for mismatch repair gene mutations: a review. *Journal of medical genetics* **50** 785–793.