European Journal of Human Genetics (2019) 27:1800–1807
https://doi.org/10.1038/s41431-019-0467-5

ARTICLE

Exploring the effect of ascertainment bias on genetic studies that use clinical pedigrees

John Michael O. Ranola[1] · Ginger J. Tsai[1] · Brian H. Shirts[1]

Received: 6 November 2018 / Revised: 9 June 2019 / Accepted: 25 June 2019 / Published online: 11 July 2019

Abstract

Recent studies have reported novel cancer risk associations with incidentally tested genes on cancer risk panels using clinically ascertained cohorts. Clinically ascertained pedigrees may have unknown ascertainment biases for both patients and relatives. We used a method to assess gene and variant risk and ascertainment bias based on comparing the number of observed disease instances in a pedigree given the sex and ages of individuals with those expected given established population incidence. We assessed the performance characteristics of the method by simulating families with varying genetic risk and proportion of individuals genotyped. We implemented this method using SEER cancer incidence data to assess clinical ascertainment bias in a set of 42 pedigrees with clinical testing ordered for either breast/ovarian cancer or colorectal/endometrial cancer at the University of Washington and negative sequencing results. In addition to expected biases consistent with the stated testing purpose, there were trends suggesting increased colorectal and endometrial cancer in pedigrees tested for breast cancer risk and trends suggesting increased breast cancer in families tested for colon cancer risk. There was no observed selection bias for prostate cancer in this set of families. This analysis illustrates that clinically ascertained data sets may have subtle biases. In the future, researchers seeking to explore risk associations with clinical data sets could assess potential ascertainment bias by comparing incidence of disease in families that test negative under given ordering criteria to expected population disease frequencies. Failure to assess for ascertainment bias increases the risk of false genetic associations.

Introduction

As clinical sequence information becomes more available, a more-detailed understanding of genetic risk conferred by clinically tested genes and individual variants observed will become more important and more feasible to obtain [1]. Several recent studies have found novel cancer associations using cohorts of clinically ascertained patients who have had cancer panel testing [2–8]. In addition, penetrance estimates have often been obtained from families selected for having a high disease burden [9–15]. Each of these situations has potential to lead to biased overestimates of specific disease risk. When dealing with bias inherent in proband ascertainment, one can simply omit probands who were selected for both genotype and phenotype. This leads to calculating risk from analysis of relatives, who are typically assumed to be ascertained with less bias. However, for research studies that ascertain families based on multiple affected individuals, this assumption is not true. Similarly, in clinically ascertained pedigrees with specific genotype findings there may be an unknown ascertainment bias for both patients and relatives with both disease-specific and non-specific phenotypes being overrepresented because ordering providers may justify diagnostic testing based on either expected or intriguing patterns of familial disease. A system to compare incidence in ascertained families to expected population incidence could help identify unexpected biases.

Similarly, analysis attempting to classify individual variants in genes with established effects may suffer from clinical ascertainment bias. Large clinical data sets may be ideal for defining hypomorphic or low-risk variants, as large

Supplementary information The online version of this article (https://doi.org/10.1038/s41431-019-0467-5) contains supplementary material, which is available to authorized users.

✉ John Michael O. Ranola
ranolaj@uw.edu

[1] Department of Laboratory Medicine, University of Washington, Seattle, WA 98109, USA

samples sizes are needed to prove low level risk. However, in the clinical variant classification scenario, all variants in a gene are typically assumed to either have a known "pathogenic" effect or have no effect [16]. Assumed "pathogenic" variant penetrance may be determined from variants initially used to map or identify the gene. When these family data sets are used to generate risk estimates for genes implicated in hereditary disease risk, all variants thought to have the same functional effect have typically been pooled to have a large enough sample to generate risk estimates with confidence intervals consistent with significant risk [17–19]. Comparing clinical risk with established risk estimated for variant classification may be problematic for several reasons if there are differences between clinical and research sampling criteria. Research work regarding variant classification may be biased based on family selection criteria used to discover risk genes and clinical samples may have different bias based on less-clearly defined clinical selection. Because of both of these biases, variants with low-to-moderate risk relative to the assumed risk profile may never be correctly classified regardless of the amount of data. On the other hand, benign variants may be incorrectly classified as "pathogenic" if ascertainment enriches for specific phenotypes relative to population risk as ascertained by other studies.

As a way to identify the existence and magnitude of clinical ascertainment bias, we used ascertained pedigrees to evaluate cancer risk using cumulative population incidence of disease, which for cancer is publicly available from the SEER database (seer.cancer.gov) [20]. Applying these standardized incidence ratios (SIR), or relative risk over SEER, to pedigrees can detect potential ascertainment bias by using a control data set of patients/participant pedigrees with negative panel results from the same cohort where case testing was performed. Although we focus on identifying ascertainment bias, in a cohort without ascertainment bias, the method might also be used to assess a priori variant affect without prior knowledge of penetrance by providing an estimate of the SIR. When used for variant classification, this method is conceptually similar in some ways to the family history weighting algorithm presented by Pruss and colleagues [21, 22]. A major difference is that the pedigree SIR approach does not rely on matching large numbers of case and control pedigrees, as hypothetical "SEER" control pedigrees are created using the same family structure as actual case pedigrees. This approach may be simpler than the family history weighting algorithm and will certainly be easier for others to implement, as we have made all our code publicly available.

To illustrate the need to assess ascertainment bias in clinical samples and to demonstrate the performance characteristics of the pedigree SIR method, we simulated families with varying risks and varying number of individuals genotyped, and use the method to classify them. We also assessed clinically ascertained families from hereditary cancer risk testing at the University of Washington and show clinical ascertainment is likely to have unexpected bias that may explain published associations from genetic risk analysis in clinical data sets [2–5].

Methods

Overview

Given a set of probands carrying a particular variant and their associated pedigree, we can determine variant effect in relation to cancer by comparing the number of relevant cancers to the theoretical distribution of cancers assuming the pedigrees follow SEER incidence rates. Define X as the number of individuals with the desired phenotype, in this case a specific type of cancer. The probability that an individual has cancer assuming a benign variant would be a Bernoulli random variable with probability, p, equal to the SEER probability that the specific individual has cancer based on their sex and age. This single person distribution would have the following mean and variance.

$$\mu = p$$

$$\sigma^2 = p(1-p)$$

When more individuals are included into the analysis, and an independence assumption is added, this distribution changes into a Poisson binomial distribution with mean and variance listed below.

$$\mu = \sum_i p_i$$

$$\sigma^2 = \sum_i p_i(1-p_i)$$

Here p_i represents the probability that person i has cancer given their sex and age. By comparing the observed number of affected individuals in a test group (e.g., clinically ascertained families with negative results or specific variant carriers), X, to the Poisson binomial distribution with mean and variance assumed to follow SEER incidence, we can determine the probability that X came from that distribution. If it is different than expected, or has a small p-value, then the test group is unlikely to come from the same population as the population defined in the incidence curves (e.g., There is ascertainment bias or the variant of uncertain significance (VUS) has an effect on the phenotype). To perform this comparison, we rely on the R package "poibin", which generates the Poisson binomial distribution [23].

Alternatively, we can classify a variant by estimating its standard incidence ratios, or similarly relative risk over SEER. A standard incidence ratio greater than one would mean that the variant affects function. Standard incidence ratio is defined as the observed rate of a particular disease to the age adjusted expected rate in the general population. For the Poisson binomial model this becomes

$$\alpha = \frac{X}{\sum_i p_i}.$$

Here, α is our estimate of the SIR, X is the observed number of cancers, and p_i is the individual's probability of getting cancer based on SEER data. Furthermore, the confidence interval for α can be found by determining the values of α which do not yield significant probabilities, when comparing X to the Poisson binomial with αp in place of each success probability, for the given level of confidence. In other words, the confidence interval would be the values of α that the data does not reject at the given confidence level. Note that a variant which affects function according to the first method will have a SIR with confidence interval greater than 1 in the second as the two methods are related.

SEER incidence

SEER incidence values were taken from seer.cancer.gov [20]. As these were given in 5-year increments, we obtained intermediate, yearly, values using a linear interpolation of the adjacent points. These were done assuming the values given in the table were for the right end points of each interval with the right end point of "85+" set to 120. Bias estimates for different cancers and sites were calculated separately.

Probability of having cancer by sex and age

The probability, p, of having cancer is based on age, sex, and affection status. For an unaffected person, p is equal to the probability that the individual did not get cancer for their sex up to their current age or age at death. For an affected person, p is equal to the product of the probability that the individual did not get cancer for their sex up to the age they became affected. In other words, if k is the individual's current age, age at death, or age of affection (if affected and R_i is the gender specific yearly incidence rates from SEER for the individual, then $p = 1 - \prod_{i=1}^{k} (1 - R_i)$.

Weighting individuals by probability of having a genotype for variant classification

In the case of variant classification, in which many individuals will have unknown genotypes, it would be advantageous to count only individuals with the variant or with some chance of having the variant. This can be done by weighting individuals based on the probability they have the variant given individuals with known variant status and familial relationships. In this case we would first find all obligate carriers and then find the probability that the individuals with unknown genotypes have the variant and weight them accordingly. An R script to calculate these weights is included in the Supplementary File.

Evaluating potential bias

To determine if a data set may be biased we can use the method above to compare the number of cancers ascertained in "negative" families to the number theoretically expected in these families, assuming they follow the SEER distributions. Appropriate negative families are those that have the same enrollment criteria and genetic testing performed as case samples that will be used for a proposed analysis, but that have no genetic cause of cancer identified. Significant enrichment in the number of cancers in the data set of control families would indicate ascertainment bias.

Simulations

Using the CoSeg R package, we simulated pedigrees using published United States of America demographics. These demographics included age at marriage, age at death, and number of offspring surviving to adulthood for males and females during each decade since 1900. Our goal was to generate pedigrees that would be similar in size and shape to those of individuals receiving genetic testing for hereditary breast, ovarian, or colon cancer at the clinic. In brief, we began with a seeded age that is sampled from a skewed normal distribution derived from the age distribution of individuals receiving hereditary cancer testing at the University of Washington [24], extended up three generations to create a founder with the variant in question and then expanded the pedigree down with each descendant having 0.5 probability of inheriting the variant of interest from a parent with the variant. We used published population demographic measures for average marriage age, number of offspring living to adulthood, and mortality [25, 26]. Phenotypes were sampled based on age and genotype status using the published *MLH1* penetrance [27, 28] and various SIR's.

We simulated 1000 sets of three-generation families ranging in size from 1 to 15 with relative risk (RR) of 1, 5, 10, and 15 times the SEER incidence and calculated the probability we would see a greater number of cancers than seen in each set using the CoSeg R program [29, 30]. To give some perspective, we also compared these with families simulated with literature-reported *MLH1* penetrance [27, 28]. Next, we simulated 1000 sets of five families of varying RR and varied the proportion of relatives with known genotype from 100 to 0% (the proband always had known genotype, of course). We again calculated the probability we would see the number of cancers seen given that we were sampling from the SEER values and plotted them as box and whisker plots. To demonstrate the effect of ascertainment bias, we simulated sets of five families of varying RR and selected 1000 that had more than one affected individual. We again calculated the probability we would see the number of cancers given the SEER distribution and plotted the results.

Clinical sample

The BROCA panel is a pan-cancer risk panel that is optimized for individuals with hereditary breast and ovarian cancer. We evaluated 21 sequential families with BROCA panel and billing codes indicating ordering for breast and ovarian cancer evaluation. The ColoSeq panel is a similar panel designed for evaluating colorectal cancer risk. We evaluated an additional 21 sequential families had ColoSeq panel testing or pan-cancer panel testing where billing codes indicated ordering for colorectal cancer risk. For all families selected in both groups the results of clinical testing were negative (no cancer risk variants or VUS identified in any tested gene). These pedigrees were coded and deidentified for subsequent analysis. Without any pre-existing estimates of potential bias, we coded an arbitrary number, 21, in each group for our initial analysis. None of these families had identified hereditary cancer risk variants. Testing was referred from patients across the country, with a preponderance of orders coming from Washington State. The study was approved by the University of Washington IRB (#00005392).

For cancers not predicted to be part of clinically expected risk profiles, observed cancers should follow the SEER distribution, if there is no bias present. Using the pedigree structure and reported or estimated ages of individuals in these families, we ran 100,000 simulations to generate expected cancer count for breast, colorectal, endometrial, ovarian, pancreatic, and prostate cancer using published SEER age-specific cancer incidence [20]. We excluded probands from actual and expected cancer counts so that data would illustrate bias in relative selection, rather than bias in proband selection. We then created a histogram of the expected cancer count values and compared them to the actual number of cancers seen in the families.

Software availability and timing

All code used for this manuscript is publicly available as a supplementary R file to be used with the freely available R program. Each analysis takes less than a minute to run using 100,000 simulations to create an expected SEER distribution.

Results

Simulated sample performance characteristics: variant effect without ascertainment bias

To demonstrate the performance characteristics of the method in the absence of bias, we used 1000 sets of simulated 3-generation pedigrees with variants which affect function that conferred 1, 5, 10, and 15 times the SEER incidence of colorectal cancer. For these plots, we assumed a 30% genotyping rate apart from the proband and weighted individuals based on the probability they carry the variant. With a single family, it is unlikely that a correct classification will be made for any relative risk regardless of ascertainment bias. At seven families, we are able to correctly classify more than half of the relative risk 15 sets and the *MLH1* sets. At 13 families, we are able to correctly classify more than half of the relative risk 10 sets as well. Furthermore, only five families were incorrectly classified as having an effect on risk when using families simulated with relative risk 1, sets across all family sizes (8000 families total). Figure 1 shows a box and whisker plot of these results.

We further explored the performance characteristics of this method. To deal with incomplete genotyping, we experimented with weighting ungenotyped individuals

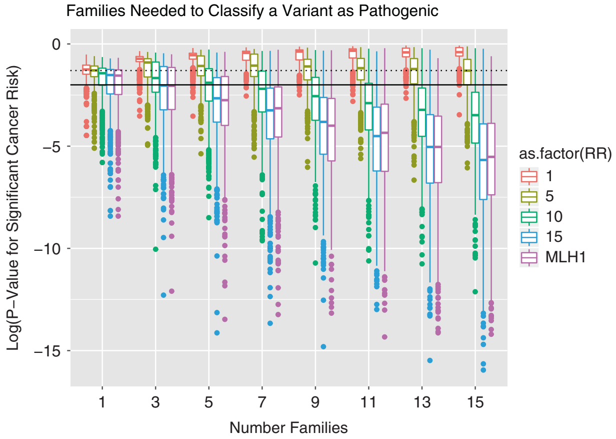

Fig. 1 Number of families required for classification of variants with varying relative risks including colon cancer caused by a variant which affects function in *MLH1*. A horizontal dashed line is drawn at $P = 0.05$ and a solid line at $P = 0.01$ for convenience. Here, each set of families is simulated 1000 times to obtain a distribution for the boxplot. The underlying simulation used colon cancer risk from the SEER population as baseline and assumes that 30% of individuals are genotyped

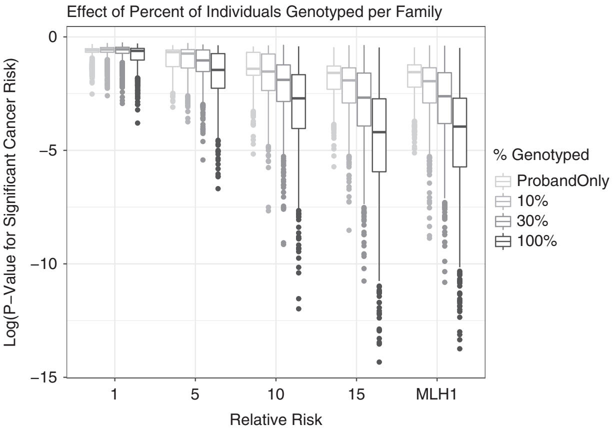

Fig. 2 Effect of genotyping on determining the class of a variant for varying relative risks. Here, each set of 5 families is simulated 1000 times to obtain a distribution. The underlying simulation used colon cancer risk from the SEER population

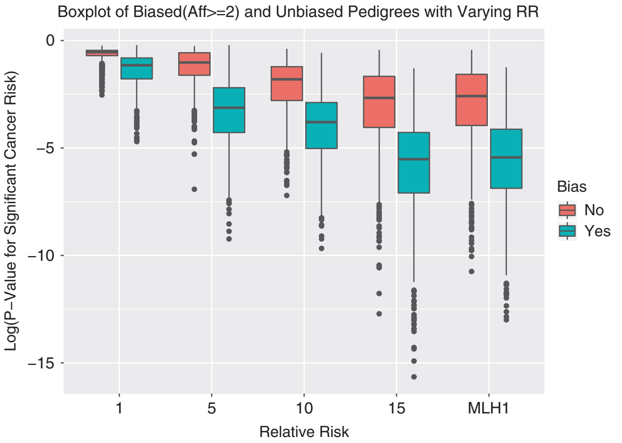

Fig. 3 This figure shows boxplots of the probability of correct "pathogenic" classification in unbiased and biased pedigrees, ones which were selected to have more than two affecteds with only one accounted for, for varying relative risks. Here, probabilities were obtained from a sample of five families and repeated 1000 times to obtain distributions. The underlying simulation used colon cancer risk from the SEER population as baseline and assumes that 30% of individuals are genotyped

based on the probability that they have the variant. Appendix Fig. 1 shows the results of unweighted, weighted, and inverse-weighted (gives weight based on the probability individuals do not have the variant) analysis for several relative risks including *MLH1*. The number of individuals genotyped in a pedigree can vary, depending on study design. Having fully genotyped pedigrees naturally provides the most correct classifications, though classification can still be made with proband-only genotypes, which is the scenario that is most common for clinically ascertained pedigrees. Increasing the number of genotyped individuals provides non-linear increases in information as family structure often allows imputation of genotype status of ungenotyped individuals. Figure 2 shows the boxplots obtained from varying the percent of genotyped individuals in the pedigrees.

Simulated sample performance characteristics: detecting ascertainment bias

To evaluate the power of this method to demonstrate bias, we simulated pedigrees across a range of relative risks and selected only those with at least two affected individuals for analysis. We analyzed the data set and show that for known benign data sets the pedigree SIR approach is able to detect bias. Figure 3 shows the result for benign, relative risk 1, variants along with other relative risks. Pedigrees with no risk and biased ascertainment appeared like pedigrees with relative risk of 5. For variants causing increased risk, including bias leads to a p-value for variant effect that would be expected with a higher actual risk. Pedigrees simulated with 5-fold risk and ascertainment bias appeared to have greater risk than pedigrees with 10-fold risk and no ascertainment bias (Fig. 3).

Detecting ascertainment bias in a clinical sample

To assess whether we could use clinically ascertained families to determine cancer risk associated with specific variants, we analyzed 42 pedigrees from the University of Washington. Half of these pedigrees (21) were sent in with clinical orders suggesting either breast/ovarian and the other half (21) with orders for colorectal/endometrial cancer risk sequencing. No patient or relative in the pedigrees was found to have variants that affect function or VUS in any gene tested. The probability that these pedigrees are unbiased for each cancer type is shown in Fig. 4.

The set of pedigrees where breast/ovarian cancer risk was interrogated was significantly biased ($P < 0.01$) for breast, ovarian, and pancreatic cancer, as expected. Unexpectedly, these families also had trends suggesting selection for colorectal and endometrial cancer in relatives. There was no observed bias for prostate cancer in this small set of families. Had these sets not been biased, estimates of the relative risk would be 4.25, 9.30, 3.91, 1.88, and 2.61 for breast, ovarian, pancreatic, colorectal, and endometrial cancer, respectively (Appendix Table 1).

The set of pedigrees with clinically ascertained for colorectal/endometrial cancer risk was significantly biased ($P < 0.01$) for colorectal and endometrial cancer, as expected. There was a trend for bias towards breast cancer with no apparent bias for ovarian and prostate cancer in pedigrees. Had these sets not been biased, estimates of the relative risk would be 2.47, 4.55, and 1.55 for colorectal, endometrial, and breast cancer, respectively (Appendix Table 1).

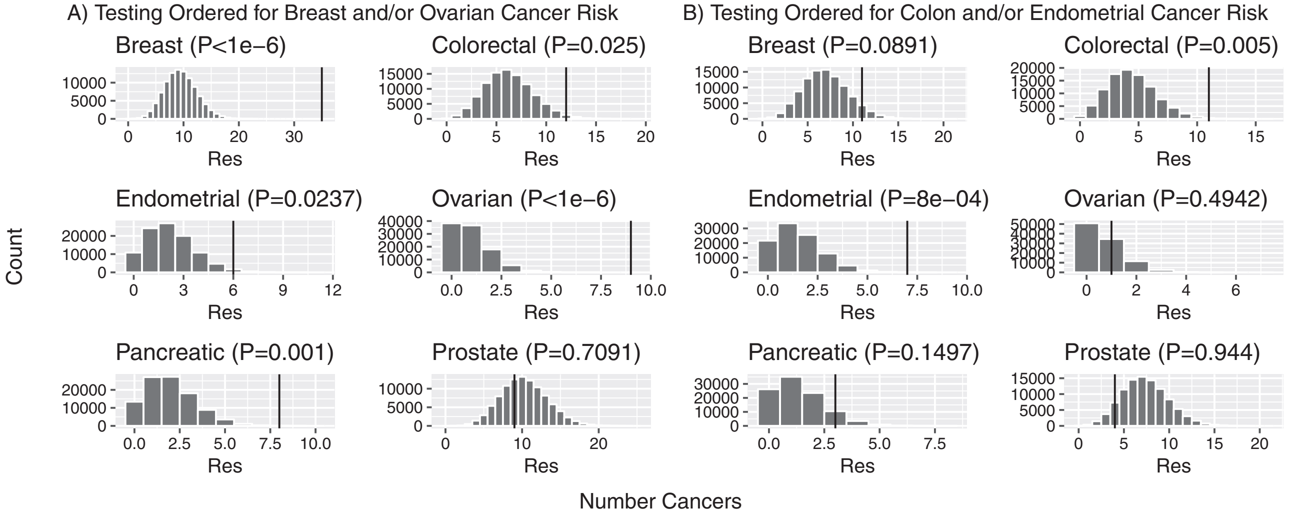

Fig. 4 Expected distributions of the number of affected individuals for various cancer types for 21 clinical families with testing ordered for breast or ovarian cancer risk **a**, and colorectal or endometrial cancer risk **b**. Also included are the number of observed cancer cases for each type represented by a vertical black line and the corresponding p-value. Note that these histograms of expected cancer cases in the group of families were made by sampling the distribution 100,000 times

Discussion

Clinical pedigrees are increasingly used for variant classification, genetic association, and penetrance estimates, but these may be biased owing to a higher likelihood of families enriched for expected and unexpected phenotypes being sent to the clinic [2–5]. We developed a relatively simple method to assess this potential bias. Even a small number of pedigrees showed trends suggesting unexpected bias patterns for pedigrees from clinical testing performed by the University of Washington Department of Laboratory Medicine. In addition to the bias expected based on stated rationale for test orders, these small groups of pedigrees showed trends towards enrichment for specific unrelated phenotypes, with colorectal and endometrial cancer potentially being overrepresented in families being tested for breast cancer risk and breast cancer potentially being overrepresented in families being tested for colorectal cancer risk. Neither group showed enrichment for prostate cancer, indicating that assessments of prostate cancer risk in patients found to have positive genetic findings during this period are less likely to be biased. However, ordering patterns may change in the future after as a result of manuscripts showing variants in *BRCA2,* which affects function in patients with metastatic prostate cancer [31]. The exact values for the amount of clinical ascertainment bias presented here should not be interpreted as definitive or generalizable. Bias in different samples may be different, so studies of genetic effect should conduct independent evaluations of sample-specific bias.

We evaluated clinically ascertained families seeking to evaluate novel associations of other clinical laboratories and potentially using clinical pedigrees for individual variant assessment. With only a small set of pedigrees we observed a high probability for false positives, which diminished our initial enthusiasm for using clinically ascertained pedigrees to assess novel associations. It was intriguing to us that the level of ascertainment bias observed in clinically ascertained samples sent to the University of Washington might explain recently reported associations of known colorectal cancer risk genes *MSH6* and *PMS2* with breast cancer [2, 4]. The observed bias is not surprising given national guidelines for medical professionals about appropriate hereditary cancer risk ordering [32]. The question of ascertainment bias depends on what population the inference is to be made. We assumed that risk relative to the general population is ideal. Risk estimates using clinical samples may produce correct comparisons with clinically ascertained "controls", even if they are not ideal estimates of risk relative to the general population. If risk relative to the general population is measured, larger studies that look for low levels of risk will need larger control samples to check for subtle sources of ascertainment bias.

In the absence of bias, using SEER data to estimate the expected number of cancers in a family might help classify variants of uncertain significance over a range of relative risk factors, known genotypes, and number of families without any prior estimates of penetrance. This method for variant assessment may have some distinct benefits over traditional methods. We developed this method to not rely on previous estimates of penetrance, so it would be capable of identifying variants that are less penetrant than those identified in published literature. In addition, it incorporates data from the entire pedigree. (Avoiding the proband, who is known to be ascertained for both phenotype and genotype, is already generally done.) The last feature, which has turned out to be critically important in pre-study data assessment, is that comparison of baseline pedigrees with SEER incidence can highlight ascertainment bias. This bias may be seen in research ascertainment for cohorts used to define pathogenicity as well as clinical cohorts used to classify VUS or identify new genotype–phenotype correlations.

This method has two major assumptions. The first assumption is that the underlying disease risk in the population studied is well-defined (e.g., SEER incidence). The second is that the individuals in the data set are independent. With different assumptions more complex segregation analysis using software, such as MENDEL, could be used similarly with variant "negative" families to assess ascertain bias. There are several limitations to this method. In populations with different levels of underlying risk, this method will generate skewed results. This could be corrected by using population-based incidence estimates instead of the general population estimates, if these are known. Although we have been able to show there is likely to be bias in one clinically ascertained data set, we have not shown how to account or correct for it. Being able to account for that bias might allow one to use and combine variant data from a variety of ascertainment strategies. Furthermore, the independence assumption is likely not completely correct as individuals within families are correlated with each other. However, as many families are needed for the analysis, the independence of families is likely to outweigh the dependence within them, especially with larger samples of families. Another minor limitation is that for variant classification, whereas we have shown that weighting counts by probability of having the variant gives higher power (See Appendix Fig. 1), this leads to comparing fractions to the discrete valued Poisson binomial distribution. We have taken a conservative approach to this issue by rounding down, as there are clear gains in power by incorporating genotype probability information from ungenotyped individuals. It is worth noting that this method is conceptually similar to the logistic regression approach by Easton et al. [33].

Although convenient, clinically ascertained data sets may have subtle biases. In the future, researchers seeking to explore risk associations with clinical data sets should perform more thorough evaluations of ascertainment bias in their samples. The process of evaluating families that test negative in parallel with those that test positive with the same ordering criteria is simple and could be an important standard for association studies that rely on clinical ascertainment. Comparisons of clinical cases with population controls from public databases that are ascertained using very different criteria will lead to many false positives without this or a similar check for ascertainment bias. Although we have shown how one may effectively ascertain clinical ascertainment bias relative to general population risk, we are not aware of a method to correct for false associations that may occur due to this bias. Appropriately correcting for biased ascertainment will be an important topic for future statistical genetics work.

Acknowledgements Primary funding for this project was provided by the University of Washington Department of Laboratory Medicine. In addition, the research group receives funding from the Damon Runyon Cancer Research Foundation (DRR-33-15) and the Fred Hutchinson/University of Washington Cancer Consortium (NCI 5P30 CA015704-39).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Shirts BH, Pritchard CC, Walsh T. Family-specific variants and the limits of human genetics. Trends Mol Med. 2016;22:925–34.
2. Espenschied CR, LaDuca H, Li S, McFarland R, Gau C-L, Hampel H. Multigene panel testing provides a new perspective on lynch syndrome. J Clin Oncol. 2017;35:2568–75.
3. Rana HQ, Gelman R, LaDuca H, McFarland R, Dalton E, Thompson J, et al. Differences in TP53 mutation carrier phenotypes emerge from panel-based testing. J Natl Cancer Inst. 2018;110:863–70.
4. Roberts ME, Jackson SA, Susswein LR, Zeinomar N, Ma X, Marshall ML, et al. MSH6 and PMS2 germ-line pathogenic variants implicated in Lynch syndrome are associated with breast cancer. Genet Med. 2018;20:1167–74.
5. Shimelis H, LaDuca H, Hu C, Hart SN, Na J, Thomas A, et al. Triple-negative breast cancer risk genes identified by multigene hereditary cancer panel testing. J Natl Cancer Inst. 2018;110:855–62.
6. Couch FJ, Shimelis H, Hu C, Hart SN, Polley EC, Na J, et al. Associations between cancer predisposition testing panel genes and breast cancer. JAMA Oncol. 2017;3:1190–6.
7. Lu H-M, Li S, Black MH, Lee S, Hoiness R, Wu S, et al. Association of breast and ovarian cancers with predisposition genes identified by large-scale sequencing. JAMA Oncol. 2018.
8. Sutcliffe EG, Bartenbaker Thompson A, Stettner AR, Marshall ML, Roberts ME, Susswein LR, et al. Multi-gene panel testing confirms phenotypic variability in MUTYH-associated polyposis. Fam Cancer 2019;18:203–9.
9. Chen S, Parmigiani G. Meta-analysis of BRCA1 and BRCA2 penetrance. J Clin Oncol 2007;25:1329–33.
10. Siegmund K, McKnight B. Modeling hazard functions in families. Genet Epidemiol 1998;15:147–71.
11. Gong G, Whittemore AS. Optimal designs for estimating penetrance of rare mutations of a disease-susceptibility gene. Genet Epidemiol 2003;24:173–80.
12. Wang Y, Ottman R, Rabinowitz D. A method for estimating penetrance from families sampled for linkage analysis. Biometrics 2006;62:1081–8.

13. Kraft P, Thomas DC. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. Am J Hum Genet. 2000;66:1119–31.
14. ten Broeke SW, Brohet RM, Tops CM, van der Klift HM, Velthuizen ME, Bernstein I, et al. Lynch syndrome caused by germline PMS2 mutations: delineating the cancer risk. J Clin Oncol. 2015;33:319–25.
15. Goodenberger ML, Thomas BC, Riegert-Johnson D, Boland CR, Plon SE, Clendenning M, et al. PMS2 monoallelic mutation carriers: the known unknown. Genet Med J Am Coll Med Genet. 2016;18:13–9.
16. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med J Am Coll Med Genet. 2015;17:405–24.
17. Pharoah PD, Guilford P, Caldas C. International Gastric Cancer Linkage Consortium. Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. Gastroenterology 2001;121:1348–53.
18. Lin KM, Shashidharan M, Thorson AG, Ternent CA, Blatchford GJ, Christensen MA, et al. Cumulative incidence of colorectal and extracolonic cancers in MLH1 and MSH2 mutation carriers of hereditary nonpolyposis colorectal cancer. J Gastrointest Surg J Soc Surg Aliment Tract. 1998;2:67–71.
19. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. Am J Hum Genet. 1998;62:676–89.
20. Noone A, Howlader M, Krapcho M, Miller D, Brest A, Yu M, et al. SEER Cancer Statistics Review, 1975–2015 [Internet]. Bethesda, MD: National Cancer Institute; 2018 Apr. Available from: https://seer.cancer.gov/csr/1975_2015/.
21. Pruss D, Morris B, Hughes E, Eggington JM, Esterling L, Robinson BS, et al. Development and validation of a new algorithm for the reclassification of genetic variants identified in the BRCA1 and BRCA2 genes. Breast Cancer Res Treat. 2014;147:119–32.
22. Morris B, Hughes E, Rosenthal E, Gutin A, Bowles KR. Classification of genetic variants in genes associated with Lynch syndrome using a clinical history weighting algorithm. BMC Genet. 2016;17:99.
23. Hong Y. On computing the distribution function for the Poisson binomial distribution. Comput Stat Data Anal. 2013;59:41–51.
24. Shirts BH, Casadei S, Jacobson AL, Lee MK, Gulsuner S, Bennett RL, et al. Improving performance of multigene panels for genomic analysis of cancer predisposition. Genet Med J Am Coll Med Genet. 2016;18:974–81.
25. McNicoll G World Population Prospects: The 1998 Revision. Volume I: Comprehensive Tables; Volume II: The Sex and Age Distribution of the World Population [Internet]. Population and Development Review. 1999 [cited 2019 Mar 25]. Available from: http://link.galegroup.com/apps/doc/A63296795/AONE?sid=googlescholar.
26. Bell FC, Wade A, Goss SC. Life Tables for the United States Social Security Area: 1900–2080, Actuarial Study No. 107, US Department of Health and Human Services. Soc Secur Adm Off Actuary SSA Pub. 1992; (11–11536).
27. Quehenberger F, Vasen HFA, van Houwelingen HC. Risk of colorectal and endometrial cancer for carriers of mutations of the hMLH1 and hMSH2 gene: correction for ascertainment. J Med Genet. 2005;42:491–6.
28. Wei EK, Colditz GA, Giovannucci EL, Fuchs CS, Rosner BA. Cumulative risk of colon cancer up to age 70 years by risk factor status using data from the Nurses' Health Study. Am J Epidemiol. 2009;170:863–72.
29. Rañola JMO, Liu Q, Rosenthal EA, Shirts BH. A comparison of cosegregation analysis methods for the clinical setting. Fam Cancer. 2018;17:295–302.
30. R-Forge: Cosegregation Analysis: R Development Page [Internet]. [cited 2018 Sep 11]. Available from: https://r-forge.r-project.org/R/?group_id=2174.
31. Pritchard CC, Mateo J, Walsh MF, De Sarkar N, Abida W, Beltran H, et al. Inherited DNA-repair gene mutations in men with metastatic Prostate Cancer. N Engl J Med. 2016;375:443–53.
32. NCCN - Evidence-Based Cancer Guidelines, Oncology Drug Compendium, Oncology Continuing Medical Education [Internet]. [cited 2018 Sep 10]. Available from: https://www.nccn.org/.
33. Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, Allen-Brady K, et al. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. Am J Hum Genet. 2007;81:873–83.