Review

Genetic Risk Scores and Missing Heritability in Ovarian Cancer

James P Brody * and Yasaman Fatapour

Posted Date: 28 January 2023

doi: 10.20944/preprints202301.0519.v1

Keywords: ovarian cancer; machine learning; germ line; genetic risk scores

Review

Genetic Risk Scores and Missing Heritability in Ovarian Cancer

Yasaman Fatapour and James P. Brody *

Department of Biomedical Engineering, University of California, Irvine
* Correspondence: jpbrody@uci.edu

Abstract: Ovarian cancers are curable by surgical resection when discovered early enough. Unfortunately, most ovarian cancers are diagnosed in the later stages. One strategy to identify early ovarian tumors is to screen women who have the highest risk scores. This mini review summarizes the accuracy of different methods used to assess the risk of developing ovarian cancer, including family history, BRCA genetic tests, and polygenic risk scores. The accuracy of these is compared to the maximum theoretical accuracy, revealing a substantial gap. We suggest that this gap, or missing heritability, could be caused by epistatic interactions between genes. An alternative approach to computing genetic risk scores, using chromosomal-scale length variation should incorporate epistatic interactions. Future research in this area should focus on this and other alternative methods of characterizing genomes.

Keywords: copy number variation; ovarian cancer; machine learning; h2o; germ line; UK Biobank; TCGA

1. Background

Ovarian cancer is known as the silent killer. The symptoms of ovarian cancer in the initial stages are minimal and non-specific. Constipation, heartburn, fatigue, and bloating are early signs of ovarian cancer, but also associated with other common maladies. Because of these non-specific symptoms, ovarian cancer is often un-diagnosed until the tumor has grown large, spread to nearby organs, and invaded the lymph system. At these later stages, treatment options are limited, and so is survival time. Ovarian tumors, like most solid tumors, can be surgically removed if found early. Removal of the tumor often leads to a complete cure[1]. However, most early detection strategies for ovarian cancer are ineffective for screening average risk women [2].

Current risk assessment tools for ovarian cancer do not work well enough. Specific genetic tests on BRCA1/BRCA2 status are available and work well for ovarian cancer, but only a small fraction (about 10%) of ovarian cancers are associated with those variants [3]. Otherwise, risk assessment is usually based on family history, but many people have limited knowledge of their family history and in any case a germ line genetic test should work better than a perfect family history. Development of a genetic test to identify women at high-risk of ovarian cancer could lead to a reduction in the number of ovarian cancer deaths.

2. Quantifying the accuracy of predictive tests.

Predictive tests often produce a numerical score that can be a continuous value, for instance from 1-100. From this score, one has to choose a cutoff value to make a prediction, which is a binary choice. Parameters like the sensitivity, specificity, positive predictive value, and negative predictive value are all a function of both the test and the choice of a cutoff value. The best way to characterize such a predictive test is with a Receiver Operating Characteristic curve[4,5]. This curve represents all cut off values, and one can read the sensitivity and specificity for the test for a given cutoff value.

The area under the curve of the receiver operating characteristic curve, or AUC, characterizes different predictive tests. The AUC, sometimes called a c-statistic, reduces the receiver operating characteristic curve to a single number, which is useful for comparing different tests. However, the

 doi:10.20944/preprints202301.0519.v1

complete ROC curve can show that two tests with similar AUC are not equivalent in some instances. Thus it is always best to examine the ROC curve for a test when judging its effectiveness.

The AUC can vary from 0.5, which is equivalent to random guessing, to 1.0, which indicates a perfect test that is always correct. The AUC is equivalent to the accuracy, when the two classes have equal numbers. The AUC is insensitive to class imbalance.

One example that illustrates how a predictive test with a low AUC can still be effective is the BRCA1 test for breast and ovarian cancer. This test works very well but only in a small subpopulation. Although the AUC is small, the test is quite valuable for that subpopulation.

3. Theoretical Maximum Accuracy of an Ovarian Cancer Genetic Risk Score

The **highest possible AUC for predicting ovarian cancer in women is about 0.99** [6]. According to [6], the discriminative accuracy of a genetic test depends on two factors, the heritability and prevalence of the trait. The Nordic Twin Study measured the heritability of ovarian cancer at about 40% [7]. Based on this heritability measurement and the prevalence of ovarian cancer, an ovarian cancer genetic test could have a maximum discrimination accuracy (AUC) in excess of 0.99. A substantial gap exists between the current best genetic risk tests and what should be possible.

4. Predicting Risk: Family History

Understanding a patient's family history is the first step in predicting whether a woman will develop ovarian cancer. Predictions based solely on family history have not been well characterized for ovarian cancer, but breast cancer predictive models have. For instance, one commonly used predictive model, the Gail model [8], has an **AUC of 0.58** (95% confidence interval [CI]=0.56 to 0.60) [9]. The Gail model incorporates several parameters including first degree relatives who were diagnosed with breast cancer but does not include any genetic information. Certain germ line mutations in BRCA1 and BRCA2 are known to increase the risk of ovarian cancer.

The Tyrer-Cuzick model includes a more detailed picture of genetics including BRCA1/BRCA2 status and a hypothetical low-penetrance gene that is designed to encompass all other genetic factors [10]. The Tyrer-Cuzick model is an improvement over the Gail model and has an **AUC = 0.62**, with a 95% CI of (0.60 to 0.64) [11].

Several mutations in the BRCA1/BRCA2 genes are known to increase the risk of developing ovarian cancer. However, these mutations account for only about 10% of ovarian cancers in the general population[3,12]. Similarly, the fraction of breast cancers attributable to mutations in BRCA1 or BRCA2 is about 10%. Thus, the best AUC we could expect for ovarian cancer predictive tests based on family history and supplemented with information on BRCA1/BRCA2 mutation status is probably similar to breast cancer, or about **AUC=0.60-0.65** [13–19].

The BRCA1/2 genetic tests are used to predict women at a high risk for breast and ovarian cancers. Some women whose BRCA test indicates a high risk of breast cancer choose to surgically remove their breasts to avoid breast cancer. Although less common, some women also choose a prophylactic oophorectomy--the surgical removal of the ovaries--to avoid ovarian cancers.

A positive BRCA1/2 test is highly predictive of breast/ovarian cancer, but a negative test is not very predictive of not having these cancers. In the US, only about 5-10% of breast and ovarian cancers are associated with mutations in BRCA1/2. A need exists to develop an effective genetic test for these other 90-95% of breast and ovarian cancers.

5. Predicting Risk: Polygenic Risk Scores

To fill this need, the most common approach is to use polygenic risk scores [13–19] . These are linear combinations of single nucleotide polymorphisms (SNPs) found more often in breast/ovarian cancer patients than in the general population. Models based on detailed germline genetics should perform better than models based on family history alone, since family history is often incomplete; limited to just a generation or two, and genetic factors present in relatives might not be inherited.

 doi:10.20944/preprints202301.0519.v1

The polygenic risk scores used today originate from Genome Wide Association Studies (GWAS)[20–22]. These GWAS studies were designed to find genes that drive disease, not for predictive tests. These polygenic risk scores are usually computed as a linear combination of the "hits," each with a different weight, found in GWAS studies. Different algorithms use slightly different criteria to decide on which "hits" to include and how to weigh them.

The current state of research knowledge on ovarian cancer genetic risk scores is best represented by two recent papers. The first was published in *JNCI* in 2020 [23] and the second was published in the *European Journal of Human Genetics* in 2022 [24].

The 2020 paper [23] evaluated polygenic risk scores for ovarian cancer, and seven other common cancers, using the UK Biobank. In this dataset, they identified 358 women who had been diagnosed with ovarian cancer. They constructed a polygenic risk score based upon 31 different SNPs. Then, they evaluated the performance of this polygenic risk score to predict ovarian cancer using the UK Biobank dataset. This test had a predictive accuracy of **AUC=0.568 (95% CI 0.537 to 0.598)**.

The second paper, with over 150 authors, is a *tour-de-force* [24]. Compared to the first paper, they increase the number of ovarian cancer subjects by nearly a factor of 100, using 23,564 cases. They thoroughly explored different combinations of SNPs and different algorithms for combining these SNPs into a polygenic risk score. The second paper [24] describes the best model found to be one based on measurements of 27,240 SNPs, almost 1000 times more than the 2020 paper [23]. After all that optimization, they found an **AUC of 0.588**. (They did not report a 95% confidence interval for the AUC).

Comparing the two papers, one can see that despite the extraordinary efforts of the second paper, the AUC of the test was not significantly higher than the first paper (**AUC=0.588** vs **95% CI 0.537 to 0.598)**. From this comparison, we can conclude that most of the useful information for predicting ovarian cancer has been extracted from SNP data using current algorithms. It seems unlikely that the AUC can be significantly improved with different algorithms, a different set of SNPs, or more patients in a dataset. This AUC is substantially lower than the theoretical maximum; something is missing.

6. Missing heritability?

Many human diseases, including ovarian cancer, are known to be inherited. It was thought that the advent of large scale genome wide association studies (GWAS) would reveal the underlying genes that led to this inheritance for different disease[25,26]. However, GWAS results have consistently shown that a substantial gap exists between the heritability that could be attributed to known factors by GWAS and the heritability observed by studying inheritance in families. The size of this gap varies by disease or trait, but it can be as large at a factor of ten [27]. The general missing heritability problem, and potential solutions, is well described by [26], in the specific case of ovarian cancer, Flaum *et al* put it succinctly:

> "However, a significant proportion of women who develop ovarian cancer with a strong family history of breast and/or ovarian cancer still do not have a known variant to explain their increased risk, and there must be other genetic factors at play that we do not yet understand."[12]

7. Beyond polygenic risk scores.

Epistatic interactions are one factor usually cited as part of the missing heritability problem[26,28]. The methods used in GWAS studies ignore non-linear interactions between genes, which are necessary to measure epistatic interactions. Modern statistical techniques, or machine learning, allow one to consider non-linear interactions between features, but these techniques inevitably require substantially more features (SNPs) than samples (patients), which is not useful when a few thousand patient samples is considered large, and genomes are characterized by millions of SNPs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2023 doi:10.20944/preprints202301.0519.v1

One approach to the problem is to construct a different representation of the genome as an alternative to SNPs. A more compact representation that still accounts for the variability in humans would allow the use of machine learning algorithms.

One example of this approach is to use measures of chromosome-scale length variation[29]. Chromosome-scale length variation can be computed from SNP array data. SNP arrays provide calibrated intensity values for each SNP location. This intensity data is usually processed into copy number variation data, which is represented by a multiplicity number (where two is the normal multiplicity) and chromosome segment. Instead, one can take this intensity data and compute an average multiplicity across an entire chromosome. By measuring this multiplicity across an entire chromosome for many people, one finds a distribution in values (See Figure 1). A person's germ line genome, then, can be characterized by a series of twenty-three numbers where each number represents the average multiplicity across each chromosome.

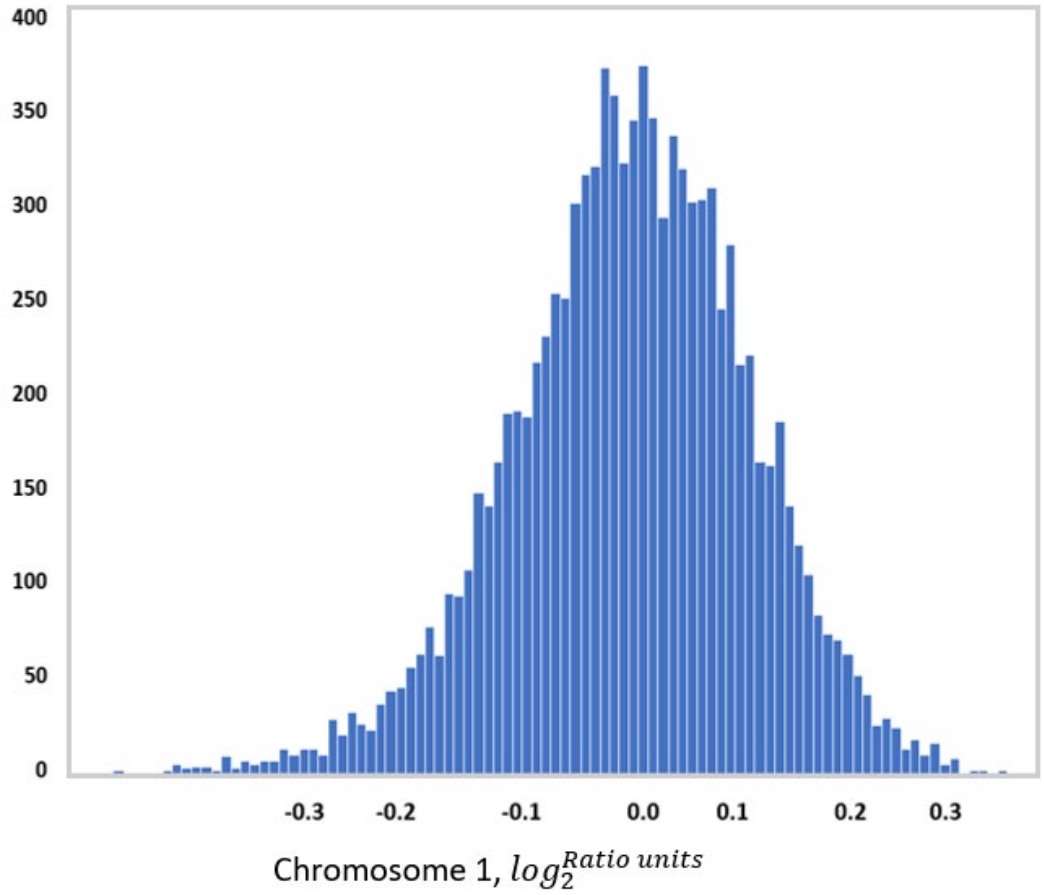

Figure 1. This figure shows a histogram of chromosome scale length variation measurements of Chromosome 1 for 10,000 people in the UK Biobank. "Chromosome length" is measured by averaging calibrated intensity measurements taken from SNP arrays for many SNPs located on Chromosome 1. These calibrated intensity measurements are representative of local copy number. Chromosomes can have many deletions, insertions, and translocations that affect copy number. The values measured in log_2(Ratio Units) represent the overall length of the chromosome, where a value of zero indicates the nominal average chromosome length. By measuring this parameter for all chromosomes, one can characterize each person's germ line genetic makeup with these 23 numbers. This compact representation of a person's genome can then be used to.

This representation of a person's genome, twenty-three decimal numbers, has some advantages over the conventional SNP representation of a genome. It is more compact, but still sufficiently complex to capture the enormity of the human population. The compactness allows one to use

modern machine learning techniques. It is extensible; you can split the chromosomes into arbitrarily small sections.

Using a data set acquired as part of the Cancer Genome Atlas (TCGA) project, we evaluated a genetic risk score computed from chromosomal scale length variation. In this data set, it had an AUC of 0.88 (95% CI of 0.86-0.91)[29]. Women with the highest 20% had 160 times the risk of developing ovarian cancer as compared to the lowest 20%. Although these numbers showed extraordinary discrimination, it is unclear whether these results are generalizable to the general population. The TCGA data set only contains people who had been diagnosed with cancer, so this work really distinguished one form of cancer from other forms of cancer. It is also possible that the TCGA contains subtle batch effects, leading to falsely high discrimination[30,31].

8. Conclusions

Ovarian cancer is completely curable in the early stages. The propensity to develop ovarian cancer appears to be transmitted through the genome. Thus, identification of signatures in the germline genome that indicate future diagnosis of ovarian cancer should be a primary and important target of research.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72.
2. Grossman DC, Curry SJ, Owens DK, Barry MJ, Davidson KW, Doubeni CA, et al. Screening for ovarian cancer US preventive services task force recommendation statement. JAMA - Journal of the American Medical Association. 2018.
3. Ramus SJ, Gayther SA. The Contribution of BRCA1 and BRCA2 to Ovarian Cancer. Mol Oncol. 2009.
4. Moons KGM, de Groot JAH, Linnet K, Reitsma JBR, Bossuyt PMM. Quantifying the added value of a diagnostic test or marker. Clin Chem. 2012.
5. Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. Clin Biochem Rev [Internet]. The Australian Association of Clinical Biochemists; 2008 [cited 2018 Jan 18];29 Suppl 1:S83-7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18852864
6. Janssens ACJW, Aulchenko YS, Elefante S, Borsboom GJJM, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases using multiple genes: Fact or fiction? Genetics in Medicine [Internet]. Nature Publishing Group; 2006 [cited 2020 Mar 12];8:395–400. Available from: http://www.nature.com/doifinder/10.1097/01.gim.0000229689.18263.f4
7. Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. JAMA [Internet]. American Medical Association; 2016 [cited 2018 May 27];315:68. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2015.17703
8. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. JNCI: Journal of the National Cancer Institute [Internet]. Oxford Academic; 1989 [cited 2021 Aug 25];81:1879–86. Available from: https://academic.oup.com/jnci/article/81/24/1879/1019887
9. Chlebowski RT, Anderson GL, Lane DS, Aragaki AK, Rohan T, Yasmeen S, et al. Predicting risk of breast cancer in postmenopausal women by hormone receptor status. J Natl Cancer Inst. 2007;99.
10. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med [Internet]. John Wiley & Sons, Ltd; 2004 [cited 2021 Aug 25];23:1111–30. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.1668
11. McCarthy AM, Guan Z, Welch M, Griffin ME, Sippo DA, Deng Z, et al. Performance of Breast Cancer Risk-Assessment Models in a Large Mammography Cohort. J Natl Cancer Inst. 2020;112.
12. Flaum N, Crosbie EJ, Edmondson RJ, Smith MJ, Evans DG. Epithelial ovarian cancer risk: A review of the current genetic landscape. Clin Genet [Internet]. Blackwell Publishing Ltd; 2020 [cited 2022 Mar 9];97:54–63. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/cge.13566
13. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19:581–90.
14. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med [Internet]. BioMed Central; 2020 [cited 2020 Jun 17];12:44. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00742-5
15. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Hum Mol Genet. 2019.

16. Hughes E, Tshiaba P, Gallagher S, Wagner S, Judkins T, Roa B, et al. Development and Validation of a Clinical Polygenic Risk Score to Predict Breast Cancer Risk. JCO Precis Oncol [Internet]. American Society of Clinical Oncology; 2020 [cited 2020 Aug 27];585–92. Available from: https://ascopubs.org/doi/10.1200/PO.19.00360
17. Khera A v., Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet [Internet]. Nature Publishing Group; 2018 [cited 2019 Apr 7];50:1219–24. Available from: http://www.nature.com/articles/s41588-018-0183-z
18. Sugrue LP, Desikan RS. What Are Polygenic Scores and Why Are They Important? JAMA [Internet]. American Medical Association; 2019 [cited 2020 Aug 30];321:1820. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2019.3893
19. Li R, Zhang X, Li B, Feng Q, Kottyan L, Luo Y, et al. Polygenic risk vectors (PRV) improve genetic risk stratification for cardio-metabolic diseases. medRxiv. 2022;
20. Goode EL, Chenevix-Trench G, Song H, Ramus SJ, Notaridou M, Lawrenson K, et al. A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. Nat Genet. 2010;
21. Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet [Internet]. 2009 [cited 2018 Jan 31];18:3525–31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19553258
22. Reid BM, Permuth JB, Chen YA, Fridley BL, Iversen ES, Chen Z, et al. Genome-wide Analysis of Common Copy Number Variation and Epithelial Ovarian Cancer Risk. Cancer Epidemiol Biomarkers Prev [Internet]. NIH Public Access; 2019 [cited 2020 Aug 20];28:1117–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30948450
23. Jia G, Lu Y, Wen W, Long J, Liu Y, Tao R, et al. Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. JNCI Cancer Spectr [Internet]. Oxford Academic; 2020 [cited 2022 Mar 13];4. Available from: https://academic.oup.com/jncics/article/4/3/pkaa021/5803653
24. Dareng EO, Tyrer JP, Barnes DR, Jones MR, Yang X, Aben KKH, et al. Polygenic risk modeling for prediction of epithelial ovarian cancer risk. European Journal of Human Genetics 2021 30:3 [Internet]. Nature Publishing Group; 2022 [cited 2022 Mar 13];30:349–62. Available from: https://www.nature.com/articles/s41431-021-00987-7
25. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009.
26. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010.
27. Génin E. Missing heritability of complex diseases: case solved? Hum Genet. 2020.
28. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A [Internet]. National Academy of Sciences; 2012 [cited 2018 Mar 26];109:1193–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22223662
29. Toh C, Brody JP. Genetic risk score for ovarian cancer based on chromosomal-scale length variation. BioData Min. BioMed Central Ltd; 2021;14.
30. Jain S, Mazaheri B, Raviv N, Bruck J. Short Tandem Repeats Information in TCGA is Statistically Biased by Amplification. bioRxiv. 2019;
31. Jain S, Mazaheri B, Raviv N, Bruck J. Glioblastoma signature in the DNA of blood-derived cells. PLoS One. 2021;16.