Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

A Review of Cancer Risk Prediction Models with Genetic Variants

Xuexia Wang[1,*], Michael J. Oldani[2], Xingwang Zhao[1], Xiaohui Huang[3] and Dajun Qian[4]

[1]Joseph J. Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA. [2]Criminology and Anthropology Department, University of Wisconsin-Whitewater, Whitewater, WI, USA. [3]Sanofi-Aventis, Bridgewater, NJ, USA. [4]City of Hope, Durate, CA, USA.

ABSTRACT: Cancer risk prediction models are important in identifying individuals at high risk of developing cancer, which could result in targeted screening and interventions to maximize the treatment benefit and minimize the burden of cancer. The cancer-associated genetic variants identified in genome-wide or candidate gene association studies have been shown to collectively enhance cancer risk prediction, improve our understanding of carcinogenesis, and possibly result in the development of targeted treatments for patients. In this article, we review the cancer risk prediction models that have been developed for popular cancers and assess their applicability, strengths, and weaknesses. We also discuss the factors to be considered for future development and improvement of models for cancer risk prediction.

KEYWORDS: cancer, risk prediction models, genetic variants, cancer risk prediction, cancer intervention

SUPPLEMENT: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

CITATION: Wang et al. A Review of Cancer Risk Prediction Models with Genetic Variants. *Cancer Informatics* 2014:13(S2) 19–28 doi: 10.4137/CIN.S13788.

RECEIVED: March 2, 2014. **RESUBMITTED:** June 30, 2014. **ACCEPTED FOR PUBLICATION:** July 1, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Review

FUNDING: Authors disclose no funding sources.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: xuexia@uwm.edu

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants. Provenance: the authors were invited to submit this paper.

Introduction

Cancer is one of the leading causes of death worldwide. A large percentage of patients are diagnosed at an advanced stage, making the removal of tumors in this population problematic. As a result, the overall 5-year survival rate is low for this cohort of patients.[1] Therefore, early stage detection would be helpful in reducing cancer mortality because treatment might be most effective at the earliest stages of the disease. For this reason, a well-established assessment model would greatly benefit patients, clinicians, and researchers because it would allow individuals at high risk to be identified at the earliest stages.

Cancer is a polygenic disease in which many genetic factors appear to play important roles in disease development in its different subtypes of cancer.[2] To date, more than 50 cancer genome-wide association studies (GWAS) incorporating more than 15 different malignancies have been reported identifying over 100 genomic cancer susceptibility regions.[3] The cancer-associated genetic variants identified in GWAS or candidate gene association studies have been shown to collectively enhance cancer risk prediction, improve our understanding of carcinogenesis, and possibly result in the development of targeted treatments for patients. For example, clinicians already use these kinds of guidelines in making decisions about assessments in order to identify carriers of *BRCA1* and *BRCA2* mutations, which indicate very high risks of breast and ovarian cancer.[4] The number of rapidly discovered cancer-associated genetic variants continues to rise and is reflected by the increasing number of published articles looking closely at the performance of genetic variants in popular cancer risk prediction models. These studies have prompted an updated assessment of the associations between genetic variants and cancer risk. Nevertheless, to date there has been no literature review concerning these publications, which provides an initial

assessment. This paper examines in detail the performance of cancer risk prediction models with genetic variants by examining the relevant studies through PubMed, Medline, and Web of Science. This review article summarizes what has been learned regarding the contribution of genetic variants as an alternative or as a supplement to the components of risk prediction models for cancer including breast cancer, prostate cancer, testicular cancer, lung cancer, and bladder cancer, as well as cancers of the head and neck.

Breast Cancer

Although the incidence rate of breast cancer has been declining since 1998–1999, there will still be 232,670 new female cases and 2,360 new male cases in the US in 2014 (http://www.cancer.gov/cancertopics/types/breast). Early stage detection of breast cancer is very important because treatment can be more effective at the early stages. For this reason, a well-established assessment model that could identify individuals at high risk would greatly benefit patients, clinicians, and researchers in the prevention and intervention of breast cancer.

Risk prediction models have been widely used to identify individuals with high risk of breast cancer. The Gail model,[5,6] for example, is used by FDA to screen women with high risk for chemopreventive use of tamoxifen. Traditional risk factors like family history, age at menarche, age at first live birth and number of previous breast biopsies, mammographic density, and diagnosis of atypical hyperplasia have been used to predict breast cancer as well.

Recently, genetic susceptibility risk prediction has been improved with the discovery of more than 40 risk-associated single-nucleotide polymorphisms (SNPs) from GWAS. Several breast cancer susceptibility genes have now been identified, including *BRCA1*,* *BRCA2*, *TP53*, and *PTEN/MMAC1*. Approximately, 60% of women with an inherited mutation in *BRCA1* or *BRCA2* will develop breast cancer sometime during their lives, compared with about 12% of women in the general population. Women with inherited *BRCA1* or *BRCA2* gene mutations also have an increased risk of ovarian cancer. Thus, evaluating these genetic susceptive predicting models has become crucial during clinical decision-making in order to help physicians and patients to determine whether a genetic testing is warranted.

Wacholder et al.[7] evaluated the Gail model using 5,590 cases and 5,998 control subjects, which are from four US cohort studies as well as from a Poland case control study. The range of age of all the subjects is from 50- to 79-years old. The area under the receiver-operating-characteristic (ROC) curve (AUC)** is 58% for four traditional risk factors. After incorporating 10 genetic variants into the prediction using a logistic regression model, the Wacholder study achieved a 61.8% AUC, a 3.8% increase over the model without genetic variants.

Another notable study in breast cancer risk prediction was done by Machiela et al.[8] In this study, a total of 1,145 breast cancer cases and 1,142 controls from the Nurses' Health Study were used to build and evaluate polygenic risk scores (PRSs) with 10–60,000 independent SNPs showing the strongest evidence of association with breast cancer. No significant evidence was found that polygenic risk score (PRS) using common variants could improve risk prediction for breast cancer over replicated SNP scores that had been robustly replicated across several independent sample sets.

Some polymorphisms identified in GWAS were also associated with an increased risk of breast cancer for *BRCA1* or *BRCA2* mutation carriers. Another study by Antoniou et al.[9] reanalyzed the association between breast cancer and six susceptibility polymorphisms in gene *FGFR2*, *TNRC9/TOX3*, *MAP3K1*, *LSP1*, 2q35 using a sample of 12,525 *BRCA1*, and 7,409 *BRCA2* carriers. The six susceptibility polymorphisms were identified in recent large-scale association studies conducted by the Consortium of Investigators of Modifiers of *BRCA1/2*.[10] Three additional SNPs (ie, rs4973768 in *SLC4A7/NEK10*, rs6504950 in *STXBP4/COX11*, and rs10941679 at 5p12) were also evaluated in this study. The interactions between SNPs were also investigated. Of the nine polymorphisms investigated, seven SNPs were found to be associated with breast cancer for *BRCA2* carriers and two SNPs were associated with *BRCA1* carriers. Additionally, interaction existed among all risk-associated polymorphisms for mutation carriers. Based on the joint genotype distribution of seven risk-associated SNPs in *BRCA2* mutation carriers, the top 5% high-risk *BRCA2* carriers were predicted to develop breast cancer by the age of 80 with a probability of 80–96%, whereas the bottom 5% low-risk *BRCA2* carriers only have a risk of 42–50% of developing breast cancer. Thus, the author concluded that these risk differences could be used in the day-to-day clinical management of mutation carriers.

Compared to high-penetrance mutations, such as *BRCA1* or *BRCA2*, all of the genetic susceptibility loci identified in GWAS to date are low-penetrance polymorphisms, with weak associations to breast cancer risk. Although each low-penetrance variant confers only a small increase in the risk of breast cancer, a combination of single variants may act cumulatively to increase the risk. For example, Sueta et al.[11] analyzed 23 genetic variants identified in previous GWASs and conducted a case–control study with 697 case subjects and 1,394 controls matched with age and menopausal status in the Japanese population. They fit conditional regression models with genetic variants and conventional risk factors. In addition, they created a polygenic risk score, using those variants with a statistically significant association with breast cancer risk, and also evaluated the contribution of these genetic predictors using AUC. Eleven SNPs revealed significant associations with breast cancer risk. In addition, a dose-dependent association was observed between the risks of breast cancer and the genetic risk score (GRS), which was an aggregate

*Appendix B. Description for the Abbreviations of Gene Names in this Article.
**More details about basic concepts used in this article can be found in Appendix A.

measure of alleles in seven selected variants. The AUC for the regression model, which included the GRS in addition to the conventional risk factors, was 0.6933, but it was only 0.6652 for conventional risk factors ($P = 1.3 \times 10^{-4}$). The population-attributable fraction of the risk score was 33.0%. Thus, this kind of study indicates that risk models, which include a GRS, are helpful in distinguishing women at high risk of breast cancer from those at low risk, particularly in the context of targeted prevention.

Prostate Cancer

Prostate cancer, behind only lung cancer, is the second leading cause of cancer-related deaths in American men. Recent data indicate that the estimated probability of being diagnosed with prostate cancer is 2.5%, 7%, and 13% for men ages 40–59 years, 60–69 years, and 70 years and older, respectively. In 2014, 233,000 new cases will be diagnosed in the US, and more than 29,480 men die of the disease (http://www.cancer.org/cancer/prostatecancer/detailedguide/prostate-cancer-key-statistics).

Prostate cancer is also a complex and unpredictable disease, with the risk for cancer affected by advancing age, ethnic background, and family history.[12]

Prostate cancer is usually accompanied by a rise in the concentration of serum prostate-specific antigen (PSA). PSA lacks specificity, but, nevertheless, has been used for decades as a sensitive biomarker and has evolved into a controversial predictor of prostate cancer mortality. In general, prostatic biopsies are often deemed unnecessary, which underscores the need for improving prediction models with increased specificity in order to aid clinicians when deciding whether or not to recommend a biopsy for patients. This is especially relevant for men with mildly elevated PSA values (3–10 ng/mL), where the risk of being diagnosed with prostate cancer is only about 20–25%.[13] After diagnosis, some cancers are indolent and cause no clinical problems, whereas others progress and may become fatal. Therefore, it is important to search for biomarkers that signal a need for more aggressive treatments, potentially improving clinical outcomes. Recently, more than 30 discovered SNPs have been associated with prostate cancer.[14] These SNPs provide an opportunity to identify strong candidates for a predictive role. SNPs identified and associated with prostate cancer in GWAS are common but confer only small increases in the risk. The mechanisms underlying their association with prostate cancer risk remain unknown.

Xu et al.[15] used SNPs of multiple DNA sequence variants and family history to estimate the absolute risk for prostate cancer. These investigators examined a Swedish study with 2,893 cases and 1,781 controls and a study in the US – the Prostate, Lung, Colon and Ovarian Cancer Screening Trial with 1,172 cases and 1,157 controls. Individuals with more than 14 risk alleles and positive family history had almost a five-fold increase in risk compared with people who had 11 risk alleles and negative family history. The study also outlined the risk of developing prostate cancer for a 55-year-old man who has positive family history and more than 14 risk alleles as being 40% over the next 20 years, while men without family history and such genotypes saw their absolute risk reduced to 13%.

In another study by Sun et al.[16] the investigators assessed predictive performance by employing positive predictive values (PPV) as well as sensitivity using family history and three sets of SNPs associated with prostate cancer. This study was a population-based case–control study (2,899 cases and 1,722 controls) in another Swedish population. SNPs and family history emerged as factors that can differentiate individual risk for prostate cancer, while identifying men at higher risk. In this particular study, the top 18% of men had a two-fold risk, while the top 8% had developed a three-fold risk of having prostate cancer in a 20-year period (age range: 55–74 years). In addition, the study showed that including more SNPs in the risk prediction will increase sensitivity with PPV.

Other studies have combined genetic variants along with PSA to predict the risk of prostate cancer. For example, a study by Johansson et al.[17] specifically combined 33 genetic variants with PSA and evaluated the risk. This was a case control study (520 cases and 988 controls) nested within the Northern Sweden Health and Disease Cohort. The AUC was used to assess whether the GRS of 33 SNPs in addition to prediagnostic PSA improves prostate cancer prediction. Adding GRS into the model improved the AUC from 86.2% to 87.2%. Thus, it appears that including GRSs into these models may not be beneficial when competing for a clinical risk assessment of prostate cancer.

Others such as Machiela et al.[8] created a new model by applying PRSs[18] and incorporating common variants to predict the risk of prostate cancer. A total of 1,164 prostate cancer cases and 1,113 controls from the Prostate Lung Colorectal and Ovarian Cancer Screening Trial were employed in this study. PRSs with 10 to 60,000 independent SNPs were used in the PRS model to compare with a model only including 30 published risk variants. Appling a 10-fold cross-validation for PRS model, the area under the ROC curve ranged from 0.564 (60,000 SNPs) to 0.569 (10 SNPs), while the AUC using 30 published risk SNPs from the literature was 0.614. Kote-Jarai[19] also proposed a study to predict the risk of prostate cancer using a multiplicative risk model, which combines the risk variants. This study used a worldwide consortium, composed of 13 groups with 7,370 prostate cases and 5,742 controls. All of the loci contributed to 16% of the familial risk of the disease, and the top 10% of risk distribution doubled the chance of prostate cancer with an odds ratio of 2.1.

The first risk prediction model for familial prostate cancer was developed by Macinnis et al.[20] which incorporated 26 prostate cancer-associated SNPs identified in previous GWAS.[21] Family phenotypes and histories were explained by a mixed model of inheritance which can be used to predict the probability of developing prostate cancer for an individual. Combined populations from 1,832 prostate cancer patients

and relatives in Australian and 2,558 patients from prostate cancer clinics at the Royal Marsden NHS Foundation Trust (UK) were used. Using this predictive model, the risk of prostate cancer for an UK male can be predicted. For example, if a man's genotype is in the top 10th percentile of joint genotype distribution and his father was diagnosed with prostate cancer at age 70, he would have a cumulative risk of 33% of developing prostate cancer by age 85. For a male with a genotype risk within the bottom 10%, the risk to develop prostate cancer would be 23%. In comparison, even without SNP information and incorporation into this kind of model, the risk remains 22% for a UK man.

Finally, Lindström et al.[22] combined a series of risk models and estimated their performance in 7,509 prostate cancer cases and 7,652 controls within the National Cancer Institute Breast and Prostate Cancer Cohort Consortium. The investigators also calculated absolute risks based on the Surveillance, Epidemiology, and End Results incidence data. The best risk model included individual genetic markers and family history of prostate cancer. They observed a decreasing trend in discriminative ability with advancing age, with highest accuracy in men younger than 60 years. The absolute 10-year risk for 50-year-old men with a family history ranged from 1.6% (10th percentile of genetic risk) to 6.7% (90th percentile of genetic risk). For men without a family history, the risk ranged from 0.8% (10th percentile of genetic risk) to 3.4% (90th percentile of genetic risk). These results indicate that incorporating both genetic information and family history into prostate cancer risk models can be particularly useful for identifying younger men who might benefit from PSA screening.

Testicular Cancer

Testicular cancer remains the most common form of cancer in men between the ages of 15 and 35 (http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002266/). It is also the most treatable form of cancer with a survival rate greater than 95% for the least aggressive type.

The risk of this type of cancer has been reported to be 8- to 10-fold higher for brothers and two- to four-fold higher for the sons of men who previously had testicular cancer.[23–27] Familial studies have estimated that genetic effects account for nearly a quarter of testicular cancer risk, which is one of the largest estimated heritabilities reported for any type of cancer.[28] More specifically, GWASs have implicated multiple genomic regions associated with testicular cancer risk, including those containing *KITLG*, *SPRY4*, *BAK1*, *ATF7IP*, *DMRT1*, and *TERT*.[29–33]

Previously published GWAS and candidate gene data have also been used to build a multiplicative model with risk variants and estimate the AUC as a measure of discrimination between testicular cancer cases and controls. Kratz et al.'s[34] study is one such example, of using this kind of data,[29–33] where previously uncovered predisposition alleles in or near *KITLG*, *BAK1*, *SPRY4*, *TERT*, *ATF7IP*, and *DMRT1* were used to predict the risk of testicular cancer by employing ROC curve analysis. The authors claim that an AUC of 69.2% suggests that about 69.2% of the time a randomly selected testicular cancer patient had a higher estimated risk than that of a randomly selected control subject. Another study showed how several established testicular germ cell tumor risk factors, such as cryptorchidism (relative risk (RR) = 4.8) and male infertility (standardized incidence ratio (SIR) = 2.8) can be incorporated into the clinical model to predict the risk.[35] Under this kind of multiplicative model, the authors estimated that white men in the top 1% of genetic risk as defined by eight risk variants had a relative risk that was 10.5-fold greater than that for a general population of similar male subjects.

We have to be aware as well that white men are more likely to develop testicular cancer than African-American and Asian-American men. Because of this race/ethnicity effect, cancer risk prediction needs to be tailored for specific populations. Additionally, GWAS needs to be extended into different populations. Each specific population requires models developed with their specificity in mind in order to create improved methods for overall risk assessment of testicular cancer.

Lung Cancer

Lung cancer remains the most common form of human cancer with complex risk factors, including genetic and environmental effects. Heredity plays an important role, and in relatives of people with lung cancer, the risk is increased 2.4 times.[36,37] This may be due, for example, to risks associated with genetic polymorphisms.[38–43] Environmental factors, such as a history of smoking, are central to several proposed lung cancer risk assessment models. These models include the Bach model, Spize model, and Liverpool Lung Project (LLP) model[44–47] as well as the improvement models based on LLP.[48,49] The LLP risk model,[45] developed from the LLP case–control study, provides a single unified model for smokers (current and former) and nonsmokers, whereas the Bach model was developed for predicting risk only in smokers and the Spitz model[46] requires three separate models for predicting risk in current smokers, former smokers, or nonsmokers. In addition, the LLP model also accounts for important lung cancer risk factors in addition to age, sex, and smoking duration. These include history of pneumonia, a history of non-lung cancer, prior asbestos exposure, and family history. Overall, this comprehensive model is simpler to incorporate into a clinical setting than Tammemagi and colleagues' model,[47] which includes many smoking-related variables that may be difficult to obtain from patients during clinical exchanges.

Other models have been developed in order to predict the 5-year absolute risk of lung cancer. For example, model based on five epidemiologic risk factors has been developed by the LLP by Raji et al.[48] where investigators quantified the improvement in risk prediction with the addition of SEZ6L, a Met430Ile polymorphic variant linked with an increased risk of lung cancer, within the framework of the LLP risk model.

In this predictive model, the authors combined the genotypes of 388 LLP subjects on SEZ6L SNP with epidemiologic risk factors. They use multivariable conditional logistic regression, with and without SEZ6L SNP, to predict 5-year absolute risk of lung cancer. Pair-wise comparison of the AUC and the net reclassification improvements (NRI) were also used to assess the improvement in the model itself with and without the SEZ6L SNP. The authors found a modest statistically significant increase in AUC when SEZ6L was added into the baseline model. The NRI for the genetic model was 27% with the SNP, while 15% without the SNP.

Raji et al. also further evaluated the LLP risk model in terms of discrimination and its ability to demonstrate a predictive benefit for stratifying patients for computed tomography (CT) screening.[49] These investigators assessed the 5-year absolute risks for lung cancer that were predicted by the LLP model in both case–control and prospective cohort study, which used data from three independent studies – the European Early Lung Cancer (EUELC), Harvard case–control studies, and the LLP population-based prospective cohort (LLPC) study from Europe and North America. The LLP risk model produced good discrimination in both the Harvard (AUC = 0.76 [95% confidence interval, CI, 0.75–0.78]) and the LLPC (AUC, 0.82 [CI, 0.80–0.85]) studies and modest discrimination in the EUELC (AUC, 0.67 [CI, 0.64–0.69]) study. The decision utility analysis, which incorporates the harm and benefit of using a risk model to make clinical decisions, indicated that the LLP risk model performed better than smoking duration or family history alone in stratifying high-risk patients for lung cancer CT screening. However, this model cannot assess whether the incorporation of other risk factors, such as lung function or genetic markers, will improve accuracy. In particular, the lack of information on asbestos exposure in the LLPC limited the ability to validate the complete LLP risk model.

Models and risk evaluation focused on genetic susceptibility loci have conferred a small to moderate disease risk and appear to be of limited utility in risk prediction. Li et al.[50] combined multiple disease-related loci with modest effects into a GRS and identified subgroups that were at high risk of lung cancer in a Chinese population. In their case–control study, they evaluated the discriminatory and predictive ability of the cumulative effect of several SNPs associated with lung cancer risk. Five SNPs identified in previous GWA or large cohort studies were genotyped in 5,068 Chinese case–control subjects. The GRS based on these SNPs was estimated by two approaches: a simple risk alleles count (cGRS) and a weighted (wGRS) method. The AUC in combination with the bootstrap resampling method was used to assess the predictive performance of the GRS for lung cancer. Four independent SNPs were found to be associated with a risk of lung cancer. The wGRS based on these four SNPs was a better predictor than cGRS. Using a liability threshold model, they estimated that these four SNPs accounted for only 4.02% of genetic variance in lung cancer. As with other studies, smoking history contributed significantly to lung cancer risk ($P < 0.001$) (AUC = 0.619 [0.603–0.634]), with the AUC value becoming 0.639 (0.621–0.652) after incorporation with wGRS and adjustment for over-fitting. Ultimately, this model shows some promise for assessing lung cancer risk in a Chinese population.

Black/white disparities concerning in lung cancer incidence and mortality mandate an evaluation of underlying biological differences. Etzel et al.[51] have previously shown higher risks of lung cancer associated with prior emphysema in African-American populations compared with white patients with lung cancer. Spitz et al.[52] further evaluated a panel of 1,440 inflammatory gene variants in a two-phase analysis (discovery and replication), adding top GWAS lung cancer hits from white populations, and 28 SNPs from a published gene panel. The discovery set (477 self-designated African-Americans cases, 366 controls matched on age, ethnicity, and gender) was from Houston, Texas. The external replication set (330 cases and 342 controls) was from the EXHALE study at Wayne State University. In discovery, 154 inflammation SNPs were significant ($P < 0.05$) on univariate analysis. One inflammation SNP, rs950286, which is intergenic between *IRF4* and *EXOC2* genes, was successfully replicated with a concordant odds ratio of 1.46 (1.14–1.87) in discovery, 1.37 (1.05–1.77) in replication, and a combined odds ratio of 1.40 (1.17–1.68). These researchers also constructed and validated an epidemiological discovery model. Furthermore, they extended risk prediction models, with the AUC for the epidemiologic discovery model being 0.77 and 0.80 for the extended model; for the combined datasets, the AUC values were 0.75 and 0.76, respectively.

Bladder Cancer

Bladder cancer remains a major health issue worldwide. In the US, bladder cancer is the fourth most common tumor in men and an estimated 74,690 new diagnoses are expected in 2014 (http://www.cancer.org/cancer/bladdercancer/detailedguide/bladder-cancer-key-statistics). The disease generally presents in older individuals, and is more common in men than women, with higher frequency among white patients than those of other ethnicities. Smoking is the most widely recognized cause of bladder cancer and accounts for half of all cases in the US.

The first risk prediction model for bladder cancer was developed by Wu et al.[53] in 2007. Patient epidemiologic and genetic data from a case–control study were used to build risk prediction models and constructed ROC. The AUC was used to evaluate the model's discriminatory ability. The model consisted of 678 white patients and 678 controls and included mutagen sensitivity and pack-years as well as six other risk factors, while achieving a 0.80 AUC, demonstrating good discrimination ability. In 2009,[54] the same group added three bladder cancer predisposition SNPs into the risk prediction

model but found no improvement of the discrimination power. However, Chen and colleagues[54] also pointed out that with the development of computing power and statistical tools, other risk factors such as gene–gene interaction and gene–environment interactions may offer greatly improved risk prediction.

In a recent paper published in *Cancer Research*, Garcia-Closas and colleagues[55] examined how genetic variants were recently identified in GWAS for bladder cancer interaction with smoking status to influence bladder cancer risk. The authors identified a new high-risk subgroup of individuals – current smokers carrying the highest genetic risk burden – who could be targeted for behavioral interventions and/or early detection protocols. This article is the first time to evaluate gene–environment interactions on risk difference, which indicates a new direction in bladder cancer prevention. Using data from seven studies, including 3,942 patients and 5,680 controls of European ancestry, the team investigated additive and multiplicative interactions between smoking status and 12 SNPs on the risk of developing bladder cancer. The SNPs selected for inclusion were recently identified bladder cancer susceptibility hits or known smoking metabolizing variants. To determine the combined effect of the SNPs across loci, the researchers created a PRS representing lowest to highest genetic risk quartiles. Smoking was assessed as lifetime history (ever/never), and as smoking status at the time of enrolment into the study (current/former/never). To gauge the public health relevance of their findings, they calculated the absolute risks resulting from the joint effects of smoking and the SNPs, and reported gene–environment interactions on the risk difference rather than relative risk of bladder cancer. Garcia-Closas and colleagues found that the cumulative 30-year absolute risk for bladder cancer in a 50-year-old US male varied by smoking status: 1.3% in never smokers, 3.0% for former smokers, and 6.2% for current smokers, confirming the importance of smoking as a strong risk factor for bladder cancer. When they factored in the PRS quartiles, the cumulative 30-year absolute risk for bladder cancer in a 50-year-old US male who is a current smoker and who carries the highest genetic risk jumped to 9.9%. Furthermore, they reported highly significant additive interactions between risk differences for smoking status across levels of PRS. They found that over four times more bladder cancer cases would be prevented if smoking were eliminated from the highest genetic risk group (n = 8,200 per 100,000 men) compared with the lowest genetic risk group (n = 2,000 per 100,000 men; $P < 0.0001$).

Head and Neck Cancer

The incidence of head and neck cancer has increased markedly in the last 20 years. Head and neck cancers account for about 3–5% of all cancers in the US. In this year, an estimated 55,070 people (40,220 men and 14,850 women) will develop head and neck cancers and 12,000 deaths (8,600 men and 3,400 women) will occur (http://www.cancer.net/cancer-types/head-and-neck-cancer/statistics).

Cigarette smoking is associated with increased head and neck cancer risk and tobacco-related carcinogens are known to cause bulky DNA adducts. Nucleotide excision repair genes encode enzymes that remove adducts and may be independently associated with head and neck cancer risk, as well as modifiers of the association between smoking and head and neck cancer risk.[56–58]

Several studies have reported that SNPs of genes in multiple biological pathways are involved in the development of head and neck cancer.[59–63] Recently, Annah et al.[61] performed a two-stage GWAS with a total of 8,605 cases and 11,405 controls and reported that five genetic variants had significant associations with risk of upper aerodigestive tract cancers including head and neck cancer in Europeans.

With the recent increase in associated SNPs with head and neck cancer being identified, the development of risk prediction models is catching up. A study by Wu et al.[64] used a customized chip containing 9,645 chromosomal and mitochondrial SNPs (mtSNPs) to call genotypes for 150 early stage head and neck cancer patients with 300 controls. The goal is to model the second primary tumor or head and neck cancer recurrence using both clinical and epidemiological variables. Results showed that when 12 chromosomal SNPs and one mtSNP were incorporated into the model, the AUC increased from 0.64 to 0.84. The 95% CI of the AUC difference is 0.18–0.29, indicating significant improvement in discrimination power.

Discussion

Cancer is a polygenic disease in which many genetic factors appear to play important roles in disease development in its different subtypes of cancer.[2] During the past several years, more than 100 SNPs have been identified that are associated with cancer.[3]

How to effectively incorporate these genetic susceptive variants in risk predictive models has become more and more important during the clinical decision-making process because effective models can help physicians and patients determine whether a genetic testing is needed. Although enormous progress has been made in the area of genetics and the susceptive risk prediction of cancer, cautions should be made when considering the application of these models within the clinical setting. Cancer remains a fundamentally complex disease with multiple, interacting risk factors. These risk factors include components of race/ethnicity, environmental carcinogens, familial history, genetic variants, and their interactions. The studies reviewed here should be understood as an initial attempt to begin a more systematic approach to assessing predictive risk models for cancer treatment in the future. In order to more accurately predict the overall risk of cancer in patients, risk prediction models need to be continuously reexamined,

Table 1. Performance of cancer risk prediction models with genetic variants.

CANCER TYPE	RISK PREDICTION MODEL	TYPES OF GENETIC FACTOR	MEASURES OF PERFORMANCE	VARIANTS IN PREDICTION	REFERENCE
Breast	Logistic regression model	Individual SNP and GRS	AUC	Helpful	7
	Logistic regression model	PRS	AUC	Not helpful	8
	Conditional regression model	PRS	AUC	Helpful	11
Prostate	Logistic regression model	GRS	Relative risk	Helpful	15
	Multiplicative model	Individual SNPs	PPV and sensitivity	Helpful	16
	Logistic regression model	GRS	AUC	Not helpful	17
	Logistic regression and multiplicative model	Individual SNPs and GRS	Overall familial risk	Helpful	19
	Mixed recessive model	PRS	LRT and AIC	Helpful	20
	Logistic regression model	Individual SNPs and GRS	AUC	Helpful	22
Testicular	Multiplicative model	Individual SNPs	AUC	Helpful	34
Lung	Conditional logistic regression	Individual SNPs	AUC and NRI	Helpful	48
	Logistic regression model	GRS	AUC	Helpful	50
	Logistic regression model	Individual SNPs	AUC	Helpful	52
Bladder	Logistic regression model	Individual SNPs	AUC	Helpful	53
	Logistic regression model	Individual SNPs and PRS	Bootstrap resampling	Helpful	55
HNC	Cox proportional hazard model	Individual SNPs	AUC	Helpful	64

Abbreviations: NRI, the net reclassification improvements; PPV, positive predictive value; AUC, the area under the receiver operator characteristic (ROC) curve (AUC); PRS, polygenic risk score; GRS, genetic risk score; HNC, head and neck cancer; PSA, prostate-specific antigen; AIC, Akaike's A Information Criterion; LRT, Likelihood ratio tests.

comprehensively assessed, and revised, taking into consideration specific populations and emergent subtype of cancer.

Table 1 indicates that the cancer risk prediction models with genetic variants generally outperform the models without genetic variants in both discrimination and prediction of cancer. However, there are still many practical concerns on implementing genetic testing into the diagnostic process. For example, the substantial cost of genetic screening is one of the main concerns.

Table 2 summarizes the frequently used risk prediction models with genetic factors and general modeling procedures. The most commonly used model is logistic regression model. When dealing with multiple genetic factors and other covariates, logistic regression assumes a linear relationship among the predictors and uses a logit link to combine them into a one-dimensional fitted value.

Although more than 50 cancer GWAS incorporating more than 15 different malignancies have been reported, identifying over 100 genomic cancer susceptibility regions,[3] for most malignancies the number of consistently confirmed SNPs is less than a dozen. The lack of power of GWAS suggests that there may exist many more SNPs associated with some malignancies that have smaller effect sizes. However, such SNPs may be statistically insignificant in genome-wide. How to effectively incorporate these SNPs in a risk prediction model is challenging since a group of these SNPs may likely make a positive contribution to a risk prediction model while the other ones may just add some noise. Evans et al.[18] and Purcell et al.[65] have proposed methods of aggregating information on a large number of SNP alleles associated with a trait that does not achieve stringent genome-wide statistical significance or even nominal statistical significance of $P < 0.05$. These models create PRS by summing risk alleles from thousands or tens of thousands of loci spanning the genome to predict an individual's genetic risk of developing disease. Michielsla et al.[8] built a logistic regression model and used PRS to reflect the genetic effect of lists of genetic markers prioritized by their association with breast cancer in a training dataset and evaluated whether these scores could improve current genetic prediction of these specific cancers in independent test samples. However, the logistic regression model integrating PRS did not outperform the model without PRS. Whereas, the study of Sueta et al.[11] demonstrates that the regression model including PRS of seven published variants outperform the model without PRS with an increased AUC as 2.81%. Both Sueta et al.[11] and Michielsla et al.[8] use logistic regression models including PRS to predict breast cancer risk, but the performance of PRS in the two studies are quite different. Sueta et al. create the PRS based on published SNPs with a statistically marginally significant association with breast cancer risk. Michielsla et al.[8] build PRS based on 10–60,000 common SNPs in their GWAS. The comparison of the two studies indicates how effectively selecting SNPs when creating PRS is critical and will affect the performance of the prediction model.

Table 2. The frequently used risk prediction models and general modeling procedures.

RISK PREDICTION MODELS	TYPE OF GENETIC FACTOR	MODELING PROCEDURES	REFERENCES
Logistic regression model	Individual SNPs, PRS, or GRS	Assuming a linear relationship among multiple predictors and using a logit link to combine them into a one dimensional fitted value. Usually start with the main effects model. Univariate logistic regression analysis is performed to assess the main effects of each individual risk factor. Then, perform a stepwise logistic regression analysis to identify significant predictors in a multivariate model.	7,22,52 8,15,17,22
Multiplicative model	Individual SNPs	A multiplicative model is used to derive genotype relative risks by multiplying the allelic odds ratio (OR) of each SNP which is obtained from a marginal test. An individual is affected if his genotype relative risk is greater than a threshold.	15,16
Conditional logistic regression	Individual SNPs	Conditional logistic regression works in nearly the same way as regular logistic regression except we need to specify which individuals belong to which matched set or stratum.	48
Cox proportional hazard model	Individual SNPs	For each SNP, the risks of disease occurrence is estimated as hazard ratios (HRs) using multivariable Cox proportional hazard regression models adjusted for age, gender, ethnicity, smoking status, tumor site, stage, and treatment, where appropriate.	64

Compared with traditional risk factors such as family history, smoking, age, and sex, sometimes the impact of genetic variants in predicting risk is small which may reflect the small effect size of disease-associated SNPs integrated in the risk prediction model. Due to the difference in effect sizes of associated SNPs, the power of genetic variants in prediction for different cancers is different as well. A recent report uses PRS to estimate the relative risks of disease. In these reported estimates, the predictive power is higher for prostate cancer than for breast cancer, which reflects the fact that the known associated SNP effect sizes for prostate cancer are greater and account for a larger percentage of the familial relative risk.[66]

In the article, we reviewed the cancer risk prediction models by different cancer types. But many cancers share the same major oncogenic or tumor suppressor genes such as *KRAS*, *P53*, *SRC*, *HER2/neu*, *RAF*, and *MYC*. Most oncogenes display a very broad tumor spectrum. For example, abnormalities of the *P53* gene (which codes for the P53 protein) have been found in more than half of human cancers. Acquired mutations of this gene appear in a wide range of cancers, including lung, colorectal, and breast cancer. The predictive power of the same oncogenic gene might be different for different cancer types. If the incidence of the disease is low, such as ovarian cancer, the predictive power for ovarian cancer might be low as well. This advocates the need for the analysis of substantially larger numbers of cases, especially if there is significant variability across histological subtypes of the disease.

Author Contributions

Conceived and designed the experiments: XW, XH. Analyzed the data: XH, XZ. Wrote the first draft of the manuscript: XW, XH. Contributed to the writing of the manuscript: MO, XZ, DQ. Agree with manuscript results and conclusions: XW, MO, XZ, XH, DQ. Jointly developed the structure and arguments for the paper: XW, MO. Made critical revisions and approved final version: XW, MO. All authors reviewed and approved of the final manuscript.

REFERENCES

1. de Martel C, Ferlay J, Franceschi S, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol*. 2012;13:607–15.
2. Schully SD, Yu W, McCallum V, et al. Cancer GAMAdb: database of cancer genetic associations from meta-analyses and genome-wide association studies. *Eur J Hum Genet*. 2011;19:928–30.
3. Stadler ZK, Vijai J, Thom P, et al. Genome-wide association studies of cancer predisposition. *Hematol Oncol Clin North Am*. 2010;24:973–96.
4. Lu K, Kauff N, Powell CB, et al. Hereditary Breast and Ovarian Cancer Syndrome. ACOG PRACTICE BULLETIN, *Obstet Gynecol*. 2009;113:957–66.
5. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*. 1989;81:1879–86.
6. Gail MH, Benichou J. Validation studies on a model for breast cancer risk. *J Natl Cancer Inst*. 1994;86:573–5.
7. Wacholder S, Hartge P, Prentice R, et al. Performance of common genetic variants in breast-cancer risk models. *N Engl J Med*. 2010;362:986–93.
8. Machiela MJ, Chen CY, Chen C, Chanock SJ, Hunter DJ, Kraft P. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genet Epidemiol*. 2011;35:506–14.
9. Antoniou AC, Beesley J, McGuffog L, et al. Common breast cancer susceptibility alleles and the risk of breast cancer for BRCA1 and BRCA2 mutation carriers: implications for risk prediction. *Cancer Res*. 2010;70(23):9742–54.
10. Chenevix-Trench G, Milne RL, Antoniou AC, Couch FJ, Easton DF, Goldgar DE. An international initiative to identify genetic modifiers of cancer risk in BRCA1 and BRCA2 mutation carriers: the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA). *Breast Cancer Res*. 2007;9:104.
11. Sueta A, Ito H, Kawase T, et al. A genetic risk predictor for breast cancer using a combination of low-penetrance polymorphisms in a Japanese population. *Breast Cancer Res Treat*. 2012;132:711–21.
12. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer-analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med*. 2000;343:78–85.
13. Amundadottir LT, Thorvaldsson S, Gudbjartsson DF, et al. Cancer as a complex phenotype: pattern of cancer distribution within and beyond the nuclear family. *PLoS Med*. 2004;1(3):e65.
14. Goh CL and Eeles RA. Germline genetic variants associated with prostate cancer and potential relevance to clinical practice. prostate cancer prevention. *Recent Results Cancer Res*. 2014;202:9–26.

15. Xu J, Sun J, Kader AK, et al. Estimation of absolute risk for prostate cancer using genetic markers and family history. *Prostate*. 2009;69:1565–72.
16. Sun J, Kader AK, Hsu FC, et al. Inherited genetic markers discovered to date are able to identify a significant number of men at considerably elevated risk for prostate cancer. *Prostate*. 2011;71:421–30.
17. Johansson M, Holmstrom B, Hinchliffe SR, et al. Combining 33 genetic variants with prostate-specific antigen for prediction of prostate cancer: Longitudinal study. *Int J Cancer*. 2012;130:129–37.
18. 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*. 2009;18:3525–31.
19. Kote-Jarai Z, Easton DF, Stanford JL, et al. Multiple novel prostate cancer predisposition loci confirmed by an international study: the PRACTICAL Consortium. *Cancer Epidemiol Biomarkers Prev*. 2008;17:2052–61.
20. Macinnis RJ, Antoniou AC, Eeles RA, et al. A risk prediction algorithm based on family history and common genetic variants: application to prostate cancer with potential clinical impact. *Genet Epidemiol*. 2011;35:549–56.
21. Al Olama AA, Kote-Jarai Z, Giles GG, et al. Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet*. 2009;41:1058–60.
22. Lindström S, Schumacher FR, Cox D, et al. Common genetic variants in prostate cancer risk prediction – results from the NCI Breast and Prostate Cancer Cohort Consortium (BPC3). *Cancer Epidemiol Biomarkers Prev*. 2012;21(3):437–44.
23. Dieckmann KP, Pichlmeier U. The prevalence of familial testicular cancer: an analysis of two patient populations and a review of the literature. *Cancer*. 1997;80:1954–60.
24. Forman D, Oliver RT, Brett AR, et al. Familial testicular cancer: a report of the UK family register, estimation of risk and an HLA class 1 sib-pair analysis. *Br J Cancer*. 1992;65:255–62.
25. Heimdal K, Olsson H, Tretli S, Flodgren P, Borresen AL, Fossa SD. Familial testicular cancer in Norway and southern Sweden. *Br J Cancer*. 1996;73:964–9.
26. Hemminki K, Chen B. Familial risks in testicular cancer as a etiological clues. *Int J Androl*. 2009;29:205–10.
27. Hemminki K, Li X. Familial risk in testicular cancer as a clue to a heritable and environmental etiology. *Br J Cancer*. 2004;90:1765–70.
28. Czene K, Lichtenstein P, Hemminki K. (2002) Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int J Cancer*. 2002;99:260–6.
29. Nathanson KL, Kanetsky PA, Hawes R, et al. The Y deletion gr/gr and susceptibility to testicular germ cell tumor. *Am J Hum Genet*. 2005;77:1034–43.
30. Kanetsky PA, Mitra N, Vardhanabhuti S, et al. Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nat Genet*. 2009;41:811–5.
31. Kanetsky PA, Mitra N, Vardhanabhuti S, et al. A second independent locus within DMRT1 is associated with testicular germ cell tumor susceptibility. *Hum Mol Genet*. 2011;20:3109–17.
32. Rapley EA, Turnbull C, Al Olama AA, et al. A genome-wide association study of testicular germ cell tumor. *Nat Genet*. 2009;41:807–10.
33. Turnbull C, Rapley EA, Seal S, et al. Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nat Genet*. 2010;42:604–7.
34. Kratz CP, Greene MH, Bratslavsky G, Shi J. A stratified genetic risk assessment for testicular cancer. *Int J Androl*. 2011;34:e98–102.
35. Dieckmann KP, Pichlmeier U. Clinical epidemiology of testicular germ cell tumors. *World J Urol*. 2004;22:2–14.
36. Li X, Hemminki K. Familial and second lung cancers: a nation-wide epidemiologic study from Sweden. *Lung Cancer*. 2003;39:255–63.
37. Jonsson S, Thorsteinsdottir U, Gudbjartsson DF, et al. Familial risk of lung carcinoma in the Icelandic population. *JAMA*. 2004;292:2977–83.
38. McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15.33. *Nat Genet*. 2008;40:1404–6.
39. Rafnar T, Sulem P, Stacey SN, et al. Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat Genet*. 2009;41:221–7.
40. Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet*. 2008;40:616–22.
41. Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*. 2008;452:633–7.
42. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*. 2008;452:638–42.
43. Wang Y, Broderick P, Webb E, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet*. 2008;40:1407–9.
44. Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. *J Natl Cancer Inst*. 2003;95:470–8.
45. Cassidy A, Myles JP, van Tongeren M, et al. The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer*. 2008;98:270–6.
46. Spitz MR, Hong WK, Amos CI, et al. A risk model for prediction of lung cancer. *J Natl Cancer Inst*. 2007;99:715–26.
47. Tammemagi CM, Pinsky PF, Caporaso NE, et al. Lung cancer risk prediction: prostate, lung, colorectal and ovarian cancer screening trial models and validation. *J Natl Cancer Inst*. 2011;103:1058–68.
48. Raji O, Agbaje OF, Duffy SW, Cassidy A, Field JK. Incorporation of a genetic factor into an epidemiologic model for prediction of individual risk of lung cancer: the Liverpool Lung Project. *Cancer Prev Res*. 2010;3:664–9.
49. Raji O, Duffy SW, Agbaje OF, et al. Predictive accuracy of the liverpool lung project risk model for stratifying patients for computed tomography screening for lung cancer: a case-control and Cohort validation study. *Ann Intern Med*. 2012;157(4):242–50.
50. Li H, Yang L, Zhao X, et al. Prediction of lung cancer risk in a Chinese population using a multifactorial genetic model. *BMC Med Genet*. 2012;13:118.
51. Etzel CJ, Kachroo S, Liu M, et al. Development and validation of a lung cancer risk prediction model for African-Americans. *Cancer Prev Res (Phila)*. 2008;1:255–65.
52. Spitz MR, Amos CI, Land S, et al. Role of selected genetic variants in lung cancer risk in African Americans. *J Thorac Oncol*. 2013;8:391–7.
53. Wu X, Lin J, Grossman HB, et al. Projecting individualized probabilities of developing bladder cancer in white individuals. *J Clin Oncol*. 2007;25:4974–81.
54. Chen M, Cassidy A, Gu J, et al. Genetic variations in PI3 K-AKT-mTOR pathway and bladder cancer risk. *Carcinogenesis*. 2009;30(12):2047–52.
55. Garcia-Closas M, Rothman N, Figueroa JD, et al. Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. *Cancer Res*. 2013;73:2211–20.
56. Neumann AS, Sturgis EM, Wei Q. Nucleotide excision repair as a marker for susceptibility to tobacco-related cancers: a review of molecular epidemiological studies. *Mol Carcinog*. 2005;42:65–92.
57. Friedberg EC. How nucleotide excision repair protects against cancer. *Nat Rev Cancer*. 2001;1:22–33.
58. Goode EL, Ulrich CM, Potter JD. Polymorphisms in DNA repair genes and associations with cancer risk. *Cancer Epidemiol Biomarkers Prev*. 2002;11:1513–30.
59. McKay JD, Truong T, Gaborieau V, et al. A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. *PLoS Genet*. 2011;7:e1001333.
60. Hashibe M, McKay JD, Curado MP, et al. Multiple ADH genes are associated with upper aerodigestive cancers. *Nat Genet*. 2008;40:707–9.
61. Annah B. Wyss, Amy H. Herring, Christy L. Avery, et al. Single-Nucleotide polymorphisms in nucleotide excision repair genes, cigarette smoking, and the risk of head and neck cancer. *Cancer Epidemiol Biomarkers Prev*. 2013; 22:1428–45.
62. Wyss AB, Herring AH, Avery CL, et al. Single-nucleotide polymorphisms in nucleotide excision repair genes, cigarette smoking, and the risk of head and neck cancer. *Cancer Epidemiol Biomarkers Prev*. 2013;22:1428–45.
63. Yuan H, Ma H, Lu F, et al. Genetic variants at 4q23 and 12q24 are associated with head and neck cancer risk in China. *Mol Carcinog*. 2013;52:e2–9.
64. Wu X, Spitz MR, Lee JJ, et al. Novel susceptibility loci for second primary tumors/recurrence in head and neck cancer patients: large scale evaluation of genetic variants. *Cancer Prev Res (Phila)*. 2009;2(7):617–24.
65. International Schizophrenia Consortium, Purcell SM, Wray NR, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009;460:748–52.
66. Bahcall O. Risk prediction and population screening for breast, ovarian and prostate cancers. 2014:doi: 10.1038/ngicogs.5.
67. Moons KGM, Kengne AP, Woodward M, Royston P et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of new (bio)marker. *Heart*. 2012; 98:683–90.
68. Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet*. 2013;9(3):e1003348. doi:10.1371/journal.pgen.1003348.

Appendix A. Basic Concepts Related to Risk Prediction Models Used in this Article

Risk prediction model.[67] A statistical model is used to estimate the risk of future outcomes for individuals based on one or more underlying characteristics. Such characteristics often simply referred to as predictors traditionally not only include standard features such as age, sex, smoking, and family history but also increasingly include genetic variants identified in genetic association studies.

Polygenic risk scores.[68] Risk alleles among an ensemble of markers that do not individually achieve significance in a large-scale association study can be summarized as a score. This score can be used to test association between the selected markers and a trait and predict an individual's genetic risk of developing disease. Moreover, these scores could be used to predict an individual's outcome even without knowing which of the SNPs in the score are conclusively associated with disease.

Sensitivity. The probability that a test will indicate "disease" among those with the disease.

Specificity. The fraction of those without disease will have a negative test result.

Positive predictive value. The probability that the patient actually has the disease if the test result is positive.

Area under the receiver-operating-characteristic (ROC) curve (AUC). ROC curve analysis is commonly adopted in cancer predictive studies to evaluate the performance of a predictive test. The ROC curve plots the test's sensitivity against one specificity by continuously changing the cutoff points over the whole range of possible test results, and is often summarized by its one-dimension summary index – the area under the ROC curve (AUC). The AUC measures discrimination, that is, the ability of the predictive test to correctly classify those with and without the disease. A test with perfect discrimination has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore, the closer the ROC curve is to the upper left corner, the larger the AUC, which means the higher the overall accuracy of the test.

Appendix B. Description for the Abbreviations of Gene Names in this Article

ATF7IP – activating transcription factor 7 interacting protein

BAK1 – BCL*2*-antagonist/killer 1

DMRT1 – double sex and mab-3 related transcription factor 1

EXOC2 – exocyst complex component 2

FGFR2 – fibroblast growth factor receptors 2

HER2 – v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2

IRF4 – interferon regulatory factor 4

KITLG – KIT ligand

KRAS – Kirsten rat sarcoma viral oncogene homolog

LSP1 – lymphocyte-specific protein 1

MAP3K1 – mitogen-activated protein kinase kinase kinase 1

MYC – v-myc avian myelocytomatosis viral oncogene homolog

PTEN (*MMAC1*) – phosphatase and tensin homolog

P53(*TP53*) – tumor protein p53

RAF – Raf kinases

SEZ6L – seizure related 6 homolog (mouse)-like

SLC4A7(*NEK10*) – solute carrier family 4, sodium bicarbonate cotransporter, member 7

SRC – SRC proto-oncogene, non-receptor tyrosine kinase

STXBP4(*COX11*) – syntaxin binding protein 4

SPRY4 – sprouty homolog 4 (Drosophila)

TNRC9(*TOX3*) – trinucleotide-repeat-containing 9

TERT – telomerase reverse transcriptase