









## Methods

# Variance of age-specific log incidence decomposition (VALID): a unifying model of measured and unmeasured genetic and non-genetic risks

John L Hopper,<sup>1,\*</sup> James G Dowty,<sup>1</sup> Tuong L Nguyen ,<sup>1</sup> Shuai Li ,<sup>1</sup> Gillian S Dite ,<sup>1,2</sup> Robert J MacLinnis,<sup>1,3</sup> Enes Makalic ,<sup>1</sup> Daniel F Schmidt,<sup>1,4</sup> Minh Bui,<sup>1</sup> Jennifer Stone ,<sup>5</sup> Joonho Sung ,<sup>6</sup> Mark A Jenkins ,<sup>1</sup> Graham G Giles,<sup>3</sup> Melissa C Southey,<sup>3,7</sup> and John D Mathews <sup>1</sup>

<sup>1</sup>Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, VIC, Australia, <sup>2</sup>Genetic Technologies Ltd., Fitzroy, VIC, Australia, <sup>3</sup>Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, VIC, Australia, <sup>4</sup>Faculty of Information Technology, Monash University, Clayton, VIC, Australia, <sup>5</sup>School of Population and Global Health, University of Western Australia, Perth, WA, Australia, <sup>6</sup>Division of Genome and Health Big Data, Department of Public Health Sciences, Graduate School of Public Health, Seoul National University, Seoul, Korea and <sup>7</sup>Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC, Australia

\*Corresponding author. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, 207 Bouverie Street, Victoria 3010, Australia. E-mail: [j.hopper@unimelb.edu.au](mailto:j.hopper@unimelb.edu.au)

Received 11 July 2022; Editorial decision 15 May 2023; Accepted 16 June 2023

## Abstract

**Background:** The extent to which known and unknown factors explain how much people of the same age differ in disease risk is fundamental to epidemiology. Risk factors can be correlated in relatives, so familial aspects of risk (genetic and non-genetic) must be considered.

**Development:** We present a unifying model (VALID) for variance in risk, with risk defined as  $\log(\text{incidence})$  or  $\text{logit}(\text{cumulative incidence})$ . Consider a normally distributed risk score with incidence increasing exponentially as the risk increases. VALID's building block is variance in risk,  $\Delta^2$ , where  $\Delta = \log(\text{OPERA})$  is the difference in mean between cases and controls and OPERA is the odds ratio per standard deviation. A risk score correlated  $r$  between a pair of relatives generates a familial odds ratio of  $\exp(r\Delta^2)$ . Familial risk ratios, therefore, can be converted into variance components of risk, extending Fisher's classic decomposition of familial variation to binary traits. Under VALID, there is a natural upper limit to variance in risk caused by genetic factors, determined by the familial odds ratio for genetically identical twin pairs, but not to variation caused by non-genetic factors.

**Application:** For female breast cancer, VALID quantified how much variance in risk is explained—at different ages—by known and unknown major genes and polygenes, non-genomic risk factors correlated in relatives, and known individual-specific factors.

**Conclusion:** VALID has shown that, while substantial genetic risk factors have been discovered, much is unknown about genetic and familial aspects of breast cancer risk especially for young women, and little is known about individual-specific variance in risk.

**Key words:** Breast cancer, familial cause, familial odds ratio, familial risk ratio, genetic cause, genomic cause, major gene, non-familial cause, polygenic risk score, variance components

### Key Messages

- Risk can be defined as age-specific log(incidence) or cumulative risk.
- The key metric for defining the risk discrimination of a risk factor is  $\Delta$  = the log of the change in odds ratio per standard deviation of a possibly adjusted and transformed risk score with unit variance.
- $\Delta$  = the difference between cases and controls in mean risk score.
- $\Delta^2$  = the variance in risk attributed to this risk score.
- We show how variation in risk can be partitioned into measured and unmeasured genetic and non-genetic components.
- Variation in genetic risk is finite and its upper limit can be determined from the disease association (specifically the familial odds ratio which approximates the familial risk ratio for most diseases) within genetically identical (monozygotic: MZ) twin pairs.
- Genetic factors will not be important for risk prediction if the MZ twin pair odds ratio is weak, irrespective of disease frequency.
- Variation in non-genetic risk is unlimited.

## Background

A fundamental issue for epidemiology is the extent to which known and unknown factors explain how much people of the same age differ from one another in their disease risk. Given that risk factors can be correlated in relatives, familial risk factors—both (germline) genetic and non-genetic (e.g. shared environment)—must be considered.

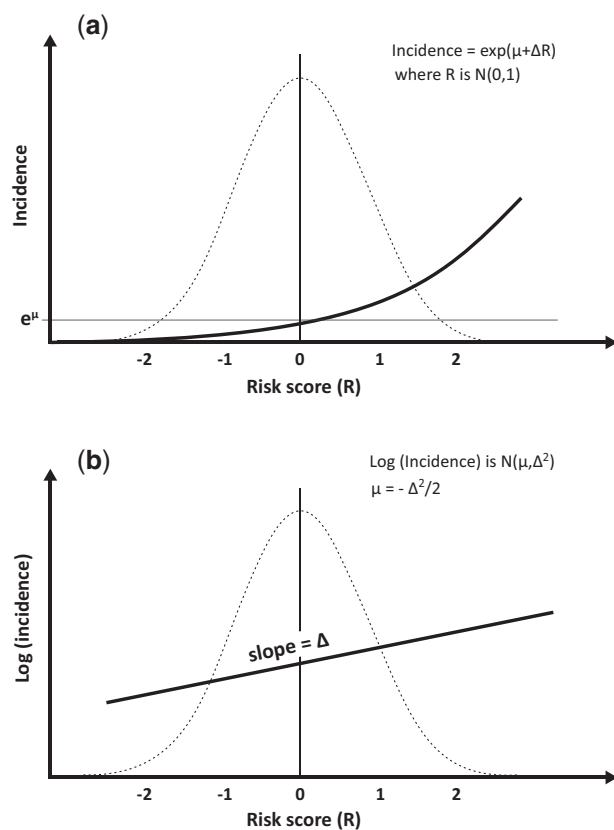
This paper introduces a unifying model called Variance of Age-specific Log Incidence Decomposition (VALID), with risk defined as the age-specific log(incidence) or logit(cumulative incidence). Variance in risk as a quantitative trait is the building block. VALID considers familial and non-familial, genetic and non-genetic, measured and unmeasured variance in risk. It therefore brings together individual-specific and familial risks, including lifestyle, polygenes, major genes and shared environment, known and unknown.

VALID is in part based on Fisher's seminal 1918 paper<sup>1</sup> that introduced the concept of unmeasured genetic and non-genetic causes of variation in measured quantitative outcomes (traits); see Historical context in the [Supplementary Material](#) (available as [Supplementary data](#)

at *IJE* online). Fisher warned that the concept of 'heritability' could be misleading,<sup>2</sup> as we found when studying large immigrant and non-immigrant sibships.<sup>3</sup> Here we essentially extend Fisher's model to disease risk, and thereby to binary traits in general. Whereas Fisher converted familial correlations into variances in measured quantitative traits, VALID converts familial odds ratios into variances in risk.

## Modelling genetic and non-genetic familial and non-familial causes of variation

For a trait with total variance  $\sigma^2$  and an additive genetic component with variance  $A$ , the trait correlation is  $r_{MZ} = A/\sigma^2$  for monozygotic (MZ) twin pairs,  $r_{DZ} = 1/2A/\sigma^2$  for dizygotic (DZ) twin pairs and other first-degree relatives,  $1/4A/\sigma^2$  for second-degree relatives, and so on.<sup>1</sup> This model was extended to include environmental (i.e. non-genetic) causes shared by (or common to) relatives, whose variance has historically been denoted by  $C$ . The classic twin model assumes  $C$  is the same for MZ and DZ pairs, so  $A = 2(r_{MZ} - r_{DZ})\sigma^2$  and  $C = (2r_{DZ} - r_{MZ})\sigma^2$  provided  $2r_{DZ} > r_{MZ}$  (Falconer's formula).<sup>4</sup> Under a flexible parametrization fitted using, for example, the multivariate normal model,<sup>5,6</sup>



**Figure 1** Incidence as an exponential function (a) and log(incidence) as a linear function (b) of the risk score under the VALID model for a risk score with a standard normal distribution, superimposed on the risk score's density function (dotted line)

the model can be extended to families, the genetic component can be modelled as a function of measured factors either as fixed or as random effects,<sup>7</sup> and the shared environmental variance component,  $C$ , can take into account factors such as the extent to which pairs of relatives cohabit, have cohabited or have lived apart; see Modelling of the familial causes of variance in risk below.

## Development

### Risk score versus risk factor

We represent a risk factor (which might be a composite of risk factors such as genetic markers) as a risk score that has a standard normal distribution, that disease incidence increases exponentially as the risk score increases (see Figure 1a), for which log(incidence) increases linearly as the risk score increases (see Figure 1b), at least when incidence is small (see below).

These characteristics have been observed for the combined associations of common genetic variants on risk of breast cancer based on additivity on the log risk scale both within and between markers to create an 'additive'

polygenic risk score.<sup>8</sup> This model is also inherent to case-control and cohort study analyses using logistic and Cox regression, respectively; see 'Why log(incidence)?' in the Conclusion.

We are studying variation in relative risk, not absolute risk per se, so the risk score must be adjusted for age and possibly other covariates, as should be standard practice in epidemiology. This approach underlies the odds ratio per adjusted standard deviation (OPERA) concept as a population measure of risk discrimination.<sup>9</sup> For concreteness, we take risk to mean the log(incidence), although the VALID concept also applies to log(odds ratio) = logit(odds) and therefore to cumulative risk (e.g. lifetime risk or risk to a given age) or any binary trait in general. Our main interest is in diseases, not common traits.

VALID essentially follows models by us and others,<sup>10–15</sup> except that here we assume the risk score has been standardized to have unit variance. This is important when interpreting the term 'risk score'. Pharoah and colleagues<sup>12</sup> and Clayton<sup>13</sup> refer to the polygenic risk score,  $R$ , as having a log-normal distribution such that  $\log(R) = Y$  is distributed as  $N(\mu, \sigma^2)$ . VALID considers  $Z = (Y - \mu)/\sigma$ , which has a standard normal  $N(0,1)$  distribution. The difference between cases and controls in mean  $Y$  is  $\sigma^2$ ;<sup>12</sup> so the difference between cases and controls in mean  $Z$  is  $\sigma$ .

### Parameterization

Figure 2 shows the key parameters involved in the VALID model. The strength of a risk score, in terms of its ability to differentiate cases from appropriate controls on a population basis, is assessed by log(OPERA), where OPERA is the odds ratio per adjusted standard deviation. The adjusted standard deviation is the standard deviation of the residuals after the risk factor has been adjusted for age and potentially other measures.<sup>9,16,17</sup> Given that what is estimated for an adjusted risk factor is the change in risk per unit change of the risk factor, while conceptually holding constant all those measures taken into account by sampling and analysis, it is not appropriate to use the odds ratio per 'unadjusted' standard deviation.

Consider a risk score that is normally distributed for both cases and controls, and with the same variance in these two groups which, without loss of generality, we take to be 1. Let  $\Delta$  = the difference between cases and controls in mean risk score. Then:

$$\Delta = \log(\text{OPERA}). \quad (1)$$

(see Relationship between OPERA and  $\Delta$  in the Supplementary Material, and in the Supplementary

Risk factor	Factor associated with risk of disease for persons of the same age
Risk score	Risk factor normalised, adjusted for age and standardised (variance = 1)
OPERA	Odds ratio per standard deviation of adjusted and normalised risk factor
$\Delta = \log(\text{OPERA})$	Difference in mean risk score between cases and controls
$\sigma^2 = \Delta^2$	Variance in age-specific log incidence; difference in mean $\log(\text{incidence})$ between cases and controls
$\text{FRR} \sim \exp(r\sigma^2)$	Familial risk ratio generated by a risk score correlated $r$ in relatives
AUC	Area under the receiver operating characteristic score; linearly related to $\sigma = \Delta$ in the range 0 to 1.2 and less so thereafter; linearly related to $\sigma = \Delta$ after being probit transformed
Tetrachoric correlation	Linearly related to $\sigma^2 = \Delta^2$ in the range 0 to 1.4, and less so thereafter; linearly related to $\log(\text{FRR})$ after Fisher Z transformed; differs by disease frequency

**Figure 2** Definitions, descriptions, and relationships between major concepts underlying the VALID model

**Material** (available as [Supplementary data](#) at *IJE* online) in Schmidt DF and colleagues<sup>18</sup>).

Figure 1b shows the linear relationship between  $\log(\text{incidence})$  and the standardised risk score where  $\log(\text{incidence})$  has a normal distribution with mean  $\mu$  and variance  $\Delta^2$ .

There is a simple relationship between  $\Delta = \log(\text{OPERA})$  and the area under the receiver operating characteristic curve (AUC) given by:

$$\text{AUC} = \Phi(\Delta/\sqrt{2}), \quad (2)$$

where  $\Phi$  is the cumulative distribution function of the standard normal distribution (see Relationship between AUC and  $\Delta = \log(\text{OPERA})$  in [Supplementary Material](#)<sup>18</sup>). Therefore  $\Delta = \log(\text{OPERA})$  is linearly related to probit transformed AUC irrespective of the disease prevalence. It is the difference between cases and controls in the mean of the standardized risk score and is also referred to in different ways in different disciplines, such as Cohen's  $D$ .<sup>19</sup> Figure 3 shows the distribution of  $\log(\text{incidence})$  for cases and controls in the situation where  $\Delta = 1.2$  and  $\text{AUC} = 0.8$ .

The variance of the  $\log(\text{incidence})$  is:

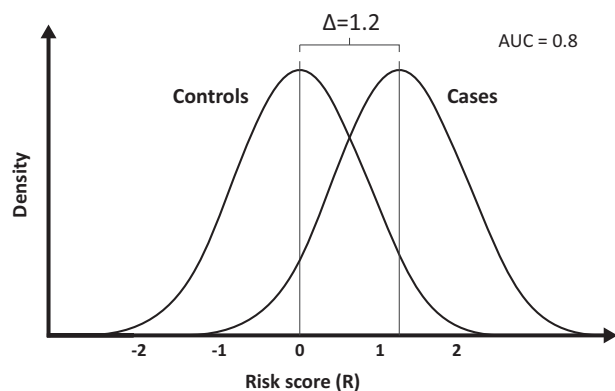
$$\sigma^2 = \Delta^2 = [\log(\text{OPERA})]^2 \quad (3)$$

and is the square of the difference in means between cases and controls on the standardized risk score scale, the difference in mean  $\log(\text{incidence})$  between cases and controls, and the square of the logarithm of the odds ratio per standard deviation of the risk score; see [Figure 2](#).<sup>12</sup>

#### Familial risk caused by familial aspects of a risk factor: unifying equation

For a given pair of relatives,  $\text{rel} = \text{twin pairs, siblings, etc}$ , let the familial odds ratio be the odds of disease for the relative of an affected person divided by the odds of disease for the same type of relative of an unaffected person. A risk factor with a correlation in risk score between relatives of  $r_{\text{rel}}$ , and a risk gradient of  $\Delta = \log(\text{OPERA})$ , generates a corresponding:

$$\text{familial odds ratio} = \exp(r_{\text{rel}}\Delta^2). \quad (4)$$



**Figure 3** Density of the risk score distribution under the VALID model for cases and controls when  $\Delta = 1.2$  and the area under the receiver operating characteristic curve (AUC) = 0.8

Given we are interested in diseases (see above), the familial odds ratio is approximately equal to the familial risk ratio ( $FRR_{rel}$ ) = the risk of disease for the unaffected relative of an affected person divided by the risk for the same type of unaffected relative of an unaffected person. In this setting:

$$FRR_{rel} \sim \exp(r_{rel}\Delta^2). \tag{5}$$

Once the relationship between  $\Delta$  and the interquartile risk ratio is understood (see Relationship between IQRR and  $\Delta$  in [Supplementary Material](#),<sup>18</sup>) it can be seen that [Equation \(5\)](#) was in effect derived by Aalen<sup>10</sup> under the assumption of a multiplicative risk and a ‘rare’ disease. For a polygenic model, [Equation \(4\)](#) was derived by Pharoah and colleagues<sup>12</sup> and Clayton proved it was a good approximation for both the multiplicative and logistic risk models.<sup>13</sup> [Equation \(4\)](#) had previously been shown to apply to specific instances by Hopper and Carlin.<sup>11</sup>

We refer to [Equation \(4\)](#) as the Unifying Equation. It is fundamental to genetic epidemiology and plays a critical role in VALID because it allows the familial aspects of any risk factor to be interpreted in terms of its contribution to the disease association for all pairs of relatives. For diseases, [Equation \(5\)](#) implies that:

$$\Delta = [\log(FRR_{rel})/r_{rel}]^{0.5} \tag{6}$$

and from (2) and (5),

$$AUC = \Phi\{[\log(FRR_{rel})/2r_{rel}]^{0.5}\}. \tag{7}$$

If the only cause of familial risk is genetic factors such that, for first degree-relatives,  $r_{rel} = 0.5$ , then:

$$AUC = \Phi\{[\log(FRR_{rel})]^{0.5}\}. \tag{8}$$

**Table 1** Comparative tabulation of different parameters of risk discrimination

AUC <sup>a</sup>	OPERA <sup>b</sup>	$\Delta^c$	$\Delta^{2d}$	$FRR_{MZ}^e$	IQRR <sup>f</sup>	UQRR <sup>g</sup>
0.50	1	0	0	1	1	1
0.55	1.2	0.18	0.03	1.04	1.6	1.2
0.60	1.4	0.36	0.13	1.14	2.5	1.5
0.65	1.7	0.54	0.30	1.34	4.0	1.8
0.70	2.1	0.74	0.55	1.74	6.7	2.1
0.75	2.6	0.95	0.91	2.50	12	2.4
0.80	3.3	1.19	1.42	4.12	22	2.8
0.85	4.3	1.47	2.15	8.58	49	3.1
0.90	6.1	1.81	3.28	26.8	135	3.5
0.95	10	2.33	5.41	224	706	3.8

<sup>a</sup>Area under the receiver operating characteristic curve (AUC).

<sup>b</sup>Odds ratio per standard deviation of the adjusted risk factor (OPERA).

<sup>c</sup>Difference in mean between cases and controls ( $\Delta = \log(OPERA)$ ).

<sup>d</sup>Variance in  $\log(\text{incidence})$  ( $\Delta^2$ ).

<sup>e</sup>Familial risk ratio if  $r = 1.0$  ( $FRR_{MZ} = \exp(\Delta^2)$ ).

<sup>f</sup>Interquartile risk ratio (IQRR).

<sup>g</sup>Upper-quartile risk ratio to the population average (UQRR).

Under this assumption, if the FRR for first degree relatives is 2, then the maximum AUC that can be achieved by knowing all additive genetic factors is 0.80, corresponding to  $\Delta = 1.2$  and  $\sigma^2 = 1.4$ ; see [Figure 3](#).

[Table 1](#) shows the different risk discrimination parameters for a selection of values across their ranges sufficient to allow for reasonably accurate interpolation.

### Modelling the familial causes of variance in risk

For the point of illustration, consider the classic twin model which makes the ‘equal environments assumption’ that the non-genetic effects shared by twins are the same for both MZ and DZ pairs. This assumption maximizes the proportion of familial variance attributed to genetic factors.

Suppose that the variance in risk can be decomposed into an additive genetic component (A) and a shared environment component (C) as described in Background. The risk score represents germline genetic factors for which  $r_{rel}$  can be modelled in terms of the kinship coefficients following Fisher,<sup>1</sup> and the effects of non-genetic factors shared by twins can be modelled in various ways; see below.

For monozygotic (MZ) twin pairs,  $r_{rel} = 1$ . For dizygotic (DZ) twin and sibling pairs:

$$r_{rel} = (0.5A + C)/(A + C). \tag{9}$$

This model can be extended to other relatives.<sup>3</sup>

The shared environmental variance component, C, can be modelled perhaps more informatively by taking into

account the extent to which pairs of relatives cohabit, have cohabited or have lived apart.<sup>6</sup> Non-genetic effects shared by parents and offspring,<sup>20–24</sup> spouse associations<sup>3,23,24</sup> and variations that take into account the birth order can be modelled.<sup>25,26</sup> Despite evidence that shared environment has different roles for different types of relatives, even for those of the same degree of genetic relationship,<sup>3,21,22</sup> this more nuanced modelling has not been popular among genetic researchers. Recently, we analysed epigenetic data for twins and family from across the lifespan and found evidence for non-genetic factors that would otherwise have been attributed to genes.<sup>23,24</sup> Given familial aggregation is highly age-dependent, at least for breast cancer,<sup>27,28</sup> it is also important to consider age and cohabitation aspects of both A and C.

### Combining risk factors

For two factors whose risk associations are virtually independent, in that their individual risk gradients  $\Delta_i$  ( $i = 1, 2$ ) are essentially the same whether they are fitted alone or together, let  $\Delta_{12}$  be their combined risk gradient when they are fitted together. Then:

$$\Delta_{12} \sim (\Delta_1^2 + \Delta_2^2)^{0.5}. \quad (10)$$

An exact and more general formula for  $\Delta_{12}$  is given in the [Supplementary Material](#) where its validity is shown for a special case.

Heuristic justification for the approximate formula comes from interpretation of  $\Delta$  as the difference between cases and controls in mean risk score. If two (uncorrelated) risk scores are combined, the distance in means in two-dimensional space is the hypotenuse of a right-angled triangle whose sides are the differences in means for each of the risk scores. This argument can be extended to  $n > 2$  independent risk factors in which case:

$$\Delta_{1\dots n} \sim (\Delta_1^2 + \dots + \Delta_n^2)^{0.5}. \quad (11)$$

If the two risk scores are not acting independently (i.e. their associations are correlated) their combined

associations would be attenuated, as would the third side of a less than right-angled triangle; see [Supplementary Material](#). Therefore, the risk variance for a combination of independent risk scores,  $\Delta_{1\dots n}^2$ , is approximately the sum of the variances of the independent components,  $\Delta_i^2$ . This variance will be attenuated if the risk scores capture some risk factor information in common, which can also be overcome by using the OPERA concept.

### Application

As in Hopper and Carlin<sup>11</sup> we study female breast cancer, but model variance in age-specific  $\log(\text{incidence})$ .

### Unmeasured familial factors

First, we consider unmeasured familial factors by analysing twin associations estimated by the Nordic Twin Study,<sup>28</sup> which takes into account temporal and censored aspects lacking in an earlier publication.<sup>29</sup>

Column two of [Table 2](#) shows that the FRR for MZ pairs decreases from 5.91 before age 50 years to 2.50 by age 80 years. Column four shows that, given  $r_{\text{rel}} = 1$  for MZ pairs and [Equation \(5\)](#), the maximum variance decreases from  $\log(5.91) = 1.78$  to  $\log(2.50) = 0.92$ .

Under the classic twin model and using [Equation \(8\)](#), column six shows that the additive genetic variance (A) decreases with age from 1.04 to 0.66, and column seven shows that the shared environment variance (C) decreases from 0.74 to 0.26. Therefore, on average about two-thirds of the declining familial variance is attributed to genetic factors irrespective of age.

### Measured familial factors

#### Genomic risk factors

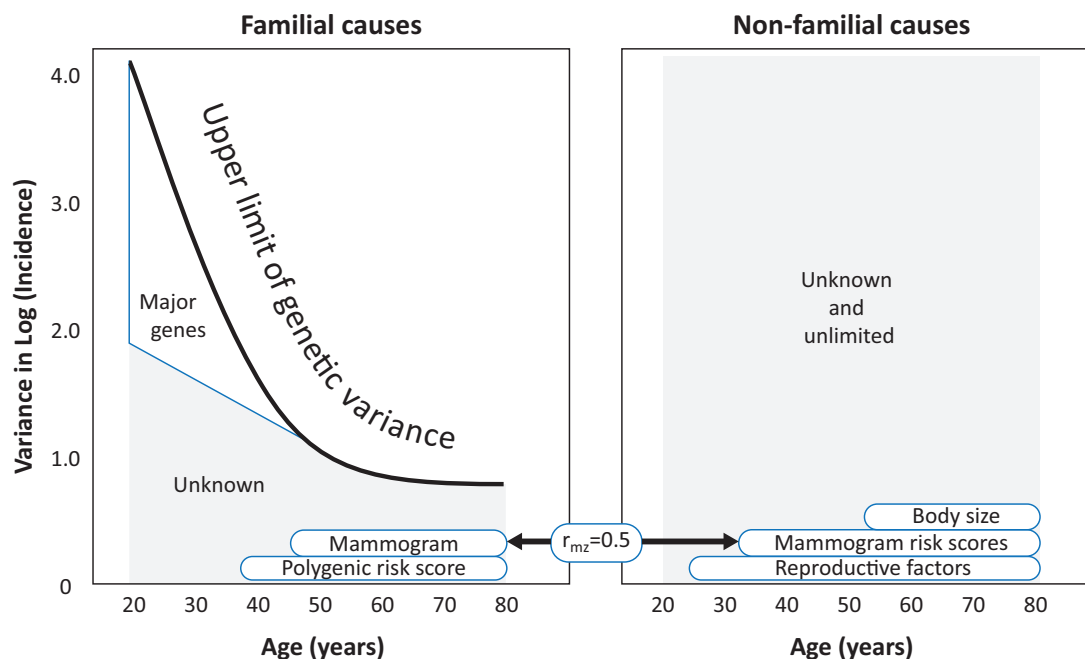
Segregation analyses of multigenerational family data have also found that the total familial variance decreases with age. A substantial proportion of variance at young ages is explained by the major breast cancer susceptibility genes

**Table 2** Familial relative risk (FRR), twin pair covariance in  $\log(\text{incidence})$ , additive genetic (A) and shared environmental (C) components of variance in  $\log(\text{incidence})$ , and maximum area under the receiver operating characteristics curve from knowing all genetic causes ( $\text{AUC}_{\text{max}}$ ) based on data from the Nordic Twin Study of Breast Cancer<sup>28</sup>

Age (years)	FRR MZ	FRR DZ	Covariance MZ	Covariance DZ	A	C	$\text{AUC}_{\text{max}}$
<50	5.91	3.51	1.78	1.26	1.04	0.74	0.83
50–59	4.93	2.77	1.60	1.02	1.15	0.44	0.81
60–69	2.98	2.24	1.09	0.81	0.57	0.52	0.77
70–79	2.5	1.8	0.92	0.59	0.66	0.26	0.75

MZ, monozygotic twin pairs; DZ, dizygotic twin pairs.





**Figure 4** Decomposition of variance in log(incidence) of breast cancer by age according to familial effects, including rare high-risk variants in major genes such as *BRCA1* and *BRCA2*, polygenic risk scores, mammogram risk scores which have a substantial familial component and some other epidemiological risk factors that are mostly non-familial, based on literature cited in the text

*BRCA1* and *BRCA2*, and a small proportion by other major genes including *ATM*, *PALB2* and *Tp53*.<sup>30</sup> These major genes explain little variance for post-menopausal women; see Figure 4.

The OPERA for the current best breast cancer polygenic risk score (PRS) is  $\log(1.65) = 0.50$  so the variance explained is  $(0.50)^2 = 0.25$ .<sup>8</sup> This association is similar across all ages, although perhaps weaker before age 40 years. For women under the age of 50 years, a PRS based on 77 single nucleotide polymorphisms (SNPs) did not explain any familial risk of breast cancer diagnosed before age 50 years.<sup>31</sup> Therefore, much remains to be learned about the polygenic risk for breast cancers diagnosed at a young age; see Figure 4.

#### Non-genomic risk factors

Many non-genomic risk factors have been identified from questionnaire data. These include reproductive factors such as number and timing of live births and ages at menarche and menopause, as well as anthropometric factors height, and for post-menopausal women weight, which have historically been combined as body mass index. The risk gradients are modest, with OPERAs in the range of 1.005 to 1.2.<sup>9</sup>

Questionnaires attempt to reveal aetiologically relevant processes which, if measured more precisely, would have greater risk gradients. Almost all these non-genomic risk factors are correlated in relatives, usually only weakly.

Therefore, they generate familial as well as mostly non-familial components of variance, and the Unifying Equation (4) describes how these are apportioned.

#### Familial aspects of non-genomic risk factors

As an example, multiple mammogram risk scores (MRSs) based on different aspects of a mammogram are being found to be associated with breast cancer risk. These include conventional mammographic density, mammographic density measured at high brightness pixel thresholds,<sup>16,32–38</sup> and textural features and other agnostic measures learned by machine learning.<sup>18,39,40</sup>

The correlation in the MRSs based on mammographic density is about 0.6 for MZ pairs and 0.3 for DZ and sister pairs.<sup>41,42</sup> The risk gradient for an MRS based on conventional mammographic density has an OPERA of about 1.5, so the variance is about 0.16, of which 0.10 would be familial and 0.06 non-familial. The risk gradient is greater for the new MRS, and when combined could be as high as 2.1,<sup>37</sup> in which case the variance would be 0.55. If the MZ twin pair correlations of these new MRS are similar to those for conventional density,<sup>42</sup> they could explain as much if not more familial variance than the current PRS.

#### Non-genomic non-familial risk factors

Most variation in questionnaire-based risk factors is individual-specific and makes minimal contribution to familial variance; see Figure 4. Greater specificity of

exposures will increase the variance due to known non-familial factors, as is being found with the new MRS being discovered by applying artificial intelligence to digital mammography.<sup>34</sup> Application to epigenetics might reveal new and mostly individual-specific risk factors.<sup>23,24,43</sup>

### Combinations of risk factors: independence and interactions

In general, the risk associations for known risk factors (i.e. relative risks for women of the same age) do not change greatly when fitted together; in epidemiological parlance, these associations are said to be ‘independent’ because they are additive on a particular scale. But this can be misleading. Given epidemiological analyses use the log or logit scales, a ‘lack of interaction’ on those scales means the associations of risk factors tend to multiply on one another on the absolute risk scale, on which there must be ‘interactions’ because the greater a woman is on one risk factor, the greater is her absolute risk gradient on another risk factor.<sup>44,45</sup>

### Combining polygenic risk scores with risk scores based on family history

Polygenic risk scores are familial, so their (relative) risk associations will not necessarily be independent of family history associations. We constructed a continuous familial risk score (FRS)<sup>46</sup> from multigenerational family history data using, for example, the BOADICEA pedigree-based model.<sup>47</sup> We estimated risk associations with and without fitting an established PRS and found that, for breast cancer diagnosed before age 50 years, the FRS and PRS were not correlated and their risk associations were independent. That is, the PRS discovered using mostly samples of post-menopausal women explains at most a small proportion of why breast cancer diagnosed at a young age runs in families. [Figure 4](#) shows that the major genes and other factors dwarf the contribution of the PRS to familial risk variance in this younger age range.<sup>30</sup>

### Combining mammographic risk scores with polygenic risk scores

We originally predicted that ~10% of the familial variance of breast cancer is explained by familial aspects of mammographic density (adjusted for age and body mass index).<sup>48</sup> This was corroborated by estimating the change in family history associations after adjusting for this MRS.<sup>49</sup> About the same proportion of SNPs associated with breast cancer have been found to be nominally associated with

this MRS,<sup>50</sup> but the current best PRS SNPs is at best only weakly correlated with this familial MRS.<sup>51,52</sup>

### Conclusion

For any risk factor, once appropriately converted into a multiplicative risk score, its ability to differentiate cases from controls is dictated by the risk gradient,  $\log(\text{OPERA})$ , the square of which is the variance in risk. The familial aspects of variance can be estimated from the familial odds ratio using the Unifying Equation. The familial risk variance can be decomposed into genetic and non-genetic components by returning to Fisher’s seminal 1918 paper that converted familial correlations into variance components; see [Supplementary Material](#) or a discussion of the historical context. VALID converts familial risk ratios into variance components of risk for familial and non-familial factors, genetic and non-genetic aspects of familial risk, genomic and non-genomic aspects of genetic risk, and familial and non-familial aspects of non-genomic risk; see [Figure 4](#).

VALID is underpinned by the OPERA concept<sup>9</sup> and the key metric is  $\Delta = \log(\text{OPERA})$ , a natural risk gradient for a risk score.  $\Delta$  can be interpreted as the difference between cases and controls in their mean risk score and is the standard deviation of  $\log(\text{incidence})$ .

VALID extends the concept of ‘polygenic’ variance in risk<sup>12,13</sup> to all other causes and can be applied to major genes by estimating the proportion of polygenic variance explained after fitting the effects of rare high-risk mutations.<sup>30</sup> VALID allows the familial variance to be due to more than genetic factors alone, for example using [Equation \(9\)](#).

[Table 1](#) allows comparisons of the risk-discriminatory strengths of risk factors, measured and unmeasured. Note we are considering variation in risk for persons of the same age. Therefore, it is inappropriate to compare, for example, AUCs derived from cohorts estimating absolute risk for diseases whose incidence is age-dependent—particularly when this age-dependence is not necessarily universal—with AUCs derived from case-control studies. [Figure 5](#) shows the receiver operating characteristic curve according to the FRR.

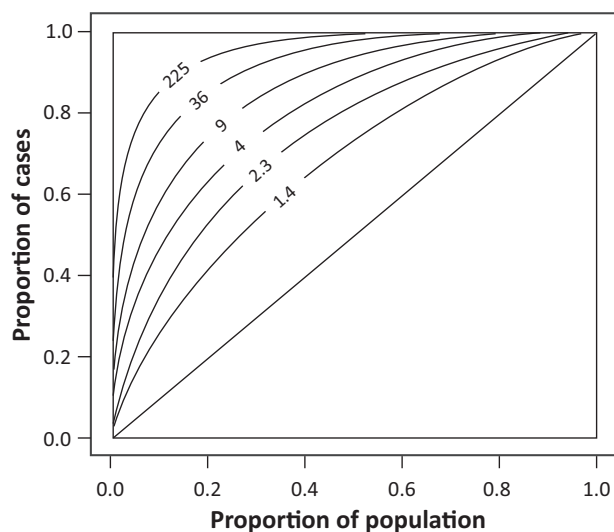
### Why $\log(\text{incidence})$ ?

$\log(\text{incidence})$  is a natural risk scale in epidemiology, and typically is highly dependent on age. The linearity or otherwise of its relationship to  $\log(\text{age})$  has been used to make biologically relevant inference about underlying stages in disease progression with application to common cancers<sup>53</sup>



and about the role of cumulative exposure to ovarian hormones in breast cancer risk.<sup>54</sup>

A major focus of epidemiology is on the causes of differences in log(incidence) between groups of the same age and the estimation of the risk gradients such as relative risk, odds ratio and hazard ratio by applying logistic regression to case-control studies or Cox proportional hazards regression to cohort studies. Variation in log(incidence) is the



**Figure 5** Receiver operating characteristic curves under the VALID model labelled according to the familial risk ratio (FRR), where Proportion of cases is the sensitivity and the Proportion of population is 1—specificity (following Clayton<sup>13</sup>), for area under the receiver operating characteristic curve (AUC) ranging 0.60, 0.73, 0.80, 0.92, 0.97 and 0.99 from lower right to upper left

basis of complex segregation analyses of pedigree data in search of evidence for, and about, major genes.<sup>47</sup> Genome-wide association studies are applying case-control analyses to create additive polygenic risk scores on this scale.<sup>8</sup>

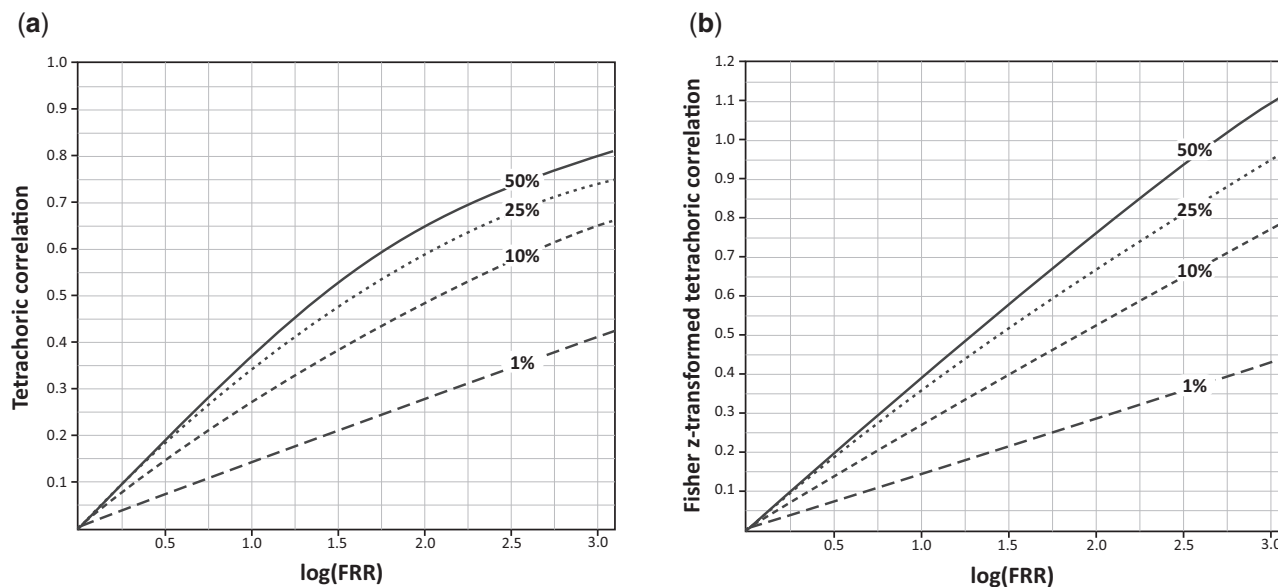
### Generality

We allow the risk score to be measured or unmeasured. Whereas our multiplicative model might not accurately represent reality for every risk factor, as a model for studying combined risk factors by, for example, familial versus non-familial or genetic versus non-genetic, we illustrated how it might be a useful approximation to reality based on empirical evidence, at least for breast cancer. The model also applies to combinations of risk scores.

### Comparison with liability model and heritability

Application of the deterministic liability model to the Nordic Twin Study<sup>28</sup> suggested that the influence of genetic factors on variation in risk, as measured by the tetrachoric correlation and heritability, is relatively stable with age. This is contrary to our findings from applying VALID. We think the discrepancy is explained because the AUC under the liability model is dependent on the disease prevalence as well as the tetrachoric correlation,<sup>55</sup> whereas under the VALID model it depends solely on the FRR.

Figure 6a shows that the relationship between the tetrachoric correlation and log(OPERA) is almost linear for log(OPERA) < 1, but not thereafter, and depends highly on the disease prevalence. Figure 6b shows that on a natural



**Figure 6** Plot of the relationship of: (a) the tetrachoric correlation calculated using the polycor package in R, and (b) the Fisher Z transform of the tetrachoric correlation, against the logarithm of the familial risk ratio (log(FRR<sub>M2</sub>)) under the VALID model for disease frequencies 1%, 10%, 25% and 50%

scale for correlations, the Z-transformed tetrachoric correlation is almost linear with log(OPERA) across unbounded scales, though the slope still depends highly on disease prevalence.

There are two important consequences. First, for diseases with a <2-fold increased risk from having an affected first-degree relative, decomposition of familial risk will be similar whether the liability or VALID models are used. Second, the liability model predicts that the role of genes is greater for more common diseases with the same FRR, and for older persons even if the FRR is independent of age, contrary to the prediction of the VALID model and empirical evidence.

## Summary

In conclusion, we propose thinking about how risk factors explain variation in risk in terms of variance in the logarithm of age-specific incidence. Genetic and non-genetic factors combine to explain greater amounts (not proportions) of variation in risk. VALID describes the finite genetic architecture and unlimited environmental landscape of disease risk using a single metric, enabling causes of risk variation (not causes per se) to be compared and combined.

The maximum variation in risk due to genetic factors is determined by studying MZ twin pairs. Genetic factors will not be important for population risk stratification if the MZ twin pair odds ratio is weak, irrespective of disease frequency. The familial odds ratio is directly related to the absolute familial variance by the Unifying Equation. This harks back to Fisher's 1918 paper<sup>1</sup> where he showed that the major issue for evolution was the magnitude of the genetic variance, not a percentage or proportion of the total variance which he described as a 'hotch-potch' of a denominator.<sup>2</sup> For risk, the denominator is in effect unlimited.

Our application of VALID to female breast cancer revealed that, whereas substantial components of variation in familial risk have been discovered, there remains much to be learned about the familial causes of breast cancer particularly for young women, and little is known about individual-specific variance in risk.

## Ethics approval

This study does not need ethical approval as it uses only data from published analyses.

## Data availability

Not applicable.

## Supplementary data

Supplementary data are available at *IJE* online.

## Author contributions

J.L.H. initiated the VALID and OPERA concepts and how they could be used to decompose familial, genetic and non-genetic variance. J.G.D. proved the relationship between OPERA and the AUC and wrote the Combining Risk Scores section of the [Supplementary Material](#). T.L.N. helped develop the OPERA concept and demonstrated its use for mammographic density research. S.L. demonstrated how the VALID concept applied to major genes and breast cancer. G.S.D. and R.J.M. applied the OPERA concept to familial risk scores and polygenic risk scores. E.M. and D.F.S. applied the OPERA concept to mammogram risk scores. M.B. applied the decomposition of familial risk ratios underlying VALID to childhood asthma and hay fever. J.S. (Australia) introduced the mammographic density research technology to Australia. J.S. (Korea) helped develop the VALID and OPERA concepts through application to Korean mammographic density research studies with T.L.N. supervised by M.A.J., who also applied familial odds ratio decomposition to asthma. G.G.G. helped create the Australian mammographic density research studies. M.C.S. led genomic work fundamental to the application of VALID to breast cancer. J.D.M. developed earlier versions of this analytical approach with J.L.H. in the 1980s.

## Funding

J.L.H. was supported by an NHMRC Fellowship (GMT1137349) and is currently a Dame Kate Campbell Professorial Fellow at the University of Melbourne. T.L.N. is supported by Cancer Council Victoria (AF7305). S.L. is a Victorian Cancer Agency Early Career Research Fellow (ECRF19020). J.S. (Australia) is supported by a National Breast Cancer Foundation Early Career Fellowship (ECF-17-010). J.S. (Korea) is supported by a National Research Foundation of Korea grant funded by the Korean government (MSIT) (No. 2020R1A2C2101041). M.A.J. is supported by a National Health and Medical Research Council Investigator grant (APP1195099). M.C.S. is supported by a National Health and Medical Research Council Fellowship (GNT1155163).

## Acknowledgements

We would like to acknowledge the work of Odd Aalen, David Clayton, Robert Elston and Ronald Fisher, which laid the foundations for the VALID concept.

## Conflict of interest

G.S.D. is employed by Genetic Technologies Limited. The other authors declare no conflict of interest.

## References

1. Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* 1919;52:399–433.
2. Fisher RA. Limits to intensive production in animals. *Br Agric Bull* 1951;4:217–18.
3. Harrap SB, Stebbing M, Hopper JL *et al*. Familial patterns of co-variation for cardiovascular risk factors in adults: The Victorian Family Heart Study. *Am J Epidemiol* 2000;152:704–15.
4. Falconer DS. *Introduction to Quantitative Genetics*. Edinburgh/London: Oliver & Boyd, 1960.

5. Lange K, Westlake J, Spence MA. Extensions to pedigree analysis. III. Variance components by the scoring method. *Ann Hum Genet* 1976;39:485–91.
6. Hopper JL, Mathews JD. Extensions to multivariate normal models for pedigree analysis. *Ann Hum Genet* 1982;46:373–83.
7. Hopper JL, Tait BD, Probert DN, Mathews JD. Genetic analysis of systolic blood pressure in Melbourne families. *Clin Exp Pharmacol Physiol* 1982;9:247–52.
8. Mavaddat N, Michailidou K, Dennis J *et al*. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet* 2019;104:21–34.
9. Hopper JL. Odds per adjusted standard deviation: comparing strengths of associations for risk factors measured on different scales and across diseases and populations. *Am J Epidemiol* 2015;182:863–67.
10. Aalen OO. Modelling the influence of risk factors on familial aggregation of disease. *Biometrics* 1991;47:933–45.
11. Hopper JL, Carlin JB. Familial aggregation of a disease consequent upon correlation between relatives in a risk factor measured on a continuous scale. *Am J Epidemiol* 1992;136:1138–47.
12. Pharoah P, Antoniou A, Bobrow M *et al*. Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* 2002;31:33–36.
13. Clayton DG. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet* 2009;5:e1000540.
14. Wentzensen N, Wacholder S. From differences in means between cases and controls to risk stratification: a business plan for biomarker development. *Cancer Discovery* 2013;3:148–57.
15. Win AK, Dowty JG, Cleary SP *et al*. Risk of colorectal cancer for carriers of mutations in MUTYH, with and without a family history of cancer. *Gastroenterol* 2014;146:1208–11.
16. Nguyen TL, Aung YK, Evans CF *et al*. Mammographic density defined by higher than conventional brightness threshold better predicts breast cancer risk for full-field digital mammograms. *Breast Cancer Res* 2015;17:142.
17. Hopper JL, Nguyen TL, Li S. Blood DNA methylation score predicts breast cancer risk: applying OPERA in molecular, environmental, genetic and analytic epidemiology. *Mol Oncol* 2021;16:8–10.
18. Schmidt DF, Makalic E, Goudey B *et al*. Cirrus: An automated mammography-based measure of breast cancer risk based on textural features. *JNCI Cancer Spectr* 2018;2:pk057.
19. Salgado JF. Transforming the area under the normal curve (AUC) into Cohen's d, Pearson's  $r_{pb}$ , odds-ratio, and natural log odds ratio: two conversion tables. *European J Psychol Applied to Legal Context* 2018;10:35–47.
20. Hopper JL, Mathews JD. Extensions to multivariate normal models for pedigree analysis. II. Modeling the effect of shared environment in the analysis of variation in blood lead levels. *Am J Epidemiol* 1983;117:344–55.
21. Clifford CA, Hopper JL, Fulker DW, Murray RM. A genetic and environmental analysis of a twin family study of alcohol use, anxiety, and depression. *Genet Epidemiol* 1984;1:63–79.
22. Hopper JL, Culross PR. Covariation between family members as a function of cohabitation history. *Behav Genet* 1983;13:459–71.
23. Li S, Wong EM, Dugué PA *et al*. Genome-wide average DNA methylation is determined in utero. *Int J Epidemiol* 2018;47:908–16.
24. Li S, Nguyen TL, Wong EM, Dugué PA *et al*. Genetic and environmental causes of variation in epigenetic aging across the lifespan. *Clin Epigenetics* 2020;12:158.
25. Bonney GE. Compound regressive models for family data. *Hum Hered* 1992;42:28–41.
26. Hopper JL, Jenkins MA, Macaskill GT, Giles GG. Analysis of familial aggregation in a binary trait by logistic regression and the regressive logistic model. In: LaBuda M, Grigorenko (eds). *On the Way to Individuality: Current Methodological Issues in Behavioural Genetics*. New Haven, CT: Yale University Press, 1997, 155–83.
27. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet* 2001;358:1389–99.
28. Möller S, Mucci LA, Harris JR *et al*. Heritability of breast cancer among women in the Nordic twin study of cancer. *Cancer Epidemiol Biomarkers Prev* 2016;25:145–50.
29. 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.
30. Li S, MacInnis RJ, Lee AM *et al*. Segregation analysis of 17,425 population-based breast cancer families: evidence for genetic susceptibility and risk prediction. *Am J Hum Genet* 2022;109:1777–88.
31. Dite GS, MacInnis RJ, Bickerstaffe A *et al*. Breast cancer risk prediction using clinical models and 77 Independent risk-associated SNPs for women aged under 50 Years: Australian Breast Cancer Family Registry. *Cancer Epidemiol Biomarkers Prev* 2016;25:359–65.
32. Hopper JL, Nguyen TL, Schmidt DF *et al*. Going beyond conventional mammographic density to discover novel mammogram-based predictors of breast cancer risk. *J Clin Med* 2020;9:627.
33. Nguyen TL, Aung YK, Evans CF *et al*. Mammographic density defined by higher than conventional brightness thresholds better predicts breast cancer risk. *Int J Epidemiol* 2017;46:652–61.
34. Nguyen TL, Choi YH, Aung YK *et al*. Breast cancer risk associations with digital mammographic density by pixel brightness threshold and mammographic system. *Radiology* 2018;286:433–42.
35. Nguyen TL, Aung YK, Li S *et al*. Predicting interval and screen-detected breast cancers from mammographic density defined by different brightness thresholds. *Breast Cancer Res* 2018;20:152.
36. Nguyen TL, Li S, Dite GS *et al*. Interval breast cancer risk associations with breast density, family history and breast tissue aging. *Int J Cancer* 2020;147:375–82.
37. Nguyen TL, Schmidt DF, Makalic E *et al*. Novel mammogram-based measures improve breast cancer risk prediction beyond an established mammographic density measure. *Int J Cancer* 2021;148:2193–202.
38. Watt GP, Knight JA, Nguyen TL *et al*. Association of contralateral breast cancer risk with mammographic density defined at

- higher-than-conventional intensity thresholds. *Int J Cancer* 2022;**151**:1304–49.
39. Tan M, Mariapun S, Yip CH *et al.* A novel method of determining breast cancer risk using parenchymal textural analysis of mammography images on an Asian cohort. *Phys Med Biol* 2019;**64**:035016.
  40. Pertuz S, Sassi A, Holli-Helenius K *et al.* Clinical evaluation of a fully-automated parenchymal analysis software for breast cancer risk assessment: a pilot study in a Finnish sample. *Eur J Radiol* 2019;**121**:108710.
  41. Boyd NF, Guo H, Martin LJ *et al.* Mammographic density and the risk and detection of breast cancer. *N Engl J Med* 2007;**356**: 227–36.
  42. Nguyen TL, Schmidt DF, Makalic E *et al.* Explaining variance in the Cumulus mammographic measures that predict breast cancer risk: a twins and sisters study. *Cancer Epidemiol Biomarkers Prev* 2013;**22**:2395–403.
  43. Kresovich JK, Xu Z, O'Brien KM *et al.* Blood DNA methylation profiles improve breast cancer prediction. *Mol Oncol* 2022;**16**: 42–53.
  44. Hopper JL, Dite GS, MacInnis RJ *et al.*; kConFab Investigators. Age-specific breast cancer risk by body mass index and familial risk: prospective family study cohort (ProF-SC). *Breast Cancer Res* 2018;**20**:132.
  45. Ye Z, Dite GS, Nguyen TL *et al.* Weight is more informative than body mass index for predicting post-menopausal breast cancer risk: Prospective Family Study Cohort (ProF-SC). *Cancer Prev Res (Phila)* 2022;**15**:185–91.
  46. Hopper JL. Disease-specific prospective family study cohorts enriched for familial risk. *Epidemiol Perspect Innov* 2011;**8**:2.
  47. Lee A, Mavaddat N, Wilcox AN *et al.* BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet Med* 2019;**21**:1708–18.
  48. Boyd NF, Dite GS, Stone J *et al.* Heritability of mammographic density, a risk factor for breast cancer. *N Engl J Med* 2002;**347**: 886–94.
  49. Martin LJ, Melnichouk O, Guo H *et al.* Family history, mammographic density, and risk of breast cancer. *Cancer Epidemiol Biomarkers Prev* 2010;**19**:456–63.
  50. Chen H, Fan S, Stone J *et al.*; NBCS Investigators. Genome-wide and transcriptome-wide association studies of mammographic density phenotypes reveal novel loci. *Breast Cancer Res* 2022;**24**:27.
  51. Li S, Nguyen TL, Nguyen-Dumont T *et al.* Genetic aspects of mammographic density measures associated with breast cancer risk. *Cancers* 2022;**14**:2767.
  52. Nguyen TL, Li S, Dowty JG *et al.* Familial aspects of mammographic density measures associated with breast cancer risk. *Cancers* 2022;**14**:1483.
  53. Armitage P, Doll R. The age distribution of cancer and a multistage theory of carcinogenesis. *Br J Cancer* 1954;**8**:1–12.
  54. Pike MC, Krailo MD, Henderson BE *et al.* 'Hormonal' risk factors, 'breast tissue age' and the age-incidence of breast cancer. *Nature* 1983;**303**:767–70.
  55. Wray N, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet* 2010;**6**:e1000864.