THE IMPACT OF MULTIFACTORIAL GENETIC DISORDERS ON CRITICAL ILLNESS INSURANCE: A SIMULATION STUDY BASED ON UK BIOBANK

BY

ANGUS MACDONALD, DELME PRITCHARD AND PRADIP TAPADAR

ABSTRACT

The UK Biobank project is a proposed large-scale investigation of the combined effects of genotype and environmental exposures on the risk of common diseases. It is intended to recruit 500,000 subjects aged 40-69, to obtain medical histories and blood samples at outset, and to follow them up for at least 10 years. This will have a major impact on our knowledge of multifactorial genetic disorders, rather than the rare but severe single-gene disorders that have been studied to date. What use may insurance companies make of this knowledge, particularly if genetic tests can identify persons at different risk? We describe here a simulation study of the UK Biobank project. We specify a simple hypothetical model of genetic and environmental influences on the risk of heart attack. A single simulation of UK Biobank consists of 500,000 life histories over 10 years; we suppose that case-control studies are carried out to estimate age-specific odds ratios, and that an actuary uses these odds ratios to parameterise a model of critical illness insurance. From a large number of such simulations we obtain sampling distributions of premium rates in different strata defined by genotype and environmental exposure. We conclude that the ability of such a study reliably to discriminate between different underwriting classes is limited, and depends on large numbers of cases being analysed.

KEYWORDS

Case-control Study, Critical Illness Insurance, Gene-environment Interaction, Odds Ratio, Premium Rating, Simulation, UK Biobank.

1. INTRODUCTION

1.1. Objective

Much of human genetics is concerned with studying the genetic contribution to diseases, and this leads to a profound distinction between the single-gene disorders and the multifactorial disorders.

Astin Bulletin 36(2), 311-346. doi: 10.2143/AST.36.2.2017924

(a) Single-gene disorders are caused, as their name suggests, by a defect in a single gene. Because most genes are inherited in a simple way according to Mendel's laws, these diseases show characteristic patterns of inheritance from one generation to the next, known to geneticists and underwriters alike as a 'family history'. Single-gene disorders are quite rare but often severe.

(b) Multifactorial disorders are (mostly) common diseases, such as coronary heart disease and cancers, whose onset or progression may be influenced by variations in several genes, acting in concert with environmental differences. The effect is likely to be quite slight, conferring an altered predisposition to the disease rather than a radically different risk.

Most genetic epidemiology has, until now, concentrated on single-gene disorders. One reason is that the clear patterns of Mendelian inheritance identified affected families long before molecular genetics came along. When these tools emerged in the 1990s, geneticists knew where to look; affected families were studied, genes were identified, and the key epidemiological parameters were estimated. The parameter of most interest to actuaries is the age-related *penetrance*, which is the probability that a person who carries a risky version of the gene will have suffered onset of the disease by age x. It is entirely analogous to the life table probability ${}_xq_0$. (Often, the risky versions of the gene are called 'mutations', and a person carrying one is called a 'mutation carrier' or just 'carrier'.)

Studies of affected families are by definition retrospective; families are studied *because* they are known to be affected. This introduces uncontrolled sources of bias, so such studies are, if possible, avoided in favour of prospective studies, in which a properly randomised sample of healthy subjects is followed forwards in time. Despite this health warning, retrospective studies of single-gene disorders have been carried out for reasons of convenience, cost and necessity: the ready availability of known affected families was convenient and made data collection relatively cheap; and the rarity of single-gene disorders made prospective studies impractical. Moreover, a prospective study would take many years to yield results. Another consequence of the rarity of most single-gene disorders is that most studies have had quite small sample sizes, but if the penetrance is high enough this is tolerable. These studies have successfully led to many gene discoveries and a lot of progress has been made in understanding single-gene disorders.

Multifactorial disorders are not so well-studied, and are much harder to study. The clear patterns of Mendelian inheritance are lost, and any familial clustering of disease that may be observed could just as easily be the result of shared environment as of shared genes. Therefore, there is no pool of known affected families that can be studied straightaway. And, because the influence of genetic variation may be slight (low penetrance) large samples will be needed to detect such influence with any reliability.

At the risk of oversimplifying a little, single-gene disorders represent the genetical research of the past, and multifactorial disorders represent the genetical

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

research of the future. Progress will need studies that are large-scale, prospective, and long-term (therefore very expensive) and that capture both genetic and environmental variation and the incidence of common diseases. This is very ambitious.

The proposed UK Biobank project aims to achieve this. UK Biobank will recruit 500,000 individuals aged 40 to 69, chosen as randomly as possible from the UK population, and collect data on them over 10 years. We will discuss its main features in Section 1.2. A key point is that UK Biobank aims only to collect data, not to analyse it. Its data will, in due course, be made available to researchers interested in particular genes and particular diseases, who will have to obtain separate funding for their studies. This is sensible because it is impossible to predict at outset just what combinations of genes, environment and disease it will be most fruitful to study. Nevertheless, it is necessary to have in mind the kinds of statistical studies most likely to be carried out, so that UK Biobank can be set up to capture data of the correct form. The presumption is that most studies will be *case-control* studies. We outline these in the Appendix.

Given its size and significance, it is important to study the kind of results we might expect to emerge out of UK Biobank. Our particular interest is in the implications of UK Biobank for insurance. We need not rehearse the debate, often heated, that has surrounded genetics and insurance in the past 10 years, except to note that it has mainly focussed on single-gene disorders. Daykin *et al.* (2003) or Macdonald (2004) are sources. It seems plausible that awareness of genetic issues will be heightened by enrolling 500,000 people into a genetic study. If insurance questions arise, answers obtained from past actuarial research into single-gene disorders may be wholly inapplicable. But, since the single-gene disorders provide all the easily grasped examples and paradigms, there is a risk that these examples and paradigms will be grafted onto UK Biobank, however inappropriately, by the media if not by the genetics community. It will then be unfortunate that UK Biobank will not provide the evidence to refute such errors for 5-10 years.

Our plan, therefore, is to model UK Biobank itself, so that before a single person has been recruited, or gene sequenced, we may quantify the implications of its outcomes for insurance. We choose critical illness (CI) insurance as the simplest type of coverage, because the insured event is generally disease onset. We choose heart attack (myocardial infarction) as the disease of interest, because this will certainly be a major target of studies using UK Biobank data. Our approach is simple: simulate 500,000 random life histories, given an assumed model of genetic and environmental influences on the hazard rate of heart attack. Then we may analyse these simulated data just as an epidemiologist or an actuary may be expected to.

At this stage a further complication appears, very familiar to actuarial researchers who have modelled single-gene disorders. Actuaries almost never have access to the original data upon which genetic studies are based. Section 5.2 of the UK Biobank draft protocol (www.ukbiobank.ac.uk/docs/draft/protocol.pdf) says: "Data from the project will not be accessible to the insurance industry

or any other similar body." This means that actuarial researchers will have to rely on the published outcomes of medical or epidemiological research projects that use the UK Biobank data, in particular case-control studies. The ideal, given the models actuaries typically use for pricing and reserving, would be age-dependent onset rates or penetrances, corresponding to μ_x or q_x in a life table. Unfortunately, this far exceeds what is usually published, because the questions asked in a medical study can often be answered by much simpler statistics. And, it must be said, the estimation of μ_x or q_x is very demanding of the data. So we may *not*, realistically, assume that the actuary can analyse directly the 500,000 simulated life histories. Instead, an epidemiologist must first carry out a case-control study and publish the results, probably in the form of odds ratios (see the Appendix). Then the actuary must take these odds ratios and, using whatever approximate methods come to hand, estimate onset rates or penetrances suitable for use in an actuarial model. We will model this process, with two results:

(a) We will be able to estimate the impact on CI insurance premiums of representative multifactorial modifiers of heart attack risk.

(b) Having simulated the data from a *known* model of our own choosing, we can assess the seriousness of the errors that must be made, in parameterising an actuarial model from published odds ratios rather than from the raw data. As mentioned before, previous actuarial studies have done exactly that (see Macdonald & Pritchard (2000) for an example), but only in the context of relatively high penetrances. We will be interested to see if robust actuarial modelling of relatively low-penetrance disorders is possible using published case-control studies.

The plan of the paper is as follows. In the remainder of this section we describe the main features of UK Biobank and our general approach. A model representing heart attack will be introduced and parameterised in Section 2, including a simple hypothetical 2×2 gene-environment interaction model affecting heart attack risk.

In Section 3, we present (in summary form) and analyse a set of simulated UK Biobank data, namely 500,000 life histories. A model epidemiologist carries out a case-control study, then our model actuary uses these 'published' figures to find critical illness premium rates allowing for genetic variability and environmental exposures.

Despite its great size, UK Biobank is essentially an unrepeatable single sample. Any estimated quantity based upon its data is subject to the usual sampling error — and a premium rate is just such an estimated quantity. We can assess directly the sampling properties of estimates based on UK Biobank data, simply by repeating the simulation of 500,000 life histories as many times as necessary, and constructing the empirical distributions of the odds ratios and premium rates. This is in Section 4. This is directly relevant to the criteria established in the UK by the Genetics and Insurance Committee (GAIC) for assessing the reliability of premium rates based on genetic information.

Conclusions and suggestions for further work are in Section 5.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

1.2. The UK Biobank Project

The website http://www.ukbiobank.ac.uk/ is the main source of information on UK Biobank. In particular, it provides a draft protocol, which states (Section 1.2) that:

> "The main aim of the study is to collect data to enable the investigation of the separate and combined effects of genetic and environmental factors (including lifestyle, physiological and environmental exposures) on the risk of common multifactorial disorders of adult life."

UK Biobank is a cohort study, meaning that a large number of people will be recruited, as randomly as possible, and then followed over time. The main features of the study design are as follows:

(a) The cohort will consist of at least 500,000 men and women recruited from the UK general population.
(b) The chosen age range is 40 to 69 (note that earlier versions, including the draft protocol referred to above, proposed an age range 45 to 69).
(c) The initial follow-up period is 10 years.
(d) Participants will be recruited through their local general practitioners. Participants are expected to come from a broad range of socio-economic backgrounds and regions throughout the UK, with a wide range of exposures to factors of interest.
(e) The project will be conducted through the UK National Health Service.
(f) UK Biobank is funded by the Department of Health, the Medical Research Council, the Scottish Executive and The Wellcome Trust, and will cost approximately £40 million.

People registered with participating general practices will be requested to join the study by completing a self-administered questionnaire, attending an interview, undergoing examination by a research nurse and giving a blood sample, to enable DNA extraction at a later date, as and when genotyping is required.

The Office of National Statistics will provide routine follow-up data regarding cause-specific mortality and cancer incidence. Hospitalisation and general practice records will provide data regarding incident morbidity. Every two years a subset of 2,000 participants and every five years the entire cohort will be re-surveyed by postal questionnaire to update exposure data and to ascertain self-reported incident morbidity.

It is envisaged that the main study design for later analysis will be a case-control study (see the Appendix) nested within the cohort. UK Biobank will only collect and store the data, its analysis will require further funding.

1.3. A UK Biobank Simulation Model

In this section we outline how we will simulate the UK Biobank project.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

We suppose that the study population is subdivided (or stratified) into subgroups with respect to: (a) genotype; (b) level of environmental exposure; and (c) other relevant factors such as sex. Genotype defines discrete categories, and we suppose that environmental exposures or other factors defined on a continuous scale are grouped into discrete categories. Thus, we always have a small number of discrete subgroups (or strata).

The life history of each participant, including the occurrence of a heart attack, will be represented by the multiple-state model shown in Figure 1. It is parameterised by intensities denoted $\lambda_{ij}^{s}(x)$ or $\lambda_{ij}^{s}(x,t)$, functions of age x and possibly also duration t since entering state i. The superscript 's' indicates stratum, and the intensities representing heart attacks will be stratum-dependent. These intensities are the key to the whole UK Biobank project, as well as our study.

(a) The real-life epidemiologist wants to estimate them (or in practice, odds ratios) from UK Biobank data, given a hypothesis about the effect of measured exposures on the disease.

(b) The real-life actuary wants to take the estimated intensities (or in practice, approximate them from published odds ratios) and use them in pricing and reserving.

(c) We want to specify hypothetical but plausible dependencies of these intensities on genotype and other exposures, so that we can observe our model epidemiologist and model actuary at work.

1.4. Simulating UK Biobank

The steps in simulating UK Biobank are then as follows.

(a) We choose the number of genotypes and the number of levels of environmental exposure, and also the frequencies with which each appears in the population. Thus we can model simple or complex genotypes and environmental exposures, and allow them to be more or less common or rare. These define the subgroups or strata. The simplest example (used in the UK Biobank draft protocol) is to have two genotypes and two levels of environmental exposure. We also choose the intensities of onset of heart attack in each stratum ($\lambda_{12}^{s}(x)$ in Figure 1).

(b) We randomly 'create' 500,000 individuals, each equally likely to be male or female, and with ages uniformly distributed in the range 40 to 70, and allocated to strata at random according to the chosen frequencies.

(c) The life history of each individual is modelled by simulating the times of any transitions between states in the model, as governed by the intensities. We record the times of any transitions taking place within the 10-year follow-up period of UK Biobank.

We assume that the 500,000 participants are independent in the statistical sense, which is unlikely to be true. The sample is so large that some related individuals

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

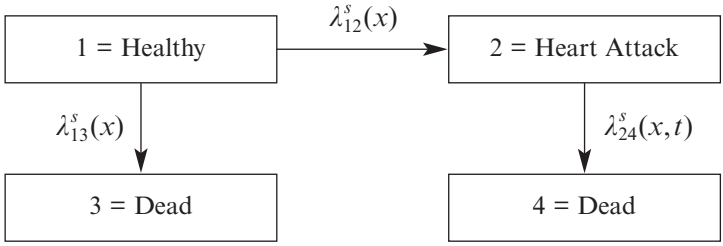

FIGURE 1: A 4-state heart attack model for stratum *s*.

are likely to be recruited by chance, but also the method of recruitment (through general practices) guarantees some level of familial and geographical clustering.

2. A Model for Heart Attack

2.1. Specification of the Model

In this section we will parameterise the heart attack model of Figure 1. Everyone is assumed to start in the Healthy state. We are interested in first heart attacks only, because this will trigger a claim under a CI policy, so any subsequent heart attacks are ignored, and the only exit from the Heart Attack state is death. It is convenient to distinguish deaths occurring after a heart attack, so states 3 and 4 are separate.

2.2. The Population Heart Attack Transition Intensity

Let $\lambda_{12}(x)$ denote the heart attack transition intensity in the general population, separately for males and females. We take $\lambda_{12}(x)$ from Gutiérrez & Macdonald (2003). For males, it is given by:

$$\lambda_{12}(x) = \begin{cases} \exp(-13.2238 + 0.152568x) & \text{if } x \leq 44 \\ \dfrac{x-44}{49-44} \times \lambda_{12}(49) + \dfrac{49-x}{49-44} \times \lambda_{12}(44) & \text{if } 44 < x < 49 \\ -0.01245109 + 0.000315605x & \text{if } x \geq 49 \end{cases} \quad (1)$$

and for females, it is given by:

$$\lambda_{12}(x) = \frac{0.598694}{\Gamma(15.6412)} \times 0.15317^{15.6412} \exp(-0.15317x)\, x^{14.6412}. \quad (2)$$

These intensities are shown in Figure 2.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

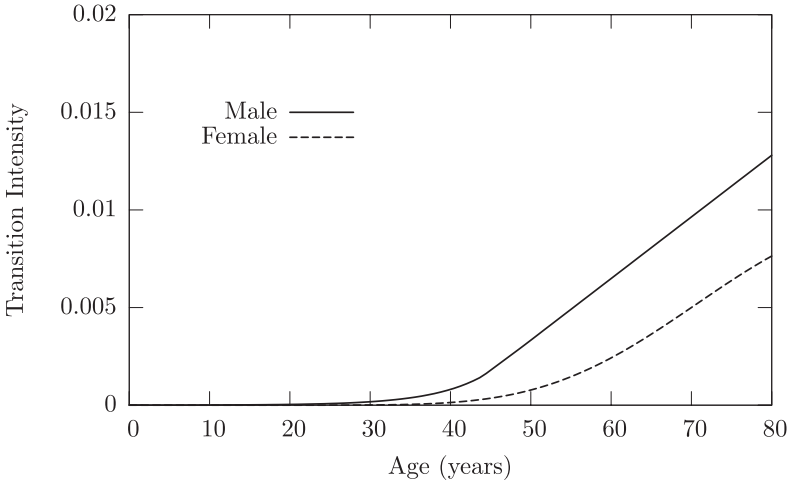

FIGURE 2: The transition intensity of all first heart attacks, by gender.

2.3. Mortality After First Heart Attacks

Many journal articles describe prognosis following heart attacks. Capewell *et al*. (2000) describe a retrospective cohort study in Scotland involving 117,718 patients admitted to hospital with first heart attacks between 1986 and 1995. This is one of the largest population-based studies describing both short and long-term prognoses.

The paper presents case-fatality rates for age-groups (at first heart attack) <55, 55-64, 65-74, 75-84 and ≥ 85, and for durations 30 days, 1 year, 5 years and 10 years following first heart attack. The age-adjusted case-fatality rates did not depend on sex. The age-specific case-fatality rates can be transformed into survival rates and parametric functions can be fitted to these. The following parametric form fits the survival functions well:

$$P_{22}^{\alpha}(t) = \frac{1}{1 + a \times t^{b} + c \times t^{d}} \tag{3}$$

where α denotes the age group (see above), t denotes the duration after the first heart attack and $P_{22}^{\alpha}(t)$ denotes the conditional probability that the individual is still in State 2 t years after the first heart attack. The parameters a, b, c and d depend on the age group. We will represent the five age groups by single representative ages, namely, 50, 60, 70, 80 and 90. We summarise the parameters in Table 1.

From the parametric form of the survival rates, the transition intensities are given as $\lambda_{24}^{\alpha}(t) = -d(\log P_{22}^{\alpha}(t))/dt$. Graphs of $\lambda_{24}^{\alpha}(t)$ are given in Figure 3, assigning each to its representative age. Also shown is the force of mortality

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

TABLE 1.

PARAMETER ESTIMATES OF THE SURVIVAL FUNCTION AFTER A FIRST HEART ATTACK.

Age Range	Representative Age	a	b	c	d
<55	50	0.0684	0.1040	0.0174	1.1919
55-64	60	0.1686	0.0911	0.0406	1.2280
65-74	70	0.4001	0.1237	0.0770	1.3370
75-84	80	0.8564	0.1732	0.1476	1.5504
≥85	90	1.5181	0.2431	0.3309	1.6727

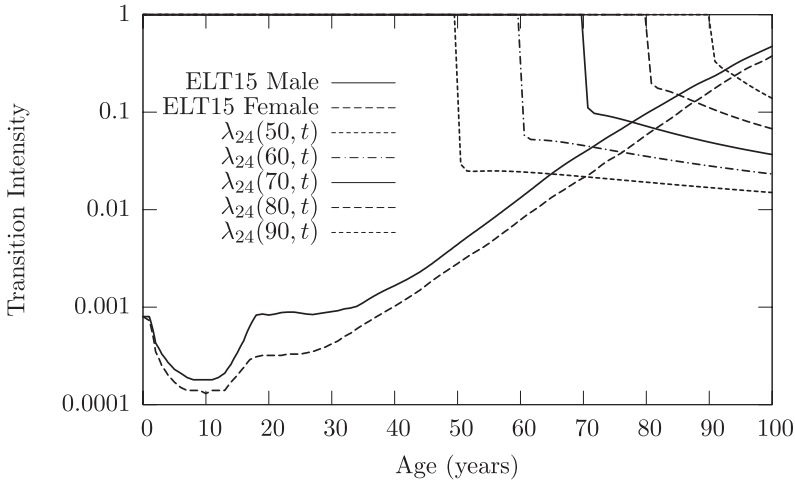

FIGURE 3: Graphs of $\lambda^{\alpha}_{24}(t)$, assigned to representative ages for each age group, and the force of mortality of the ELT15 life tables.

of the ELT15 life tables for males and females. Note that in some cases $\lambda_{24}(x,t)$ falls below the ELT15 force of mortality. This could be because survival beyond a certain duration after a first heart attack signifies better than average overall health thereafter.

To extend the definition of the transition intensity to all ages x and durations $0 \le t \le 10$, we first suppose that survival rates are the same for all strata or subgroups, so we write $\lambda_{24}(x,t)$ instead of $\lambda^{s}_{24}(x,t)$. Then we assign each $\lambda^{\alpha}_{24}(t)$ to its representative age, so $\lambda_{24}(50,t) = \lambda^{(<55)}_{24}(t)$ for all t, and so on. Then define $\lambda_{24}(x,t) = \lambda_{24}(50,t)$ for $x < 50$, $\lambda_{24}(x,t) = \lambda_{24}(90,t)$ for $x > 90$, and interpolate linearly in x between the given values for $50 < x < 90$. Capewell *et al.* (2000) do not give survival rates more than 10 years after the first heart attack, but since the follow-up period of UK Biobank is 10 years this does not matter.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

2.4. Mortality Before First Heart Attacks

The mortality intensity for persons aged x in stratum s, who do not experience a heart attack, is given by $\lambda_{13}^{s}(x)$. Again, we assume that this is the same in all strata, so we just write $\lambda_{13}(x)$. Let $P_{ij}(y,z)$ denote the conditional probability that a person is in state j at age z, given that he or she was in state i at age y. Then we have:

$$P_{13}(0,x) + P_{14}(0,x) = \int_0^x \left[P_{11}(0,z)\,\lambda_{13}(z) + \int_0^z P_{11}(0,y)\,\lambda_{12}(y)\,P_{22}(y,z)\,\lambda_{24}(y,z-y)\,dy \right] dz \tag{4}$$

$$P_{11}(0,z) = \exp\left[-\int_0^z \left(\lambda_{12}(y) + \lambda_{13}(y)\right) dy \right] \tag{5}$$

$$P_{22}(y,z) = \exp\left[-\int_0^{z-y} \lambda_{24}(y,z-y)\,dy \right]. \tag{6}$$

Further, if we assume that the overall mortality is given by the ELT15 table (for each sex) we have:

$$P_{13}(0,x) + P_{14}(0,x) = 1 - \exp\left[-\int_0^x \mu_y^{ELT}\,dy \right]. \tag{7}$$

Using these, we can solve Equation (4) numerically to obtain $\lambda_{13}(x)$. The transition intensities are given in Figure 4. For comparison, we have included the forces of mortality of the ELT15 tables.

2.5. Definition of Strata: A Simple Example

The parameters of the heart attack model estimated above are supposed to apply to the general population. However, the general population is divided into strata according to genotype, environmental exposures and other factors, and we suppose that the intensity of heart attack $\lambda_{12}^{s}(x)$ depends on the stratum s.

In this section, we will introduce the simplest possible gene-environment interactions into our model. We suppose that there is a single genetic locus with two genotypes, denoted G and g. Also, there are just two levels of environmental exposures, denoted E and e (an example might be E = 'smoker' and e = 'non-smoker'). This simple model can be used as a stepping stone to study higher-dimensional multifactorial models. Note that the UK Biobank draft protocol used the same assumptions in its examples, despite the fact that the project aims to study complex multifactorial disorders. We will suppose that G and E are adverse exposures, while g and e are beneficial. Therefore, we have four strata for each sex — ge, gE, Ge and GE — and eight in total.

We must choose plausible values for the frequencies with which each stratum is present in the population, and the stratum-specific heart attack intensities.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

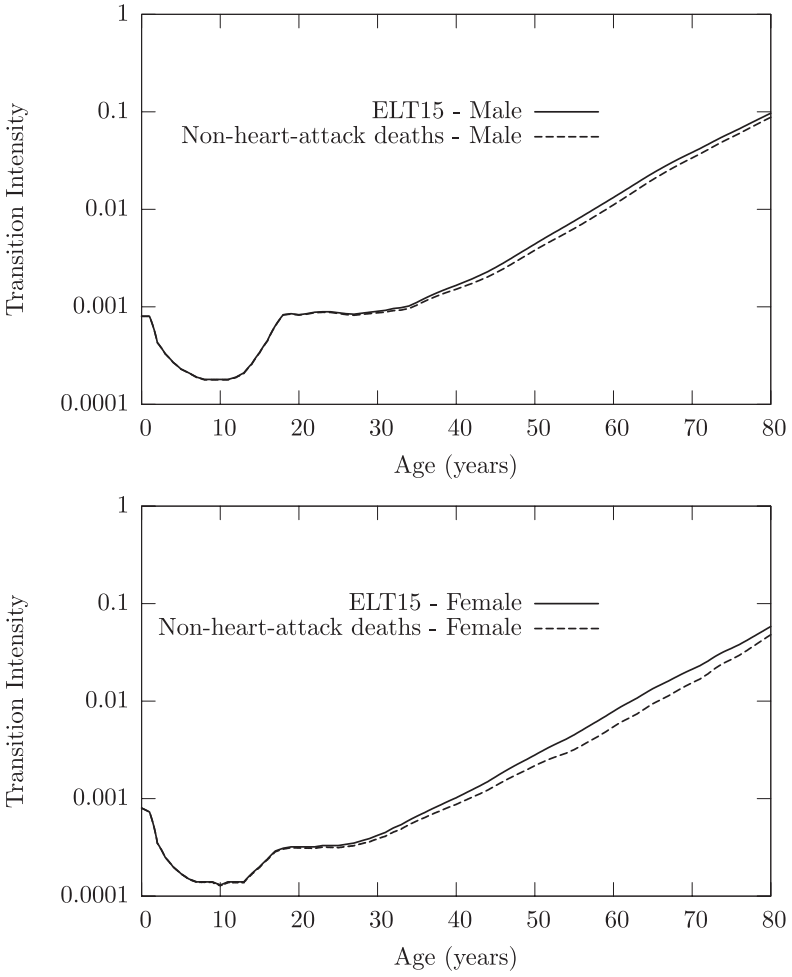

FIGURE 4: Transition intensities of non-heart-attack deaths plotted along with ELT15 for both males and females.

Since, unlike the study of single-gene disorders, we are considering common risk factors for common diseases, let us assume that the probability that a person possesses genotype G is 0.1, and the probability that a person has environmental exposure E is also 0.1. Assuming independence, the four strata (for each sex) ge, gE, Ge and GE occur with frequencies 0.81, 0.09, 0.09 and 0.01 respectively. Strictly speaking, these frequencies ought to be defined at a specific age (age 40 would be an obvious choice) and slowly change thereafter in the population of healthy persons, as higher-risk strata are depleted faster. However, given the relatively small differences we will assume, in the risk of heart attack in respect of different strata, the effect is negligible.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

TABLE 2

THE FACTOR ρ_s, IN EQUATION (8), FOR EACH GENE-ENVIRONMENT COMBINATION.

	G	g
E	1.3	0.9
e	1.1	0.7

We will suppose that the heart attack intensity in each stratum is proportional to the population average intensity. For stratum s, set:

$$\lambda_{12}^{s}(x) = k \times \rho_s \times \lambda_{12}(x) \tag{8}$$

where $\lambda_{12}(x)$ is the population intensity given in Section 2.2. We suppose, for clarity, that ρ_s does not depend on sex, but the constant k does. Noting that our interest is in genotypes of modest penetrance, we choose the values of ρ_s given in Table 2. Then, we choose k so that the strata-specific heart attack intensities are consistent, in aggregate, with the population heart attack intensities, for males and females separately. Let the proportion of the healthy population in stratum s at age x be $w_s(x)$. Then:

$$\lambda_{12}(x+t) = \frac{\sum_s w_s(x) \times \exp\left(-\int_0^t \lambda_{12}^{s}(x+y) + \lambda_{13}(x+y)\,dy\right) \times \lambda_{12}^{s}(x+t)}{\sum_s w_s(x) \times \exp\left(-\int_0^t \lambda_{12}^{s}(x+y) + \lambda_{13}(x+y)\,dy\right)} \tag{9}$$

$$= \frac{\sum_s w_s(x) \times \exp\left(-\int_0^t \lambda_{12}^{s}(x+y)\,dy\right) \times \lambda_{12}^{s}(x+t)}{\sum_s w_s(x) \times \exp\left(-\int_0^t \lambda_{12}^{s}(x+y)\,dy\right)}. \tag{10}$$

Substituting Equation (8) in Equation (10), we get:

$$\lambda_{12}(x+t) = \frac{\sum_s w_s(x) \times \exp\left(-\int_0^t \lambda_{12}(x+y)\,dy\right)^{k\rho_s} \times k \times \rho_s \times \lambda_{12}(x+t)}{\sum_s w_s(x) \times \exp\left(-\int_0^t \lambda_{12}(x+y)\,dy\right)^{k\rho_s}}. \tag{11}$$

From Equation (11) we see that k ought to depend on a specific choice of age x and duration t. However, to keep the model simple we will assume that k is constant and calculate it from Equation (11) for a representative choice of age and duration. Given that the UK Biobank protocol proposes an age range of 40 to 69 and a follow-up period of 10 years, we have chosen $x = 60$ and $t = 5$. If we assume that the weights $w_s(x)$ are equal to the population frequencies of each stratum, then for males $k = 1.317274$ and for females $k = 1.316406$. The

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

TABLE 3

THE MULTIPLIERS $k \times \rho_s$ FOR EACH STRATUM.

Stratum	*ge*	*gE*	*Ge*	*GE*
Male	0.922	1.186	1.449	1.712
Female	0.921	1.185	1.448	1.711

TABLE 4

THE TRUE RELATIVE RISKS FOR EACH STRATUM, RELATIVE TO THE BASELINE *ge* STRATUM.

Stratum	*ge*	*gE*	*Ge*	*GE*
Male	1.000	1.286	1.571	1.857
Female	1.000	1.286	1.571	1.857

constants of proportionality ($k \times \rho_s$) in Equation (8) are given in Table 3 for future reference.

We can now calculate the true values of the quantities likely to be estimated by epidemiologists, namely relative risks and odds ratios (see the Appendix, or Woodward (1999) or Breslow & Day (1980)). From now on, we define the *baseline* population to be the most common stratum, namely *ge*.

(a) The relative risk in stratum s, with respect to stratum *ge*, is denoted r_s and is:

$$r_s = \frac{k \times \rho_s}{k \times \rho_{ge}} = \frac{\rho_s}{\rho_{ge}}. \tag{12}$$

The values of r_s are given in Table 4

(b) The odds ratio at age x in stratum s, with respect to stratum *ge*, based on 1-year probabilities, is denoted $\psi_s(x)$ and is given by:

$$\psi_s(x) = \left(\frac{P_{12}^s(x, x+1)}{1 - P_{12}^s(x, x+1)} \right) \Bigg/ \left(\frac{P_{12}^{ge}(x, x+1)}{1 - P_{12}^{ge}(x, x+1)} \right) \tag{13}$$

where $P_{12}^s(x, x+1)$ is the conditional probability that a person in stratum s who was healthy at age x will suffer a heart attack before age $x+1$.

We have verified (not shown here) that the odds ratios do not vary significantly with age and are approximately equal to the corresponding relative risks. The latter is not surprising, as we have used 1-year probabilities to calculate the odds ratios.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

TABLE 5

THE SIMULATED LIFE HISTORIES OF THE FIRST 20 (OF 500,000) INDIVIDUALS SHOWING THEIR GENDERS, EXPOSURE TO ENVIRONMENTAL FACTORS, GENOTYPES AND THE TIMES AND TYPES OF ALL TRANSITIONS MADE WITHIN 10 YEARS.

ID	*Sex*	*E/e*	*G/g*	*Age*	*State*	*Age*	*State*	*Age*	*State*
1	M	e	g	41.10	1				
2	M	e	G	58.74	1	63.89	2	63.94	4
3	M	e	g	52.27	1				
4	M	e	g	68.39	1				
5	F	e	G	60.94	1	63.81	2		
6	M	e	g	62.49	1	68.18	3		
7	M	e	g	55.50	1	61.57	3		
8	F	e	G	58.95	1				
9	M	e	g	65.67	1	69.58	3		
10	M	e	g	49.79	1				
11	F	E	g	45.43	1				
12	F	e	g	57.58	1				
13	F	e	g	59.68	1				
14	F	E	g	55.14	1				
15	F	e	g	42.93	1				
16	M	e	g	56.23	1				
17	F	e	g	62.84	1				
18	M	e	g	62.29	1				
19	F	e	g	43.69	1				
20	M	e	g	45.16	1				

TABLE 6

NUMBER OF INDIVIDUALS IN EACH STATE AT THE END OF THE 10-YEAR FOLLOW-UP PERIOD.

Sex	*G/g*	*E/e*	*State 1*	*State 2*	*State 3*	*State 4*	*Total*
	G	E	1,871	126	356	115	2,468
	G	e	17,579	928	3,219	934	22,660
Male	g	E	17,588	775	3,236	702	22,301
	g	e	162,474	5,426	29,610	5,002	202,512
	G	E	2,178	70	214	52	2,514
	G	e	19,746	397	2,021	408	22,572
Female	g	E	19,811	367	2,095	330	22,603
	g	e	178,718	2,320	18,891	2,441	202,370
Total			419,965	10,409	59,642	9,984	500,000

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

3. Analysis

3.1. A Sample Realisation of UK Biobank

With the parameterised model, we simulated the life histories of 500,000 people recruited to UK Biobank and followed up for 10 years. Their ages at entry are uniformly distributed between 40 and 70. This is a much simplified representation of the true UK Biobank sampling protocol (www.ukbiobank.ac.uk/docs/draft_protocol.pdf, Section 2.3). In principle sampling should be without replacement from the UK population at these ages, whereas we effectively sample with replacement. In practice recruitment will be *via* participating medical practices, and there may be attempts (not defined very precisely) to select these so as to obtain a more uniform sample of different ages. Our simple assumption should adequately represent the UK Biobank sample; we doubt it would be worthwhile to go further in trying to reproduce it.

3.2. Epidemiological Analysis

The life histories of the first 20 people are shown in Table 5. Consider person No. 2. He is a male with the adverse allele G, exposed to the beneficial environment e. He entered the study healthy (State 1) at age 58.74. He had a heart attack (moved to State 2) at age 63.89 and died (moved to State 4) at age 63.94. The numbers of people in each state at the end of the 10-year follow-up period are given in Table 6.

Apart from the 500,000 life histories, the following information is available to the epidemiologist to carry out a matched case-control study:

(a) the framework of the UK Biobank project;
(b) the structure of the 4-state heart attack model given in Section 2.1;
(c) the transition intensities given in Sections 2.2 to 2.4;
(d) the stratum to which each person is allocated; and
(e) the proportion $w_s(x)$ of individuals in each stratum at a particular age x, say 60.

The first step is to define the cases and controls. Here, clearly, the cases are persons who had first heart attacks during the study period.

In real studies, epidemiologists will face problems such as missing data and cost constraints, and in most circumstances they will use only a subset of all cases for their analysis. Here, we have no such problems, unless we choose to model them. So, in the first instance, we will include all cases in the analysis. Later, we will consider the more realistic possibility that a subset of all cases is used.

An appropriate matching strategy is particularly important for a matched case-control study. Firstly, we match controls with cases by age. Suppose, for example, that we are comparing stratum s with the baseline stratum ge, and that

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

TABLE 7

ODDS RATIOS WITH RESPECT TO THE *ge* STRATUM AS BASELINE, BASED ON A 1:5 MATCHING STRATEGY USING ALL CASES AND 5-YEAR AGE GROUPS. APPROXIMATE 95% CONFIDENCE INTERVALS ARE SHOWN IN BRACKETS. THERE WERE NO CASES AMONG FEMALES AGE 45-49 IN STRATUM *GE*.

MALES			
Age	*gE*	*Ge*	*GE*
40-44	1.043 (0.527,2.065)	2.628 (1.561,4.423)	2.375 (0.712,7.917)
45-49	1.069 (0.816,1.400)	1.670 (1.317,2.118)	1.929 (0.940,3.959)
50-54	1.330 (1.117,1.583)	1.578 (1.336,1.865)	1.725 (1.121,2.654)
55-59	1.358 (1.168,1.579)	1.665 (1.448,1.914)	2.133 (1.486,3.062)
60-64	1.175 (1.020,1.352)	1.708 (1.507,1.935)	1.976 (1.417,2.753)
65-69	1.267 (1.116,1.438)	1.592 (1.416,1.789)	1.721 (1.251,2.368)
70-74	1.362 (1.179,1.574)	1.542 (1.348,1.764)	1.907 (1.334,2.726)
75-79	1.487 (1.160,1.907)	1.534 (1.187,1.983)	1.667 (0.910,3.052)

FEMALES			
Age	*gE*	*Ge*	*GE*
40-44	1.167 (0.301,4.520)	1.333 (0.463,3.836)	5.000 (0.313,79.942)
45-49	0.944 (0.523,1.702)	1.869 (1.139,3.067)	—
50-54	0.947 (0.659,1.361)	1.298 (0.929,1.814)	4.167 (1.800,9.644)
55-59	1.243 (0.967,1.597)	1.280 (0.999,1.641)	2.324 (1.282,4.211)
60-64	1.634 (1.343,1.988)	1.867 (1.538,2.267)	1.842 (1.112,3.053)
65-69	1.321 (1.111,1.571)	1.601 (1.359,1.887)	2.457 (1.637,3.689)
70-74	1.257 (1.045,1.511)	1.538 (1.296,1.825)	2.354 (1.528,3.626)
75-79	1.203 (0.893,1.620)	1.220 (0.896,1.659)	1.773 (0.788,3.986)

a case entered the study at age x last birthday and had a heart attack at age $x + t$ last birthday. A matched control is a person chosen randomly from persons in these two strata who also entered the study at age x last birthday and remained healthy at least until age $x + t + 1$ last birthday. Once chosen as a control, that person cannot be chosen as a control again. As controls are plentiful compared with cases, we will match 5 controls to each case, called a 1:5 matching strategy. In Section 1.2, we mentioned that the genotyping of individuals will be done as and when it is required. So, it might be necessary to genotype a large number of people to ensure that enough controls are available for a 1:5 case-control study. Other matching strategies with fewer controls per case will obviously be cheaper to implement.

To calculate odds ratios, we need to group ages sensibly. Note that epidemiological studies often use quite wide age groups, much wider than actuaries are accustomed to using. We will use 5-year age bands as a reasonable compromise between accuracy and sample size. The results are given in Table 7. We

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

TABLE 8

THE AGE-ADJUSTED ODDS RATIOS CALCULATED FOR BOTH MALES AND FEMALES.

Strata	*gE*	*Ge*	*GE*
Male	1.285 (1.209,1.365)	1.625 (1.536,1.719)	1.880 (1.620,2.182)
Female	1.298 (1.188,1.418)	1.538 (1.413,1.674)	2.250 (1.814,2.790)

can see no particular trend with respect to age, so we calculate the age-adjusted odds ratio for each stratum (a weighted average of the age-specific odds ratios, using the Mantel-Haenszel method described in Woodward (1999)), which are shown in Table 8. Comparing these with the true odds ratios in Table 4 the estimates are better for strata *gE* and *Ge* (with more cases) than for stratum *GE*. However all the true odds ratios lie within the 95% confidence intervals in Table 8.

3.3. An Actuarial Investigation

The actuary starts with the model of Figure 1 in mind, and wishes to estimate the intensity $\lambda^s_{12}(x)$ for each stratum. We assume, realistically, that the best available data are the published odds ratios. The 'estimation' procedure, therefore, consists of finding a reasonably robust way to estimate transition intensities from odds ratios. There is no simple mathematical relationship, so approximations must be made.

Supposing that the actuary knows the rates of heart attack in the general population $\lambda_{12}(x)$ (separately for males and females) a simple assumption is that the heart attack intensity for each stratum is proportional to $\lambda_{12}(x)$. In stratum s, define:

$$\gamma^s_{12}(x) = c_s(x) \times \lambda_{12}(x) \tag{14}$$

where $\gamma^s_{12}(x)$ is the actuary's 'estimate' of $\lambda^s_{12}(x)$. Assuming that the odds ratios (denoted $\psi_s(x)$) are good approximations of the relative risks, which is reasonable as long as the age groups are not too broad, we have:

$$\psi_s(x) = \frac{\gamma_s(x)}{\gamma_{ge}(x)} = \frac{c_s(x)}{c_{ge}(x)} \tag{15}$$

which leads to:

$$c_s(x) = \psi_s(x) \times c_{ge}(x). \tag{16}$$

As observed from Table 7, the odds ratios do not appear to depend strongly on age. So we further assume that $c_s(x)$ is a constant c_s (hence also $\psi_s(x)$ is a constant ψ_s), so:

$$c_s = \psi_s \times c_{ge} \tag{17}$$

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

TABLE 9

THE ESTIMATED MULTIPLIERS c_s FOR EACH STRATUM.

Stratum	*ge*	*gE*	*Ge*	*GE*
Male	0.918	1.179	1.492	1.726
Female	0.920	1.194	1.415	2.070

where ψ_s is the age-adjusted odds ratio. Thus Equation (14) becomes:

$$\gamma^s_{12}(x) = c_{ge} \times \psi_s \times \lambda_{12}(x). \tag{18}$$

Now Equation (11) can be written:

$$\lambda_{12}(x+t) = \frac{\sum_s w_s(x) \exp\left(-\int_0^t c_{ge}\,\psi_s\,\lambda_{12}(x+y)\,dy\right) c_{ge}\,\psi_s\,\lambda_{12}(x+t)}{\sum_s w_s(x) \exp\left(-\int_0^t c_{ge}\,\psi_s\,\lambda_{12}(x+y)\,dy\right)}. \tag{19}$$

Let us assume that at age $x = 60$, the $w_s(x)$ are given by the population frequencies of the respective strata. Now we can solve Equation (19) for the multiplier c_{ge} for a particular choice of age x and any duration t. Then we can use Equation (17) to obtain c_s for $s = gE$, Ge and GE. We find (not shown here) that the results are very similar for different values of t. In Table 9, we show the 'estimated' c_s for representative age $x = 60$ and duration $t = 5$, based on the age-adjusted odds ratios in Table 8. These values can be compared with the true values given in Table 3. They are in good agreement for strata $s = ge$, gE and Ge. The agreement for stratum $s = GE$ is not so good, but it was based on a small number of cases, 241 males and 122 females.

3.4. Premium Rating for Critical Illness Insurance

The actuary will use the intensities $\gamma^s_{12}(x)$ 'estimated' in Section 3.3 to calculate CI insurance premiums. We use the CI insurance model from Gutiérrez & Macdonald (2003), assuming that all intensities except those for heart attack are as given there. For heart attack, we use the intensities $\gamma^s_{12}(x)$. We compute expected present values by solving Thiele's differential equations numerically, with a force of interest of $\delta = 0.044017$ (see Norberg (1995)).

Table 10 shows the true premiums for the strata $s = ge$, Ge and GE, as a percentage of the premiums for stratum ge, for males and females and for different ages and terms. Here, 'true' means that they have been computed using the intensities $\lambda^s_{12}(x)$, not the actuary's estimates. Table 11 then shows the corresponding premiums, as a percentage of those charged for stratum ge, using

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

TABLE 10

THE TRUE CRITICAL ILLNESS INSURANCE PREMIUMS FOR DIFFERENT STRATA AS A PERCENTAGE OF THOSE FOR STRATUM *ge*.

Stratum	*Males*					*Females*				
		Term					Term			
	Age	5	15	25	35	Age	5	15	25	35
gE	45	112%	111%	109%	107%	45	103%	103%	104%	104%
	55	110%	108%	107%		55	104%	105%	105%	
	65	107%	106%			65	105%	106%		
	75	106%				75	106%			
Ge	45	124%	121%	117%	115%	45	105%	107%	108%	108%
	55	119%	116%	114%		55	109%	110%	110%	
	65	114%	112%			65	111%	111%		
	75	111%				75	111%			
GE	45	136%	131%	126%	122%	45	108%	110%	112%	112%
	55	129%	124%	121%		55	113%	115%	115%	
	65	120%	118%			65	116%	117%		
	75	117%				75	117%			

TABLE 11

THE ACTUARY'S ESTIMATED CRITICAL ILLNESS INSURANCE PREMIUMS FOR DIFFERENT STRATA AS A PERCENTAGE OF THOSE FOR STRATUM *ge*.

Stratum	*Males*					*Females*				
		Term					Term			
	Age	5	15	25	35	Age	5	15	25	35
gE	45	112%	110%	109%	107%	45	103%	104%	104%	104%
	55	110%	108%	107%		55	105%	105%	105%	
	65	107%	106%			65	106%	106%		
	75	106%				75	106%			
Ge	45	126%	123%	119%	116%	45	105%	106%	108%	108%
	55	121%	117%	115%		55	108%	109%	109%	
	65	115%	113%			65	110%	110%		
	75	112%				75	111%			
GE	45	137%	132%	126%	123%	45	111%	115%	118%	118%
	55	129%	124%	121%		55	119%	121%	121%	
	65	121%	119%			65	124%	124%		
	75	117%				75	125%			

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

the actuary's estimates $\gamma_{12}^{s}(x)$. The results are similar to those in Table 10. The estimates are good for strata *gE* and *Ge*, but not as accurate for females in stratum *GE*. As mentioned before, this stratum had relatively few cases.

4. Simulation Results

4.1. Varying the Genetic and Environment Model

In the last section, we estimated parameters of a heart attack model and the resulting CI insurance premiums, based on a simulated realisation of UK Biobank. The underlying 'true' model (chosen by us) was particularly simple — two genotypes, two environmental exposures and proportional hazards of heart attack — and by great good luck, our model epidemiologist hit upon exactly the correct hypotheses in fitting his/her model. So it is not surprising that he/she obtained good parameter estimates, with the possible exception of those in respect of the smallest stratum, *GE*.

In reality, the epidemiologist faces more difficult problems:

(a) There is likely to be more than one gene, many with more than two variants, as candidates for influencing the disease.
(b) Similarly, there are likely to be several environmental exposures of interest.
(c) Model mis-specification is always possible (indeed, it may be the norm).
(d) On grounds of cost, the number of cases and the number of controls per case may be limited.
(e) As mentioned earlier, UK Biobank will be a single unrepeatable sample, hence sampling error will be present. Although 500,000 seems like a huge sample, it may not be when smaller numbers of cases are sampled from within it.

In a simulation study, we are in a position to explore these problems. In particular, we can address (d) and (e) above, because we can replicate the entire UK Biobank dataset many times, and repeat the epidemiological and actuarial analyses using each realisation. Thus we can estimate the sampling distributions of parameter estimates and premium rates, while the analysis of the single realization in Section 3 only gave us point estimates of the latter. (We did give approximate confidence intervals of the estimated odds ratios, because they can be derived on theoretical grounds. This is not possible for such a complicated function of the model parameters as a premium rate, and simulation is one of the few practical approaches.) We concentrate on this question in the rest of this paper, because it is directly relevant to the approach adopted by GAIC in the UK, and likely to be adopted by similar bodies elsewhere, which demands that the reliability of prognoses based on genetic information must be demonstrated if they are to be used in any way. In the case of multifactorial disorders, we assume that this requirement is to be interpreted in the statistical sense rather than as applying to individual applicants. Our exploration of (a), (b) and (c) above will be the subject of a future paper.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

In addition to simulating many replications of UK Biobank, we will consider the effect of stronger or weaker genetic and environmental effects, and of more common and less common adverse genotypes. We call each such variant of the underlying model a 'scenario', which should not be confused with the simulation procedure discussed above. We will hold each scenario fixed, and then simulate outcomes of UK Biobank under those assumptions.

We have already introduced one set of assumptions in Section 2, which we will refer to as our Base scenario. The details of all the scenarios are given in Table 12. The parameters that must be specified are:

(a) The population frequency of each stratum (the same for males and females).
(b) The parameters k for each sex and ρ_s for each stratum. Although ρ_s does not depend on sex, for convenience Table 12 shows the combined constants of proportionality $k \times \rho_s$ for each sex.

TABLE 12

THE MODEL PARAMETERS FOR DIFFERENT SCENARIOS. ODDS RATIOS ARE ALSO SHOWN.

Parameters	*Stratum*	*Base*	*Penetrance*		*Frequency*	
			Low	*High*	*Low*	*High*
Population Frequency	*ge*	0.81	0.81	0.81	0.9025	0.64
	gE	0.09	0.09	0.09	0.0475	0.16
	Ge	0.09	0.09	0.09	0.0475	0.16
	GE	0.01	0.01	0.01	0.0025	0.04
ρ_s	*ge*	0.70	0.85	0.55	0.70	0.70
	gE	0.90	0.95	0.85	0.90	0.90
	Ge	1.10	1.05	1.15	1.10	1.10
	GE	1.30	1.15	1.45	1.30	1.30
k (Male)	All	1.317274	1.136603	1.568090	1.370745	1.221620
k (Female)	All	1.316406	1.136463	1.564821	1.370230	1.220385
$k \times \rho_s$ (Male)	*ge*	0.922	0.966	0.862	0.960	0.855
	gE	1.186	1.080	1.333	1.234	1.099
	Ge	1.449	1.193	1.803	1.508	1.344
	GE	1.712	1.307	2.274	1.782	1.588
$k \times \rho_s$ (Female)	*ge*	0.921	0.966	0.861	0.959	0.854
	gE	1.185	1.080	1.330	1.233	1.098
	Ge	1.448	1.193	1.800	1.507	1.342
	GE	1.711	1.307	2.269	1.781	1.587
Odds Ratio	*ge*	1.000	1.000	1.000	1.000	1.000
	gE	1.286	1.118	1.545	1.286	1.286
	Ge	1.571	1.235	2.091	1.571	1.571
	GE	1.857	1.353	2.636	1.857	1.857

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

Although the odds ratios are derived quantities rather than parameters, they are also shown in Table 12 for convenience.

The Low and High Penetrance scenarios assume smaller and larger differences, respectively, between the effects of the different strata, governed by ρ_s. The Low and High Frequency scenarios assume that disadvantageous G genotype and E environment have population frequencies half (0.05) or double (0.2) those in the baseline scenario (0.1), respectively.

In Section 3.2, we noted that problems like missing values and cost constraints might limit the number of cases that can be used for analysis. So we will also examine the effect of limiting the number of cases used in the analysis. From Table 6, around 20,000 individuals were eligible to be considered as cases (in that particular realisation). For each scenario, we will show results based on 1,000, 2,500, 5,000 and 10,000 cases as well as those based on all cases.

4.2. Outcomes of 1,000 Simulations: The Base Scenario

We will make 1,000 simulations of UK Biobank. The outcomes will be the empirical distributions of the parameters of the epidemiologist's model, and of CI insurance premium rates. Let us first consider the Base scenario, all cases included, for males aged 45 taking out a CI insurance policy with term 15 years. Figure 5 shows scatter plots of the CI insurance premium rates per unit sum assured

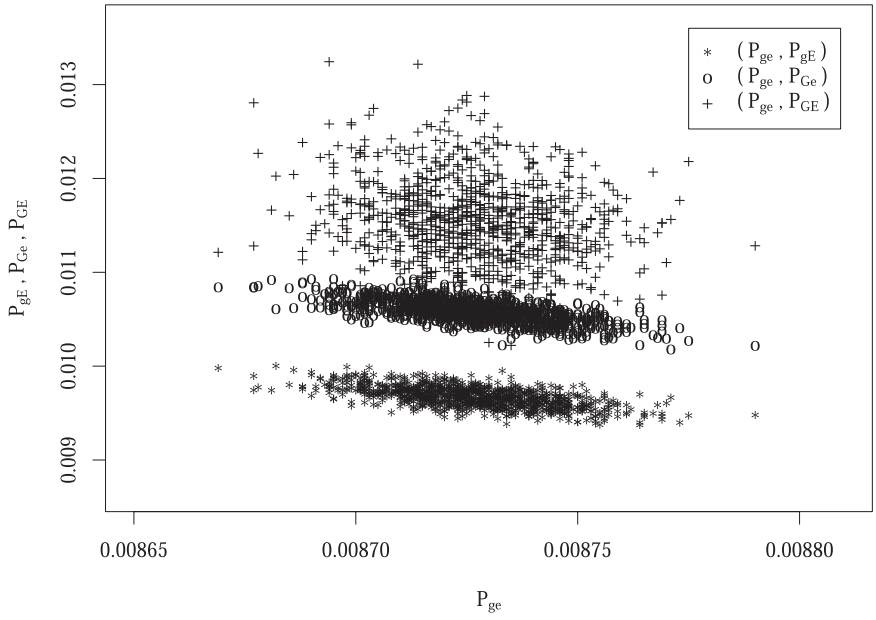

FIGURE 5: Scatter plots of CI insurance premium rates for strata *gE*, *Ge* and *GE* versus that of *ge* under the Base scenario for males aged 45 and policy term 15 years.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

TABLE 13

THE CORRELATION MATRIX OF THE STRATA-SPECIFIC PREMIUM RATES FOR MALES AGED 45 AND POLICY TERM 15 YEARS UNDER THE BASE SCENARIO, ALL CASES INCLUDED.

Stratum	*ge*	*gE*	*Ge*	*GE*
ge	1.000			
gE	–0.604	1.000		
Ge	–0.656	–0.123	1.000	
GE	–0.194	–0.057	–0.095	1.000

TABLE 14

THE CORRELATION MATRIX OF THE PREMIUM RATINGS FOR MALES AGED 45 AND POLICY TERM 15 YEARS UNDER THE BASE SCENARIO, ALL CASES INCLUDED.

Rating	R_{gE}	R_{Ge}	R_{GE}
R_{gE}	1.000		
R_{Ge}	0.095	1.000	
R_{GE}	0.013	–0.018	1.000

for strata *gE*, *Ge* and *GE* versus those of *ge*. More precisely, the outcome of the ith simulation is a drawing $p^i = (p^i_{ge}, p^i_{gE}, p^i_{Ge}, p^i_{GE})$ from the sampling distribution of the 4-dimensional random variable $P = (P_{ge}, P_{gE}, P_{Ge}, P_{GE})$, where P_s is the premium rate in stratum s.

The scatter plots show clearly that the premium rate pairs (P_{ge}, P_{gE}) and (P_{ge}, P_{Ge}) are more strongly correlated than the pair (P_{ge}, P_{GE}). This is true, as the correlation matrix given in Table 13 shows, but note that the scale of the x-axis is greatly compressed compared with that of the y-axis. The reason they are correlated is that, as outlined in Section 3.3, the actuary uses the three odds ratios published by the epidemiologist, plus the overall population intensity of heart attack, to obtain the heart attack intensities for the four strata, so the four premium estimates are not independent. The reason that the correlations are negative is that the overall level of the four intensities is adjusted so that their aggregate effect is consistent with the general population. So, if the intensities in any of the strata are high, the intensities in the others will tend to fall to restore consistency with the aggregate intensity.

We also consider the premium rates for strata *gE*, *Ge* and *GE* as a proportion of those for stratum *ge*, namely $P_{gE}/P_{ge}, P_{Ge}/P_{ge}$ and P_{GE}/P_{ge}. These correspond to premium ratings, if we take the standard premium rate to be that of stratum *ge*, and we will refer to them as such. For brevity, define $R_s = P_s/P_{ge}$ to be the premium rating for stratum s with respect to stratum *ge*. The correlation matrix of these premium ratings is given in Table 14 and the corresponding scatter plots are given in Figure 6. Both suggest correlations are small enough

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

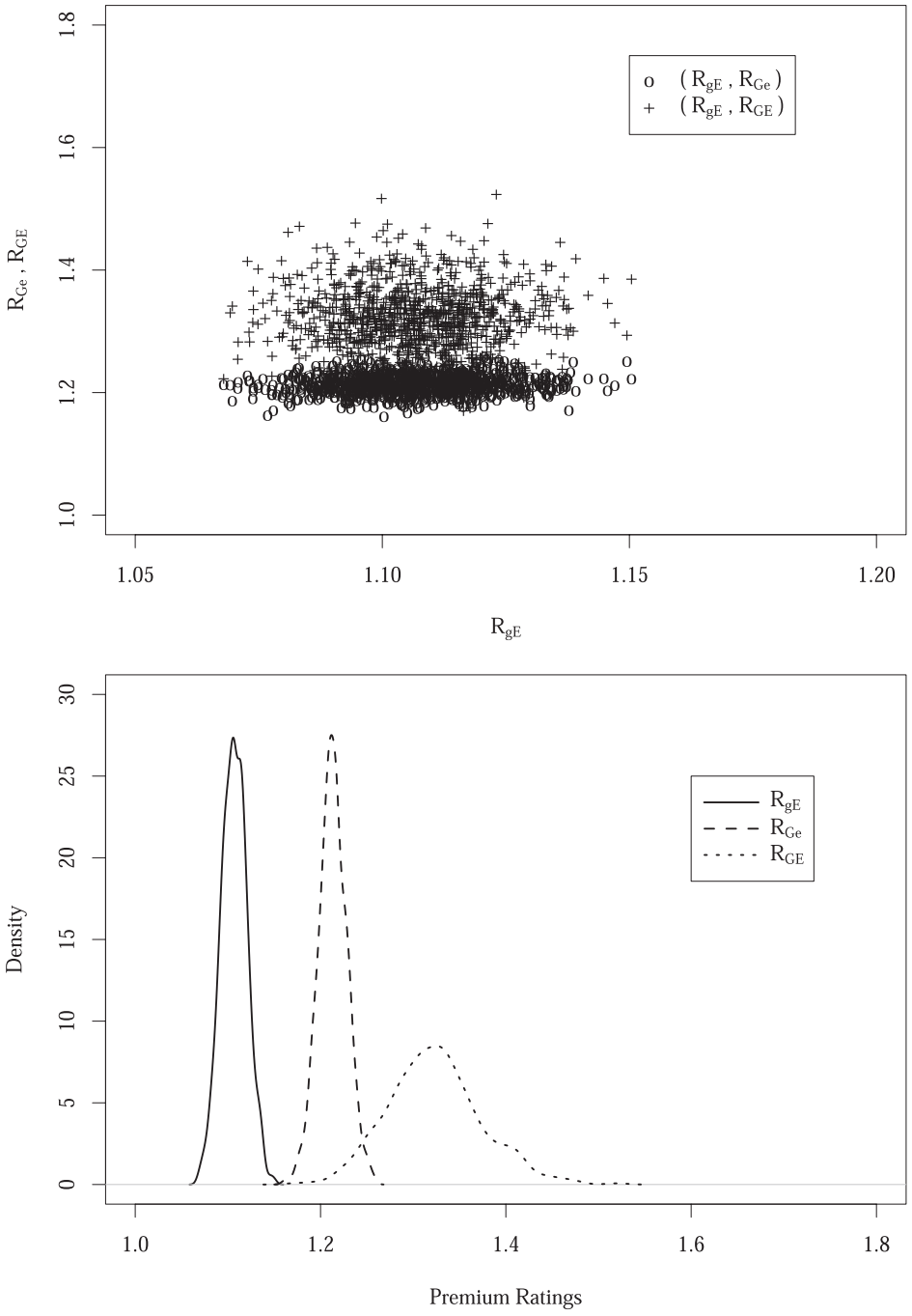

FIGURE 6: The scatter plots of the premium ratings *Ge*/*ge* and *GE*/*ge* versus *gE*/*ge* and the corresponding density plots for males aged 45 and policy term 15 years under the Base scenario, all cases included.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

to neglect, which means that instead of always considering the full joint distribution of the premiums P, we can obtain all the information of interest by separate examination of the marginal distributions of the premium ratings. The densities of these marginal distributions are given in Figure 6. This immediately suggests a simple approach to the questions that GAIC must ask, because the reliability of the premium rating in each stratum — in terms of its distinguishability from the premium ratings in the other strata — is revealed by the degree to which its marginal density overlaps the marginal densities of the others. Presented with Figure 6, we might expect GAIC to agree that strata Ge and GE had premium ratings distinct from that of stratum gE, but to ask whether or not they had premium ratings reliably distinct from each other.

4.3. A Measure of Confidence

Our precise formulation of the question that GAIC might now ask is: are the marginal empirical distributions of premium ratings in different strata sufficiently different to support charging different premiums (when doing so is allowed)? In this section, we suggest a simple measure to address this.

Let X and Y be two continuous random variables with cumulative distribution functions F_X and F_Y respectively. We can find u such that $F_X(u) + F_Y(u) = 1$. If the ranges of X and Y overlap, u lies in both and is unique, otherwise any u that lies between their ranges will do. This can be rewritten as $F_X(u) = 1 - F_Y(u)$, or $\mathrm{P}[X \leq u] = \mathrm{P}[Y > u]$.

Without loss of generality, let us also assume that $F_X(u) \geq F_Y(u)$. Define our measure of confidence to be $2 \times F_X(u) - 1$, which gives a measure of the overlap of F_X and F_Y. Denote this $O(X, Y)$, or just O if the context is clear. If $F_X(u) = F_Y(u) = 0.5$, then we are as unsure as we can be that F_X and F_Y are distinct, and $O = 0$. As $F_X(u)$ increases to 1, the area of overlap decreases. If the ranges of X and Y do not overlap at all, $F_X(u) = 1$ and we have high confidence in deciding that F_X and F_Y are distinct; in this case $O = 1$.

4.4. Results

In this section, we simulate 1,000 realisations of UK Biobank under each scenario outlined in Table 12. Our aim is to examine how reliably UK Biobank might identify differences in premium ratings, as a body like GAIC might require. This is measured by the three quantities $O(R_{gE}, R_{Ge})$, $O(R_{Ge}, R_{GE})$ and $O(R_{gE}, R_{GE})$. We have verified (not shown here) that these do not vary significantly by age or policy term, so in Table 15, we present results for a representative policy for males aged 45 and policy term 15 years.

Note that it is impossible to calculate an odds ratio for a given age group unless there is at least one case in that age group in each stratum. A few of the 1,000 simulations failed this criterion, and these were omitted from the results in Table 15. Those affected were the Base and the Low Penetrance scenarios

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

TABLE 15

THE MEASURE OF OVERLAP O FOR CI INSURANCE PREMIUM RATINGS FOR MALES AGED 45, WITH POLICY TERM 15 YEARS, FOR DIFFERENT SCENARIOS.

Scenario	*Cases*	$O(R_{gE}, R_{Ge})$	$O(R_{gE}, R_{GE})$	$O(R_{Ge}, R_{GE})$
Base	All	1.000	1.000	0.924
	10,000	0.968	0.962	0.632
	5,000	0.872	0.850	0.484
	2,500	0.718	0.698	0.356
	1,000	0.490	0.416	0.176
Low Penetrance	All	0.918	0.904	0.572
	10,000	0.662	0.658	0.346
	5,000	0.528	0.472	0.216
	2,500	0.412	0.360	0.148
	1,000	0.250	0.222	0.076
High Penetrance	All	1.000	1.000	0.992
	10,000	1.000	0.998	0.844
	5,000	0.984	0.970	0.692
	2,500	0.906	0.886	0.540
	1,000	0.688	0.658	0.354
Low Frequency	All	0.996	0.948	0.658
	10,000	0.892	0.706	0.352
	5,000	0.712	0.516	0.208
	2,500	0.566	0.322	0.060
	1,000	—	—	—
High Frequency	All	1.000	1.000	0.994
	10,000	0.988	1.000	0.896
	5,000	0.932	0.986	0.744
	2,500	0.806	0.902	0.546
	1,000	0.594	0.716	0.358

with 1,000 cases (1 simulation omitted in each case) and the Low Frequency scenarios with 2,500 and 1,000 cases (10 and 238 simulations omitted, respectively). We omit the last of these from the table as being possibly misleading. We make the following comments on Table 15:

(a) We saw in Section 4.3 that under the Base Scenario, all cases included, the densities of R_{Ge} and R_{GE} overlap over a small region. This qualitative observation is made more concrete by Table 15, which shows that $O(R_{Ge}, R_{GE}) = 0.924$ in this case. By definition, this means that there exists x such that $\mathrm{P}[R_{Ge} < x] = \mathrm{P}[R_{GE} > x] = 0.962$, and we (or GAIC) may have high confidence in assigning these strata to different underwriting groups.

(b) Stratum GE is always the smallest, so the distribution of R_{GE} is always the most spread out. This is also evident from the scatter plots in Figure 6.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

(c) We expect real case-control studies to use only a subset of cases, and Table 15 shows that the effect of this is very great. For example, in the Base scenario, $O(R_{Ge}, R_{GE})$ falls from 0.924 to 0.176 as the number of cases used falls from 'All' to 1,000. Figure 7 shows, for the Base scenario, the marginal densities with different numbers of cases. The densities overlap considerably if the number of cases is small (and bear in mind that 1,000 cases is not a very small investigation by normal standards).

(d) Figure 8 shows the empirical distribution functions of the premium ratings for males under the Base scenario. For each premium rating, we show the effect of using different numbers of cases. For example, if only 1,000 cases were used, there is about a 30% chance that underwriters would

TABLE 16

THE MEASURE OF OVERLAP O FOR CI INSURANCE PREMIUM RATINGS FOR FEMALES AGED 45 WITH POLICY TERM 15 YEARS, FOR DIFFERENT SCENARIOS.

Scenario	*Cases*	$O(R_{gE}, R_{Ge})$	$O(R_{gE}, R_{GE})$	$O(R_{Ge}, R_{GE})$
Base	All	0.990	0.990	0.734
	10,000	0.958	0.948	0.626
	5,000	0.850	0.844	0.494
	2,500	0.728	0.706	0.378
	1,000	0.466	0.488	0.244
Low Penetrance	All	0.778	0.762	0.402
	10,000	0.680	0.646	0.302
	5,000	0.528	0.506	0.222
	2,500	0.392	0.326	0.122
	1,000	0.238	0.198	0.078
High Penetrance	All	1.000	1.000	0.906
	10,000	1.000	0.998	0.836
	5,000	0.992	0.984	0.696
	2,500	0.914	0.884	0.484
	1,000	0.716	0.656	0.320
Low Frequency	All	0.932	0.800	0.436
	10,000	0.896	0.676	0.298
	5,000	0.748	0.486	0.192
	2,500	0.552	0.340	0.134
	1,000	—	—	—
High Frequency	All	0.998	1.000	0.922
	10,000	0.994	1.000	0.884
	5,000	0.922	0.986	0.756
	2,500	0.814	0.914	0.576
	1,000	0.598	0.678	0.348

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

TABLE 17

THE MEASURE OF OVERLAP O FOR CI INSURANCE PREMIUM RATINGS FOR MALES AGED 45, WITH POLICY TERM 15 YEARS, FOR DIFFERENT SCENARIOS AND A 1:1 MATCHING STRATEGY.

Scenario	*Cases*	$O(R_{gE}, R_{Ge})$	$O(R_{gE}, R_{GE})$	$O(R_{Ge}, R_{GE})$
Base	All	0.990	0.990	0.774
	10,000	0.886	0.872	0.454
	5,000	0.740	0.720	0.374
	2,500	0.554	0.544	0.248
	1,000	0.378	0.400	0.222
Low Penetrance	All	0.808	0.820	0.456
	10,000	0.558	0.526	0.220
	5,000	0.372	0.378	0.188
	2,500	0.288	0.308	0.184
	1,000	0.232	0.204	0.048
High Penetrance	All	1.000	1.000	0.908
	10,000	0.988	0.978	0.680
	5,000	0.898	0.902	0.494
	2,500	0.762	0.742	0.366
	1,000	0.548	0.480	0.222
Low Frequency	All	0.954	0.856	0.474
	10,000	0.738	0.558	0.284
	5,000	0.574	0.464	0.228
	2,500	—	—	—
	1,000	—	—	—
High Frequency	All	1.000	1.000	0.950
	10,000	0.944	0.986	0.746
	5,000	0.826	0.932	0.592
	2,500	0.668	0.802	0.456
	1,000	0.474	0.594	0.306

TABLE 18

THE NUMBER OF SIMULATIONS REJECTED DUE TO THE INABILITY TO CALCULATE THE ODDS RATIOS FOR A 1:1 MATCHING STRATEGY.

Scenario	*Number of Cases*				
	All	10,000	5,000	2,500	1,000
Base	0	0	0	0	13
Low Penetrance	0	0	0	0	16
High Penetrance	0	0	0	0	36
Low Frequency	0	0	6	123	630
High Frequency	0	0	0	0	0

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

incorrectly assume R_{GE} to be 150% or higher. If instead 10,000 cases were used the chance of making this error is very small.

(e) Figure 9 shows, for 5,000 cases, the effect of the different scenarios. Reduced frequency of the adverse genetic and environmental exposures, or reduced penetrance of the adverse genotype, both reduce the ability to discriminate between different underwriting classes. Changes in the opposite direction improve the discrimination. This qualitative observation is backed up in a more quantitative way by Table 15.

Table 16 gives the corresponding results for females (again, we omit the results for the Low Frequency scenario with 1,000 cases because of a large number of simulations with undefined odds ratios). When a fixed number of cases is used the results are very similar to those for males. This is as expected, as we assumed that the effects of genotype and environmental exposures were the same for males and females, albeit acting on different baseline risks of heart attack. However, when all cases are included, the values of O are smaller than those for males. This is because the lower incidence of heart attack among females results in fewer cases, therefore estimates with higher variances.

Until now, we have used a 1:5 matching strategy for all case-control studies; that is, five controls per case. However, cost constraints might dictate the use of fewer controls. In Table 17, we show the values of O for males if a 1:1 matching strategy is used. As expected these are decreased significantly under all scenarios.

As we mentioned when discussing Table 15, we may find simulations under which the odds ratios cannot be calculated because of a lack of cases. Also, note that the odds ratio can only be calculated if there are enough exposed controls. This is more demanding under a 1:1 matching strategy, as fewer controls are available than in 1:5 matching strategy. (At first sight this is surprising; it ought to be easier to find a smaller number of controls. This is true, but there is also a higher chance that one of the cells in the 2×2 table used to calculate the odds ratio will be empty, see Table 19 in the Appendix.) Table 18 shows the numbers of simulations rejected for this reason. The numbers are rather high for the Low Frequency scenarios where 1,000 and 2,500 cases were used. The results based on the remaining simulations may not be reliable and so these are not given in Table 17.

5. Conclusions

In this paper we ask the question: how well may UK Biobank distinguish between different levels of risk associated with the influence of genes, environment and their interactions on a given multifactorial disorder?

On the basis of our simple model, we conclude that the ability of case-control studies based on UK Biobank to identify distinct CI underwriting classes was marginal. If a very large number of cases was used, quite reliable

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

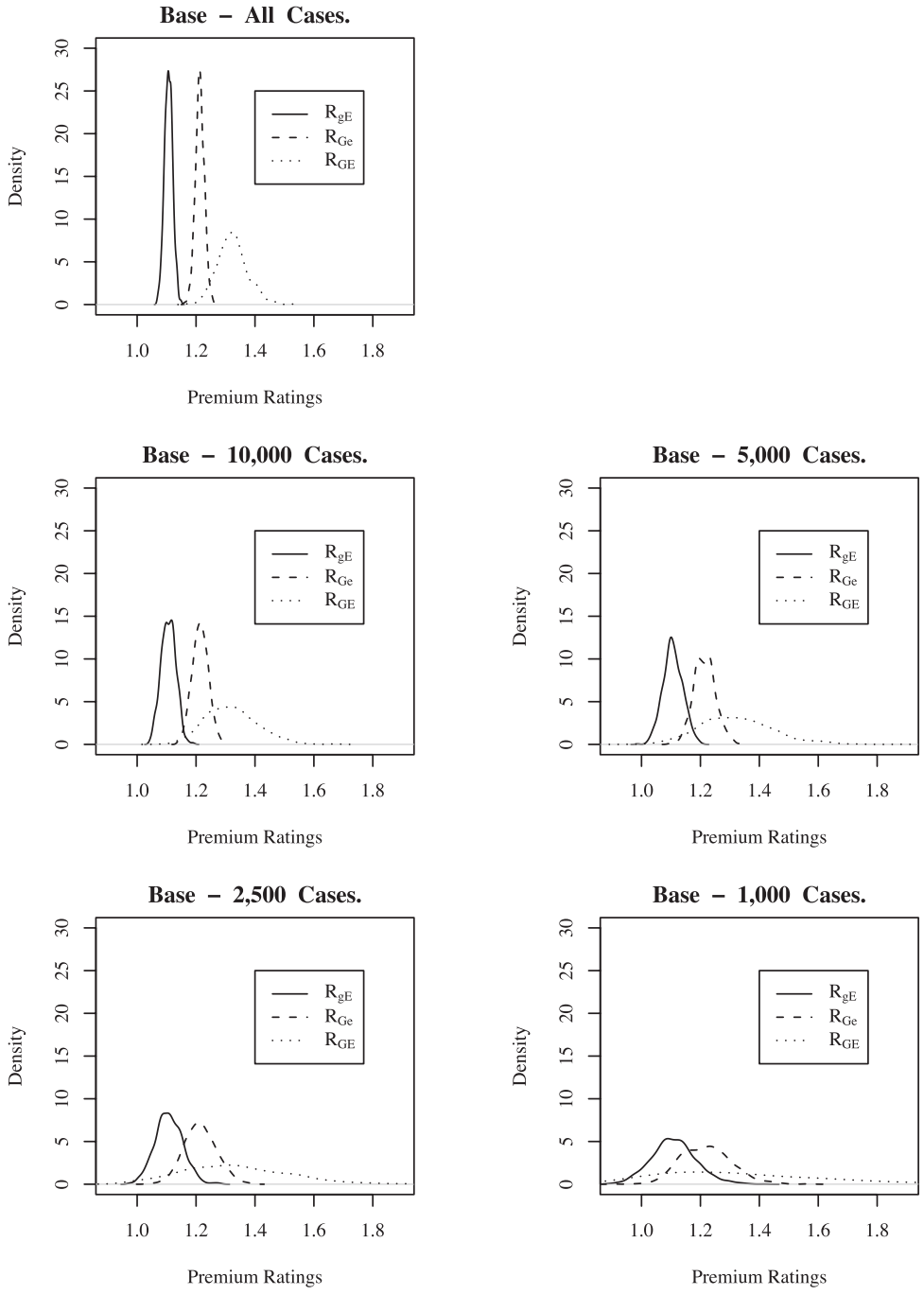

FIGURE 7: Marginal densities of premium ratings in the Base scenario (males) with different numbers of cases in the case-control study.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

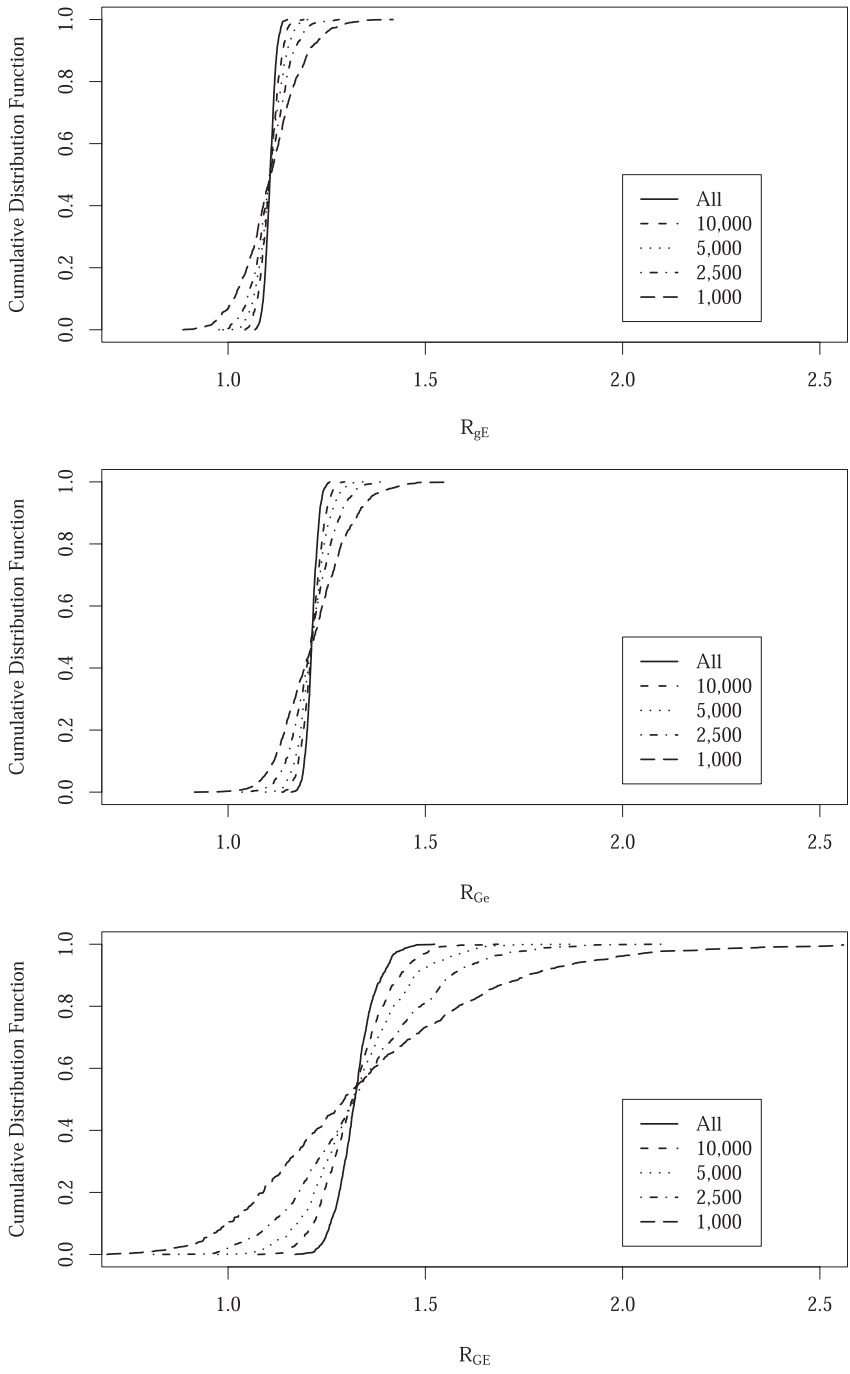

FIGURE 8: The empirical cumulative distribution function of the premium ratings *gE*/*ge*, *Ge*/*ge* and *GE*/*ge* for males aged 45 and policy term 15 years under the Base scenario.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

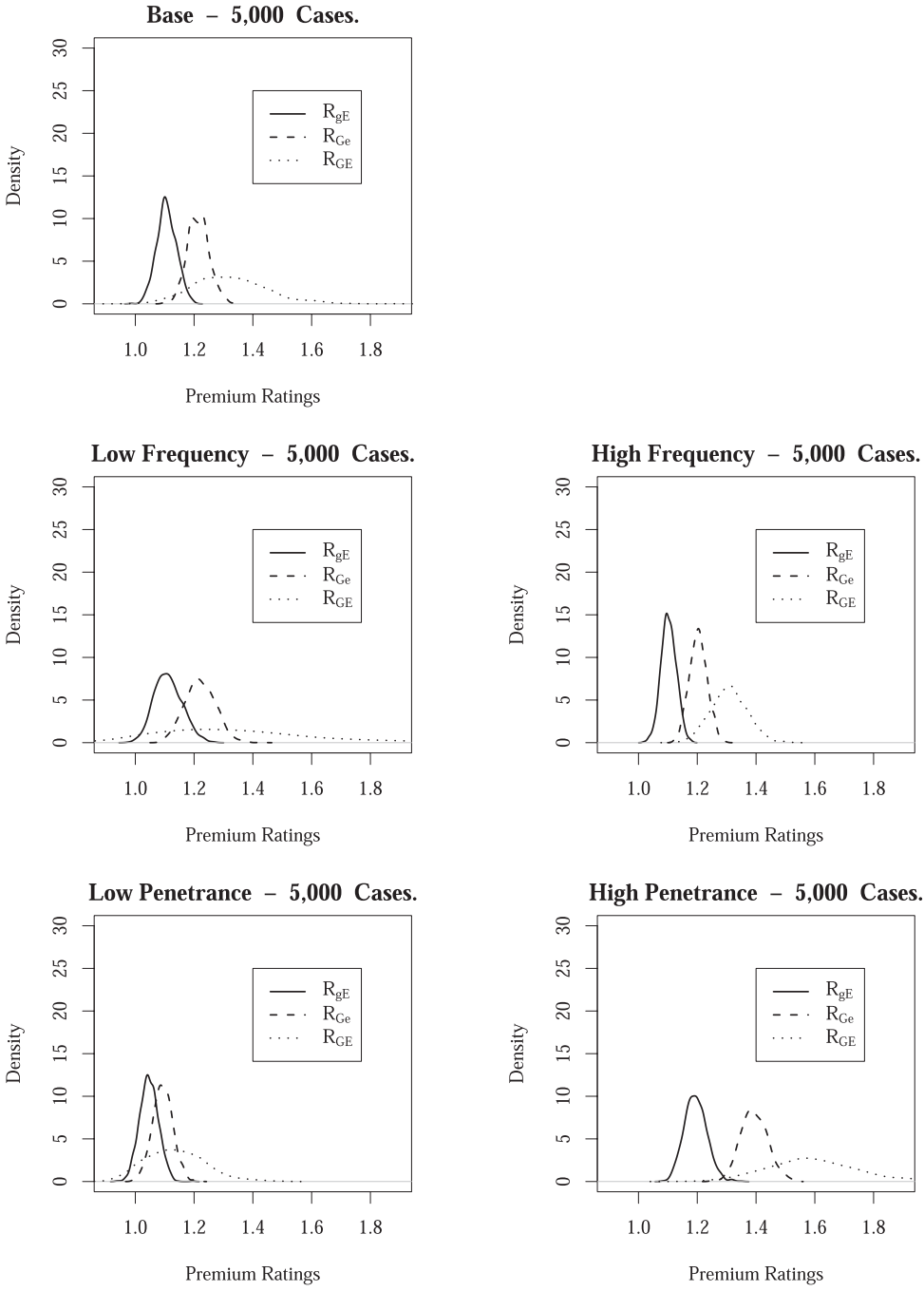

FIGURE 9: Marginal densities of premium ratings in different scenarios (males), with 5,000 cases in the case-control study.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

discrimination was achieved, but this is a very expensive option. If a more realistic number of cases was used — a few thousands — the power to discriminate quickly diminished. In particular, it was clear that if the effects of the adverse genotype and adverse environment were any less than we had assumed, the power to discriminate would be rather poor.

This conclusion ought to bring comfort to those who are worried about insurers' use of genetic information, and to insurers themselves. This is particularly important during the 5 to 10 years that must pass before UK Biobank itself starts to yield results. We have found no support for the idea that very large-scale genetic studies like UK Biobank will lead to significant changes in underwriting practice.

Our study has been very simple and idealised in several respects mentioned above. Most obviously, our genetic model is not truly multifactorial, although it does allow for a basic environmental interaction. Further research is in hand to extend the model to a more realistic, though still hypothetical, representation of a multifactorial genetic contribution to heart attack. Our aim will be to find out whether this will strengthen or weaken the discriminatory power of genetic tests, along the lines that GAIC has pioneered for single-gene disorders. Another point that will repay further study is the possibility of model mis-specification.

Acknowledgements

This work was carried out at the Genetics and Insurance Research Centre at Heriot-Watt University. We would like to thank the sponsors for funding, and members of the Steering Committee for helpful comments at various stages.

References

Breslow, N.E. and Day, N.E. (1980) *Statistical Methods in Cancer Research: Volume 1 – The analysis of case-control studies*. International Agency for Research on Cancer.

Capewell, S., Livingston, B.M., MacIntyre, K., Chalmers, J.W.T., Boyd, J., Finlayson, A., Redpath, A., Pell, J.P., Evans, C.J. and McMurray, J.J.V. (2000) Trends in case-fatality in 117,718 patients admitted with acute myocardial infarction in Scotland. *European Heart Journal*, **21**, 1833-1840.

Daykin, C.D., Akers, D.A., Macdonald, A.S., McGleenan, T., Paul, D. and Turvey, P.J. (2003) Genetics and insurance – some social policy issues (with discussions). *British Actuarial Journal*, **9**, 787-874.

Gutiérrez, C. and Macdonald, A.S. (2003) Adult polycystic kidney disease and critical illness insurance. *North American Actuarial Journal*, **7(2)**, 93-115.

Macdonald, A.S. (2004) Genetics and insurance management, in *The Swedish Society of Actuaries: One Hundred Years*, ed. A. Sandström, Svenska Aktuarieföreningen, Stockholm.

Macdonald, A.S. and Pritchard, D.J. (2000) A mathematical model of Alzheimer's disease and the ApoE gene. *ASTIN Bulletin*, **30**, 69-110.

Norberg, R. (1995) Differential equations for moments of present values in life insurance. *Insurance: Mathematics and Economics*, **17**, 171-180.

Woodward, M. (1999) *Epidemiology: Study Design and Data Analysis*. Chapman & Hall.

APPENDIX

A BRIEF INTRODUCTION TO CASE-CONTROL STUDIES

This appendix describes the main features of a case-control study, one of the main tools in epidemiology, but which is not well-known to actuaries. Standard references for this material are Woodward (1999) and Breslow & Day (1980). The question asked is: how do genetic and environmental risk factors interact to affect the risk of a disease or other outcome?

We usually want answers that are valid for the general population. The best way to proceed is to carry out a prospective cohort study — recruit a properly randomised sample of healthy people, observe them over time, and see how the suspected risk factors correlate with cases of the disease. Such a study should be free of any selection biases. However, it will be time consuming and (especially for rare diseases) prohibitively expensive. It is much quicker and cheaper to select a sample of people who already have the disease of interest, and a sample of people who do not have the disease, and see if the suspected risk factor turns up more often in the diseased sample. This is a case-control study; the two samples are known as cases and controls, respectively. But, because the cases are chosen retrospectively from known sufferers of the disease, the sampling may be biased. The statistical question, therefore, is what inferences can be drawn about the general population, from a case-control study whose cases have been selected retrospectively?

Controls should be a representative sample of disease-free individuals, who had exactly the same chances as the cases of become diseased, except in respect of the risk factors of interest. For example, suppose we are studying the effect of smoking on lung cancer. If all the cases are over 60 years old, and all the controls are under 30 years old, we can hardly compare their respective risks of lung cancer. Therefore we would match controls to cases: given as a case a man age 65 who had had lung cancer, a suitable control might be a man age 65 who had not. The general approach would be to match in respect of as many factors as possible that might affect the risk of lung cancer, except smoking habits. Then the proportions of smokers among cases and controls ought to be informative.

Matching reduces confounding when comparing cases and controls, but not within each of the two groups. For example, if age is thought to affect the risk of lung cancer, we would not want to analyse as a single unit a sample of

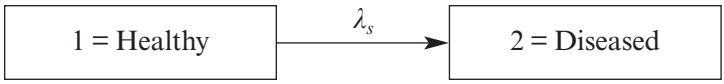

FIGURE 10: A 2-state model of a disease, in respect of a person in stratum s.

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

cases whose ages ranged from 20 to 80: to do so would be to assume, implicitly, that risk was unaffected by age. The usual approach would be to stratify the sample by age, say into 5-year or 10-year age groups. We would stratify both cases and controls with respect to risk factors other than that being studied, ensuring that matched cases and controls are in corresponding strata.

We will look at the simplest situation, a risk factor with two levels: an individual either is exposed to it or is not (for example, smokers and non-smokers). We introduce a simple model with two states, 'Healthy' and 'Diseased', see Figure 10. The (constant) transition intensity λ_s governs the probability that a healthy person in stratum s will become diseased, as follows: over a small time dt this probability is approximately $\lambda_s dt$.

The ideal outcome of a study would be estimates of all the λ_s. However, a retrospective study cannot, in general, provide unbiased estimates of the λ_s. Next best (with actuarial models in mind) might be the relative risks: the relative risk in stratum s, compared with stratum z, is $r_{sz} = \lambda_s/\lambda_z$. Then, if we could just establish λ_z in a single stratum, we could find λ_s in any stratum. Unfortunately, a retrospective study cannot, in general, provide unbiased estimates of relative risks either. Since the data are what they are, we have to seek relevant quantities that can be estimated in an unbiased fashion from them. The main example is the odds ratio.

Whenever we study the probabilities of events in epidemiology, a time interval is involved, which we denote T. Expressions such as 'the probability that X occurred' should be read as 'the probability that X occurred during the interval of length T'. For example, T might be equal to the width of the age-groups used to stratify the sample.

Let P_s be the probability that a person in stratum s suffered the disease (during a period of length T), and let $Q_s = 1 - P_s$. Choose one stratum, z say, as the baseline: the most common stratum is often chosen. Then the odds in strata s and z are P_s/Q_s and P_z/Q_z respectively, and the odds ratio in stratum s relative to the baseline is:

$$\phi_{sz} = \frac{\text{Odds in stratum } s}{\text{Odds in stratum } z} = \frac{P_s Q_z}{Q_s P_z}. \tag{20}$$

Suppose, for simplicity, we are studying the effect of two genotypes, g and G, on lung cancer. We can draw up the simple 2×2 table in Table 19. Then ad/bc is an unbiased estimate of the odds ratio ϕ_{Gg} of genotype G with respect to genotype g. This is true regardless of the retrospective sampling scheme, and is the reason why odds ratio are normally reported in case-control studies (see Woodward (1999, Chapter 6)).

If there is reason to believe that there is a common true odds ratio for all strata, it can be estimated by the Mantel-Haenszel statistic:

$$\hat{\psi} = \left(\sum_{s \neq z} \frac{a_s b_z}{N_s} \right) \Big/ \left(\sum_{s \neq z} \frac{b_s a_z}{N_s} \right) \tag{21}$$

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press

TABLE 19

TWO-WAY TABLE OF NUMBERS OF CASES AND CONTROLS BY GENOTYPE.

	Diseased (Cases)	*Disease-free (Controls)*	*Total*
Genotype G	a	b	$a+b$
Genotype g	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

where the summation is over all strata s except the baseline stratum z, a_s and b_s are the numbers of cases and controls, respectively, in stratum s, and $N_s = a_s + b_s + a_z + b_z$.

If controls are more plentiful than cases, more efficient estimates can be obtained by matching $c > 1$ controls to each case, called $1:c$ matching. However, as c increases, the marginal increase in efficiency decreases, so c is rarely greater than 5 in practice. For $1:c$ matching, the Mantel-Haenszel estimate of the odds ratio is as follows:

$$\hat{\psi} = \frac{\sum_{u=1}^{c}(c+1-u)\,m_u}{\sum_{u=1}^{c} u\,(t_u - m_u)} \tag{22}$$

where:

t_u = the number of sets with u exposures
m_u = the number of sets with u exposures in which the case is exposed.

The actuary's problem is to 'estimate' intensities from published odds ratios, plus some other information to provide a baseline, such as the intensity in one stratum or (more often) the general population. If T is reasonably short, we have:

$$\frac{P_s}{P_z} = \frac{1-\exp(-\lambda_s T)}{1-\exp(-\lambda_z T)} \approx \frac{\lambda_s}{\lambda_z} = r_{sz}. \tag{23}$$

Moreover if all the probabilities P_s are small, then $Q_s \approx 1$ and then:

$$\psi_{sz} = \frac{P_s Q_z}{Q_s P_z} \approx \frac{P_s}{P_z} \approx r_{sz}. \tag{24}$$

ANGUS MACDONALD
Maxwell Institute for Mathematical Sciences and
Department of Actuarial Mathematics and Statistics,
Heriot-Watt University,
Edinburgh EH14 4AS, U.K.
Tel.: +44(0)131-451-3209
Fax: +44(0)131-451-3249
E-mail: A.S.Macdonald@ma.hw.ac.uk

https://doi.org/10.2143/AST.36.2.2017924 Published online by Cambridge University Press