The impact of individual heterogeneity (frailty) in cancer

Morten Valberg

Dissertation for the degree of PhD

NORWEGIAN **CANCER** SOCIETY

Department of Biostatistics

Institute of Basic Medical Sciences

Faculty of Medicine

University of Oslo

Norway

Oslo, January 2014

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo
No. 1787

ISBN 978-82-8264-809-7

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Akademika Publishing.
The thesis is produced by Akademika Publishing merely in connection with the thesis defence. Kindly direct all inquiries regarding the thesis to the copyright holder or the unit which grants the doctorate.

Acknowledgments

Firstly, I am grateful for the excellent guidance and support of my supervisor Odd O. Aalen. I would also like to thank my co-supervisor Marit B. Veierød, for her meticulous feedback and great encouragement.

Furthermore, I would like to thank all my co-authors. Steinar Tretli and Tom Grotmol at the Cancer Registry of Norway, for the many fruitful discussions we have had. Tron A. Moger, for his many suggestions and advices on technical issues. Susan S. Devesa, for her important inputs and discussions on the first paper. Paola M. V. Rancoita, Clelia Di Serio, Elia Biganzoli and Romano Demicheli, for their many thoughts and discussions on the tumor dormancy phenomenon.

I would like to thank the Cancer Society of Norway for funding this project (grant number 171851).

I would also like to thank all my colleagues at the Department of Biostatistics, for creating such an enjoyable working environment. Also, I would like to thank everybody at the University Centre of Statistics for Biomedical Sciences (CUSSB), Vita-Salute San Raffaele University, for welcoming me, and taking good care of me during my stay in Milan, Italy.

Last, but not least, I would like to thank my friends and family, especially Torunn, for their encouragement and for making this period of my life so gratifying.

Oslo, January 2014

Morten Valberg

List of papers

Paper 1

Valberg M, Grotmol T, Tretli S, Veierød MB, Devesa SS, Aalen OO. Frailty modeling of age-incidence curves of osteosarcoma and Ewing sarcoma among individuals younger than 40 years. *Statistics in Medicine* 2012; **31**(28):3731–3747. DOI: 10.1002/sim.5441.

Paper 2

Valberg M, Grotmol T, Tretli S, Veierød MB, Moger TA, Aalen OO. A hierarchical frailty model for familial testicular germ-cell tumors. *American Journal of Epidemiology* 2013. DOI: 10.1093/aje/kwt267.

Paper 3

Aalen OO, Valberg M, Grotmol T, Tretli S. Understanding variation in disease risk: the elusive concept of frailty. *Revised and re-submitted, International Journal of Epidemiology.*

Paper 4

Rancoita PMV, Valberg M, Demicheli R, Biganzoli E, Di Serio C. Investigating tumor dormancy with frailty models. *Manuscript.*

Contents

1 Introduction

The fact that individuals are different is a fundamental observation of life. Some differences are obviously important, like being a female compared to a male, if you are to bear a child. Other differences are more subtle, and harder to observe. Genetic predisposition towards a disease is one example. A population consists of a mix of individuals, where some may have a high risk, some may have a low risk, or some may even be immune to the outcome we are studying. We say that the population is *heterogeneous*.

Survival data, consisting of the time from some starting point to an event, is frequently encountered in a large range of fields, from medicine to economics. *Survival analysis*, that comprises the statistical tools for analyzing such data, has become very important. Of special importance is the *hazard rate*, which expresses the instantaneous rate of having an event at a particular time (given survival up to that time). For many, the method of choice for handling survival data, is the Cox proportional hazards model [1]. The model allows for the inclusion of individuals that are lost to follow-up, i.e. censored, and the population is divided into subgroups by different *covariates* that we suspect to have an impact on the phenomenon under study. In fact, describing the differences between individuals by means of covariates is the way heterogeneity between individuals are usually taken into account in medical statistics and in epidemiology. In general it is not likely that all relevant covariates can be included in such an analysis. Some covariates are not included because of practical, ethical or economic reasons, while others are simply not suspected to have an influence. This consequently leads to some *unobserved* heterogeneity between individuals. What if the 'important' heterogeneity is caused by these unobserved covariates? Even if all relevant covariates could have

been included, there will always be some unexplained rest. Failing to account for unobserved differences may be serious.

Accounting for unobserved heterogeneity between individuals has a long history. In 1959, Beard studied mortality in a population, and introduced a random effect to model survival data [2]. The term *frailty*, however, was not introduced until 1979, by Vaupel et al. [3]. The frailty is considered a random variable that models the variation in risk between individuals. The *individual hazard rate* is defined as the frailty variable multiplied with some *basic hazard rate* that is shared by all individuals, but possibly influenced by covariates. In this way, the heterogeneity due to observed and unobserved factors are separated.

Several cancers has peaking incidence rate curves with respect to age. In the interpretation of such an observation, it can be of crucial importance to take into account possible frailty effects. One straight forward interpretation of the shape of the incidence curve could be that each individual experiences an increasing risk of developing the disease up to a certain age, before the risk starts to decline. Taking the frailty view, on the other hand, the explanation would be that the shape of the incidence curve is a result of selection effects in the population. While the risk for each individual is increasing throughout life, there are some individuals with a very high susceptibility. The highly susceptible, or frail, individuals will develop the disease early. After some time, the population will contain a smaller and smaller proportion of highly susceptible individuals, and the incidence rate will drop. The age-incidence curves of several cancers fit this description, and the incidence curves of e.g. testicular cancer [4], colorectal cancer [5], nasopharyngeal carcinoma [6] and Hodgkin lymphoma [7] have all been analyzed from a frailty point of view. Also the incidence rates of other diseases, like Schizophrenia [8], have been modeled by

this approach. It is important to note that in such studies, the frailty approach has biological underpinnings, and is not just a mathematical construct. For cancers, or other diseases, where some heritable component is thought to be present, frailty models offers a way of describing the varying genetic disposition to the disease. In a cancer setting, these models use the knowledge available on the individual level, trough e.g. a carcinogenic model like the Armitage-Doll model [9], combined with the notion of a varying susceptibility between individuals, to describe observations in a population.

In many situations it is not reasonable to assume that individuals are independent. For monozygotic twins, for instance, the times to some event are often associated in some way (e.g. when a disease that has some genetic component is under consideration). The simplest multivariate frailty model that accounts for this is the *shared* frailty model, introduced by Clayton in 1978 [10]. Rather than letting the frailty be distributed across individuals, the frailty is distributed across clusters, which could e.g. be pairs of monozygotic twins. However, in many situations the shared frailty model is not flexible enough, and it might be more reasonable to let the frailties be *correlated* rather than shared within the clusters. If a pair of monozygotic twins have another sibling, it is often the case that the twins are more correlated than a twin and the other sibling. Several other types of multivariate frailty models has been proposed to handle such situations. Hougaard introduced a multiplicative frailty model, where the frailty of an individual was the product of two independent frailties [11]. One of these frailties would be common for all siblings, while the twins share another part of the frailty independently of the ordinary sibling. Another correlated frailty model, uses an additive decomposition of each individuals' frailty, and different parts of the additive components

are shared between individuals to construct different correlation structures. Early papers introducing this type of models were Pickels et al. in 1994 [12] and Yashin et al. in 1995 [13].

This thesis considers several aspects of frailty analysis applied in a cancer setting. **Paper 1** modifies the Armitage-Doll model with random frailty, to capture the narrow peaks in the age-incidence curves of osteosarcoma and Ewing sarcoma. **Paper 2** analyzes population wide data on the incidence of testicular germ-cell tumors (TGCTs) in Norwegian families, and we emphasize the calculation of familial risk of disease. The paper considers a frailty model that randomizes a parameter in the compound Poisson distribution for individual frailty. The randomized parameter is decomposed additively to account for the correlation structure in a family with up to five children. In **paper 3**, the persuasive evidence of the presence of (unmeasured) inter-individual variation in the risk of cancer, as well as other diseases, is discussed, and several potentially surprising effects of frailty on standard epidemiological measures is pointed out. **Paper 4** introduces frailty effects as a possible explanation of the peaks seen in the hazard rates (with respect to re-occurrence and mortality) after treatment of several cancers, including breast cancer. A modified univariate frailty model for modeling bimodal hazard rates is introduced, and applied to analyze a dataset of breast cancer patients to examine whether the bimodality of the hazard rate, with respect to mortality, could be explained by this hypothesis.

Carcinogenic models are often used as a justification of the parametric form of the basic hazard rate in the frailty model, and such carcinogenic models are discussed in Section 2. In Section 3, the different frailty models mentioned in this introduction are discussed in somewhat more detail, and the *hierarchical* frailty

model is introduced [14–17]. Section 4 considers the definition, and calculation, of familial risk of diseases. The frailty relative risk (FRR) is introduced. Section 5 gives summaries of the four papers, while a discussion and some ideas for future work are given in Section 6.

2 Models of carcinogenesis

Inspired by the work of Nordling [18], Armitage and Doll postulated their famous model of carcinogenesis in 1954 [9]. Their model stated that a cell had to go through certain transitions to become malignant. They stressed the fact that these transitions did not necessarily have to be mutations, and that the actual nature of the transitions was not important in a mathematical analysis. The basic observation was that incidence rates of many cancers increased with a certain power of age, such that a plot of the logarithm of age versus the logarithm of the incidence rate gave a straight line. Such cancers are often termed *log-log cancers*. Armitage and Doll demonstrated this feature for a number of cancers, limited to the age interval 25-74 years. However, the model did not fit all cancers, especially not those with a peaking incidence at an early age. In their original article Armitage and Doll stated that

> "The relatively high rates at the younger ages could result if the population contained a group of subjects specially susceptible to cancer (...)".

After the Armitage-Doll model was first published, other models have been proposed. The Moolgavkar-Knudson model considered two mutations [19]; first a transformation from a normal cell to an intermediate cell, followed by a period of clonal expansion, culminating in a mutation of an intermediate (daughter) cell to a malignant cell. Various shapes of population incidence rates were explained by a varying number of available susceptible stem cells, and a peaking age-incidence rate were explained by the exhaustion of susceptible stem cells within each individual. Several modifications of the model have been made [20–23], including allowing for more than two mutations [20].

Even though it is simple, the Armitage-Doll model continues to play a fundamental role in the theory of carcinogenesis. Combined with a proper treatment of the unobserved heterogeneity between individuals, the Armitage-Doll model is able to capture also the peaking incidence rate of cancers [4, 5, 8], which is what Armitage and Doll foresaw in the above given quote. Rather than the exhaustion of susceptible stem cells, the heterogeneity, in the susceptibility to the disease, between individuals is considered as a possible explanation of a peaking incidence curve.

Throughout this thesis, and in the papers that comprise it, we consider an individual hazard rate as the product of a basic hazard rate and a random variable that describes the frailty of the individual. In **papers 1-3**, the basic hazard rate is assumed to increase with some power of age, which corresponds to a *Weibull* form, and the Armitage-Doll model is used as the justification. However, other approaches also give rise to a Weibull form of the basic hazard rate [24, 25]. In fact, the Weibull model is the most commonly used parametric model for carcinogenesis [25].

3 Frailty models

This section gives an overview of the models that form the starting points for methods developed and used in the papers in this thesis. However, all types of frailty models are not considered here, and the reader is referred to one of the books of the field for a complete overview [26–30].

3.1 The proportional frailty model

In the proportional frailty model the hazard rate of an individual is given as a basic hazard rate, $\lambda(t)$, multiplied with a quantity, Z, that expresses the extent of frailty in the individual (see e.g. chapter 6 in Aalen et al. [28]),

$$\alpha(t|Z) = Z\lambda(t), \tag{1}$$

where t is the time from some starting point (typically, t is age). The frailty, Z, is considered a random variable over the population. The individual survival function is given by

$$S(t|Z) = \exp(-Z\Lambda(t)),$$

where $\Lambda(t) = \int \lambda(t)\mathrm{d}t$. The population survival function is found by integrating out the frailty variable in the individual survival function,

$$S(t) = \mathrm{E}[\exp(-Z\Lambda(t))] = \int \exp(-z\Lambda(t))f(z)\mathrm{d}z,$$

where E denotes expectation, and f is the distribution function of Z. Since the Laplace transform of the distribution function of Z is given by

$$\mathscr{L}(c) = \mathrm{E}[\exp(-cZ)] = \int \exp(-cz)f(z)\mathrm{d}z,$$

the population survival function can be expressed as $S(t) = \mathscr{L}(\Lambda(t))$. The population hazard rate is then found as

$$\mu(t) = \lambda(t)\frac{-\mathscr{L}'(\Lambda)}{\mathscr{L}(\Lambda)}.$$

As mentioned above, a Weibull form of the basic hazard rate is usually assumed,

$$\lambda(t) = kat^{k-1}, \quad k > 0,$$

where a is a scale parameter. When studying cancer incidence, this is justified by carcinogenic models [9, 25], and k is interpreted as the number of events necessary for a cell to become malignant [4]. Thus, $k > 1$ is a more realistic interval for k [4]. The model combines the available information on the individual level, from the carcinogenic model, with what is actually observed in the population. The risk of having an event (i.e. cancer) will increase throughout life. The most frail, or susceptible, individuals are likely to have an event early, causing the population hazard rate to rise. With time, there will be a decreasing proportion of highly susceptible individuals left in the population, and the population hazard rate will decrease. In **paper 1**, the proportional frailty model (with Weibull basic hazard rate) is referred to as *The Armitage-Doll model with random frailty*.

3.2 Frailty distributions

The most commonly used frailty distribution is the gamma distribution [28]. Hougaard presented the *power variance function* (PVF) family of distribution that includes both the gamma distribution and several other distributions as special cases [26, 31]. Aalen showed that the compound Poisson (cP) distribution is also included in this class of distributions [32]. The cP distribution may be defined as the sum of N independent gamma distributed variables, each with scale parameter ν and scale parameter η, where N is Poisson distributed with expectation ξ. The papers in this thesis considers the gamma and the cP distributions only. The

Laplace transforms are given as

$$\mathscr{L}_{cP}(c) = \exp\left(-\xi\left(1-\left(\frac{\nu}{\nu+c}\right)^{\eta}\right)\right), \tag{2}$$

where and ξ, ν and η are all > 0, and

$$\mathscr{L}_{gamma}(c) = \exp\left(-\delta\left(\log\left(\theta+c\right)-\log\left(\theta\right)\right)\right),$$

where the shape parameter, δ, and the scale parameter, θ, are both > 0. However, the PVF distributions are part of a broader class of distribution, namely the Lévy-type distributions. Consider the Laplace transform

$$\mathscr{L}(c) = \exp(-s\phi(c)), \tag{3}$$

where s is a non-negative parameter and $\phi(c)$ is called a *Laplace exponent*, and has a parametric form depending on the frailty distribution. This is a valid Laplace transform for all Lévy-type distributions [28, 33]. For the cP, and the gamma distribution, the Laplace exponents are given as

$$\phi_{cP}(c) = \xi\left(1-\left(\frac{\nu}{\nu+c}\right)^{\eta}\right) \quad , \quad \xi > 0$$

and

$$\phi_{gamma}(c) = \delta\left(\log\left(\theta+c\right)-\log\left(\theta\right)\right). \tag{4}$$

However, for the cP distribution, s is usually subsumed in the ξ parameter, or, equivalently, the re-parameterization $\rho = s\xi$ is used, and the Laplace transform

and the Laplace exponent are consequently frequently written as

$$\mathscr{L}_{cP}(c) = \exp(-\rho\phi_{cP}(c))$$

and

$$\phi_{cP}(c) = 1 - \left(\frac{\nu}{\nu + c}\right)^{\eta},$$

respectively [28], yielding the Laplace transform in (2), with ρ considered the expectation of the underlying Poisson variable. A similar re-parameterization may be done for the gamma distribution. However, it can be useful to keep the s parameter explicit, especially in the formulation of the hierarchical frailty model (Section 3.6), as done by Moger et al. [17].

3.3 Modeling bimodal hazard rates

Several cancers, and other diseases, exhibits incidence rates with more than one peak. Taking frailty into account may be of crucial importance in these instances. For nasopharyngeal carcinoma, Haugen et al. proposed a method for modeling the two peaks seen in the incidence rate [6]. They modeled the individual hazard rate as a linear combination of two independent frailties,

$$\alpha(t|Z_1, Z_2) = Z_1\lambda_1(t) + Z_2\lambda_2(t).$$

The first frailty, Z_1 represented the risk due to unobserved "genetic and viral factors", while Z_2 represented risk due to unobserved lifestyle factors. Both λ_1 and λ_2 were assumed to have a Weibull form, and the population hazard rate was found by the same approach as in Section 3.1. The model has also been used

for modeling the bimodal hazard rate of Hodgkin lymphoma [7]. In both these applications [6, 7], data from large registries with many events were analyzed. However, it is well known that the estimation of frailty models from single event data may be accompanied by a high degree of uncertainty [28]. For situations where less data are available, this could pose a problem. Furthermore, it can be hard to justify the separation of the unobserved heterogeneity into two different sources. In **paper 4**, a simpler model for modeling bimodal hazard rates is proposed, where the individual hazard rate is given as

$$\alpha(t|Z) = Z\left(\lambda_1(t) + \lambda_2(t)\right).$$

The frailty acts multiplicatively on the sum of two basic hazard rates. The basic idea is that there are two distinct underlying biological processes that are both, over time, increasing the individual risk of having an event. One of these may, for instance, be increasing steadily from some defined starting point, while the other process might be increasing much more slowly in the beginning of the period under study, and then increase more rapidly and at some point begin to dominate the sum of the two hazards. In that case, it would mean that the frailest individuals are likely to have an event related to the first hazard, while only the strongest (least frail) will in practice be influenced by the second process. This would lead to a bimodal shape of the population hazard rate. In **paper 4**, where survival after conservative breast cancer treatment is studied, it is hard to justify the separation of the unobserved heterogeneity in the same way as by Haugen et al. [6]. Especially considering the relatively short time span in our study ($\sim$ 10 years of follow-up).

3.4 Shared frailty models

In the proportional frailty model (Section 3.1), Z expresses the level of frailty in an individual. However, in many situations it is likely that some survival times are related to each other. Organs within the same individual, or units in a hospital are examples. As mentioned in the Introduction, it is, for a genetic disease, often the case that two monozygotic twins have a similar predisposition for attaining the disease (beyond what is modified by possible measured environmental risk factors). One way of accounting for this is to let the twins *share* the same value of the frailty.

The shared frailty model takes the same form as the proportional frailty model, but the frailty is distributed over clusters (e.g. families) rather than individual entities (persons). The proportional frailty model is thus a special case of the shared frailty model, with cluster size one. The shared frailty model is discussed at great length in the books by Hougaard [26] and Duchateau and Janssen [27]. While we are interested in the distribution of the frailty in the univariate (i.e. proportional) frailty model, it is possible to estimate the level of frailty in each cluster from the shared frailty model using an empirical Bayes approach (see e.g. Aalen et al. chapter 7 [28]).

3.5 Additive frailty models

Letting each member of a cluster share the same value of the frailty may be too simplistic. In a multi-center trial, it is for instance possible that hospitals that are geographically closer to each other might be more correlated than those further apart. Different units within hospitals may have different frailties, but are likely to be correlated to some degree. Different members of a family will most

certainly have some correlation, but it is for instance obvious that a father and a son is more correlated than a father and a mother, if a disease with some genetic component is under consideration. Additive frailty models are constructed to take these aspects, i.e. correlation structures within clusters, into account, and are therefore frequently referred to as *correlated* frailty models (although correlated frailty models is a broader class of models, where the frailty of an individual is not necessarily decomposed additively).

Yashin et al. considered a frailty model where a sum of several frailty variables acted multiplicatively on the basic hazard rate [13]. They noted that in twin studies, the frailty should be correlated rather than shared, to e.g. be able to include both monozygotic and dizygotic twins. They considered three independent gamma distributed variables, Y_0, Y_1 and Y_2. The frailty of the twins was expressed as $Z_1 = Y_0 + Y_1$ and $Z_2 = Y_0 + Y_2$, respectively. In fact, studies of twins is the most frequent application in papers developing and/or applying additive frailty models [12, 13, 34–40]. The model can, however, easily be extended to model other correlation structures [41, 42]. Several types of correlated frailty models are thoroughly discussed in the book by Wienke [29].

Korsgaard and Andersen gave several examples of what they have named *additive genetic frailty models* [41]. Their simplest example considered a family of a mother, a father and a child. They considered four identically, independently distributed gamma$(\lambda/2, \lambda)$ random variables, Y_1, Y_2, Y_3 and Y_4. Here, $Z_F = Y_1 + Y_2$, $Z_M = Y_3 + Y_4$ and $Z_C = Y_1 + Y_3$ represented the additively decomposed frailties of the father, mother and the child, respectively. Thus, Y_1 and Y_3 was the parts of the parents' genes affecting the frailty transmitted to the child. There was no correlation between the frailty of the parents, whereas the child and each parent

had correlation 1/2. This simple model forms the basis of the familial correlation structure used in **paper 2**.

3.6 Hierarchical frailty models

A hierarchical frailty model is a useful combination of the models described in Sections 3.1, 3.4 and 3.5. Let the distribution function of the frailty variable Z_1 have Laplace transform

$$\mathscr{L}_{Z_1}(c) = \mathrm{E}[\exp(-cZ_1)] = \int \exp(-cz_1) f_{Z_1}(z_1)\mathrm{d}z_1 = \exp(-z_2\phi_1(c)), \qquad (5)$$

where E denotes expectation and f_{Z_1} denotes the distribution function of Z_1. The parametric form of the Laplace exponent ϕ_1 depends on the choice of frailty distribution, and z_2 is a constant parameter. The population survival function is $S(t) = \mathscr{L}_{Z_1}(\Lambda(t))$, as in Section 3.1.

Let now the parameter z_2 be randomized by Z_2, which has Laplace transform

$$\mathscr{L}_{Z_2}(c) = \exp(-z_3\phi_2(c)),$$

where ϕ_2 depends on the distribution of Z_2, and z_3 is a non-negative constant parameter. The marginal Laplace transform for the combined frailty of each individual is then given by

$$\mathscr{L}(c) = \mathrm{E}[\exp(-Z_2\phi_1(c)] = \int \exp(-z_2\phi_1(c)) f_{Z_2}(z_2)\mathrm{d}z_2 = \exp(-z_3\phi_2(\phi_1(c))),$$

where f_{Z_2} is the distribution function of Z_2. Similarly, we could add another level in the model, by randomizing z_3 by Z_3, whose distribution function has the Laplace

transform $\mathscr{L}_{Z_3}(c) = \exp(-z_4\phi_3(c))$. The Laplace transform of the combined frailty of an individual would then be

$$\mathscr{L}(c) = \mathrm{E}[\exp(-Z_3\phi_2(\phi_1(c)))] = \exp(-z_4\phi_3(\phi_2(\phi_1(c)))).$$

If even further levels are needed, one could continue by randomizing z_4, and so on. The first level could represent the frailty variation between individuals, as in the proportional frailty model, the second level could represent the frailty that varies between families (but are shared within families). The third level could e.g. represent frailty variation between neighborhoods and so on. However, in **paper 2** we have two levels only.

As discussed in Section 3.2, the parameter z_{i+1} in the Laplace transform for the variable Z_i, can be viewed directly as a parameter in the distribution of Z_i [14–16], or as a scale transformation of that parameter [17]. If Z_i is cP distributed, and

$$\phi_i(c) = 1 - \left(\frac{\nu}{\nu + c}\right)^{\eta},$$

then the parameter z_{i+1} is equivalent the underlying Poisson parameter in the cP distribution. This parameterization is used for the cP distributed first-level frailty in **paper 2**. For the gamma distribution, the parameterization given in expressions (3) and (4) is used in **paper 2**. That is, z_{i+1} is viewed as a scale transformation of a parameter in the frailty distribution. However, since no more levels were added, $z_{i+1} = 1$ in that case (and hence there is no difference between the two parameterizations of Section 3.2).

Hierarchical frailty models have been developed in recent years [14–16], and does not fit in the same group as other classes of multivariate frailty models (e.g.

additive). Interpretation-wise, letting the first-level frailty be cP distributed is very interesting. Randomizing z_2 allows the members of different families to have different probabilities of having zero frailty. Given z_2, the individuals in a family are independent. Although members of a family shares the value of z_2, the model e.g. allows for two members of a family were one has zero frailty, while the other could have a positive frailty.

It does not always make sense that all the family members have the same second-level frailty. Moger et al. proposed to decompose a level in the hierarchical frailty model, to impose a certain correlation structure in the family [17]. They used the same approach as Korsgaard and Andersen [41], although allowing for two children in a family. Furthermore, they had three levels in their model; the first one described the varying individual frailty, the second level described the correlation structure of a family in an genetic additive fashion as described by Korsgaard and Andersen, and the third level described a frailty shared among the family members due to environmental factors. The frailties on all levels (and in the additive decomposition) were gamma distributed, with the expectation on each level set to one. A consequence is that they were able to estimate the parameter of the third-level frailty due to the correlation between the parents in the family.

As mentioned above, the model in **paper 2** has two levels only; the individual variation in frailty on the first level, and the familial frailty variation on the second level. The model allows for up to five children, and decomposes the genetic second-level frailty into 16 components, as opposed to two in Korsgaard and Andersen [41], and four in Moger et al. [17]. Since females are not considered, adding another level, to serve as a shared environment term is not feasible. However, we add a term accounting for the environment shared among brothers independently

of their fathers, inspired by the work on additive models [41, 42]. In this way it is ensured that the correlation is larger between brothers than between father and son. Furthermore, the model in **paper 2** uses a cP distributed first-level frailty, that opens up for interesting interpretational aspects. The randomized z_2 parameter is correlated rather than shared between individuals in a family. This means that two individuals does not need to have the same probability of having zero frailty, although the probabilities are correlated.

4 Familial Risk

There are several cancers that shows some degree of familial clustering. Such a clustering naturally leads us to think that there are some heritable aspects involved in the development of the cancer. Families usually share a very similar environment, and environmental risk factors could also play a role in familial clustering. However, as discussed in **paper 3**, such risk factors are not likely to be the main drivers of familial risk [43, 44]. In any case, familial risk is something that both patients and their families might be concerned with, and it is something that even show up in the general media form time to time. It is therefore of great interest how such risks are estimated.

Familial risks are usually considered as the risk of developing the disease in question if a family member is affected, and compare that risk to the general risk level in the population. It is not obvious how to give such an estimate, especially for siblings, and there are several different approaches. One option is to have an index case in each family (the proband), and to compare the incidence rate in family members of these index cases to the incidence rate in the population [45]. Another approach is to let the first individual in the family who has an event be

the index case, and to let the family members be a part of the risk set from the time of the index case had the event [46].

Usually, estimation of familial risk is based on data from population wide registries. Over the last decade or so, the most dominant approach for providing estimates for sibling risks is the *cohort method* given by Hemminki et al. [47]. As mentioned in **paper 2**, it is not entirely clear how these calculations are done, because there appears to be an error in the formula in Hemminki et al. [47] (see **paper 2**). The paper is nevertheless heavily cited (see selected references in **paper 2**), and the formula is even reproduced with the same error elsewhere [48, 49]. However, the approach seems to be to pool all siblings with at least one affected sibling into one big cohort, and find the *standardized incidence ratios* (SIRs) by comparing the incidence rate in this cohort to that of the general population. This means that in a group of siblings with more than one affected individual, all individuals give a contribution to this defined cohort. It is not intuitively easy to understand the rationale behind the formula, and it needs better justification. In any case, although useful, the SIR approach provides merely summary measures.

4.1 Frailty relative risk (FRR)

The large population wide registries in the Nordic countries provides, because of the possible linkage trough the personal identification number, unique opportunities to study familial diseases. More advanced analyses of the complex data that are available could provide much more information than analyses giving summary measures. Hierarchical (and additive) frailty models can be very useful in this regard.

Consider two members of the same family, individuals A and B. Moger and

Aalen proposed a frailty relative risk (FRR) to express the familial association of a disease [15, 16],

$$FRR = \frac{\text{P(A has disease within time } t_A | \text{Family member B has disease within time } t_B)}{\text{P(A has disease within time } t_A | \text{Family member B has not had disease within time } t_B)}. \quad (6)$$

The fact that given all higher level frailties any pair of individuals are independent, the FRR can be expressed trough the Laplace transforms of the frailty variables. The relative risk in (6) resembles a traditional relative risk, comparing two exclusive groups. In studies of familial risk, however, the comparison of interest is usually the general risk level in the population. With this in mind, the FRR in **paper 2** is defined as

$$FRR = \frac{\text{P(A has disease within time } t_A | \text{Family member B has disease within time } t_B)}{\text{P(A has disease within time } t_A)}, \quad (7)$$

which is actually the same as the definition originally given by Moger et al. [14]. The FRR may be expanded to include more members in the family that develops disease or not within specified ages. In Web Appendix 4 of **paper 2** the FRRs for certain constellations are expressed in terms of the survival functions. However, the FRR may easily be calculated for any constellation. Also, the FRR can be defined to condition on the exact timing of when the family member developed then disease. The FRR would then be expressed in terms of the derivative of the survival function.

The FRR allows for great flexibility with regards to the structures and the sizes of the families, and enables us to be very precise in our definition of familial risks. The ease of which the relative risk given multiple affected and/or unaffected family members are found, once the model parameters have been estimated, is particularly appealing. As of now, the hierarchical frailty model is somewhat

computing intensive, and one advantage of the SIR approach is that it is easy to calculate.

5 Summary of papers

5.1 Paper 1

Frailty modeling of age-incidence curves of osteosarcoma and Ewing sarcoma among individuals younger than 40 years.

The Armitage-Doll model with random frailty (i.e the proportional frailty model with Weibull basic hazard rate) model the incidence rate of several cancers well [4–6]. In this paper it is demonstrated that this model is not able to capture the very steeply increasing, and later plummeting, incidence rates of the bone cancers osteosarcoma and Ewing's sarcoma. This is the case even if the cP frailty distribution is used. The peaks in the incidence rates of the two cancers coincides with the growth spurt period in adolescence, a period where the host tissue (i.e. bone) is expanding. A model that takes into account a biological mechanism that is accelerated at some, possibly short, period of life, is presented. The model fit the incidence curves of osteosarcoma and Ewing's sarcoma well when growth is seen as the accelerated process, and the results support evidence for an underlying susceptibility for these diseases. It is indicated that (susceptible) individuals with an unusually rapid growth spurt may develop the diseases earlier than they would have if their growth spurts had been closer to the average. This could lead to an excess incidence early in puberty, followed by a compensation leading to a faster decrease of the incidence rate than would have been expected if growth spurts were

more similar across individuals.

5.2 Paper 2

A hierarchical frailty model for familial testicular germ-cell tumors.

We analyzed incidence data on testicular germ-cell tumors (TGCT) in all Norwegian families registered since the personal identification number was introduced in 1960, obtained trough a linkage between Statistics Norway and the Cancer Registry of Norway. A total of 1,135,320 families were included, and 7,524 families had at least one member affected by TGCT.

Moger et al. developed a hierarchical frailty model where the frailty on a certain level were decomposed in an additive fashion to take the correlation structures within families into account [17]. We modified the model to analyze a male cancer that is also, possibly, influenced by a maternal mode of inheritance. Furthermore, we expanded the model to take into account more children in a family, and we let the correlation between brothers be larger than the correlation between father and son.

The paper highlighted how the FRR provides accurate definitions of familial risks, and the flexibility of the FRR as a measure of familial association. This was demonstrated by its calculation given multiple affected or healthy family members.

Given one affected brother, the lifetime FRR was 5.88 (95% confidence interval (CI): 4.70, 7.36). Given two affected brothers, the FRR was 21.71 (95% CI: 8.93, 52.76), and if there were two additional healthy brothers in the family, the FRR was 15.80 (95% CI: 9.56, 26.11). A borderline significantly higher FRR for non-seminoma than for seminoma (P =0.06), the two main histological subtypes of

TGCT, were found for sons of affected fathers.

Several of the different FRRs estimated in the paper (those who took multiple diseased and/or healthy siblings into account), have not previously been reported for TGCT.

5.3 Paper 3

Understanding variation in disease risk: the elusive concept of frailty.

The paper pointed out how variation in risk that goes beyond measured risk factors are present. It gave examples of how variation in risk of cancers, and other diseases, may be established early in life, or even prior to birth. It was discussed that much of this variation may be due to randomness. Also, it was discussed how even moderate familial risks are indications of very large individual variations in risk of disease. Furthermore, a number of consequences of frailty on fairly standard epidemiological measures were pointed out. This included peaking age-incidence curves, as in **paper 1**, and also how frailty may affect incidence rates over calendar time. It was also discussed how frailty may play an important role when comparing hazard rates of different groups in e.g. a clinical trial. The paper stressed that observations in a population may be totally unrepresentative for the individual, and the biological processes within. In many situations the presence of unobserved heterogeneity cannot be ignored.

5.4 Paper 4

Investigating tumor dormancy with frailty models.

Several cancers show peaking hazard rates with respect to re-occurrence and, eventually, death, after treatment of the original tumor [50, 51]. The peak in the hazard rate may be found a long time after the initial treatment. For some cancers, several peaks have been observed, and the latest peak have occurred many years after initial treatment. This phenomenon is usually explained by *tumor dormancy*, which means that malignant cells or micrometastases remains non-proliferative a long time after treatment. We studied overall survival in a dataset from the Milan National Cancer Institute, consisting of 1657 breast cancer patients having received conservative surgery followed by radiotherapy.

A frailty model for capturing bimodal hazard rates was proposed. The model was simpler than previously suggested models [6, 7], in the sense that only one frailty variable was introduced in the model. The model took on the form of a proportional frailty model, but the frailty variable was multiplied with the sum of two basic hazard rates. The two basic hazard rates represented the two main micrometastatic processes, i.e. originating from single dormant cells or from dormant micrometastases, and were both assumed to have a Weibull form. The model captured the two peaks in the hazard rate, and the paper set forth the hypothesis that this behavior could be, at least in part, due to a selection phenomenon. Although it is clear that other models based on different assumptions might fit the date equally well (or even better), the paper offered a new view on the bimodality of the hazard rates (with respect to death) following treatment of breast cancer.

6 Discussion

As the title of this thesis suggests, the aim has been to elucidate the presence, and importance, of unobserved individual differences in cancer. The term *cancer* is,

of course, very broad, and includes over 100 diseases that may arise from almost any cell type in the body [52]. Even though they have some common features, it is challenging to make general statements about such a large range of diseases. Nevertheless, it makes intuitively sense that the development of any disease is subject to individual differences beyond the ones that are known to have an effect. **Paper 3** focuses on pointing out the evidence for large variation in individual risk, and demonstrates, using only hypothetical examples, some possible implications of this variation on epidemiological measures. **Papers 1**, **2** and **4** focuses on specific cancer types. These three papers all have some degree of methodological development, and an applied side. When univariate frailty models are applied to data, there is always some speculative elements involved. Both **paper 1** and **paper 4** are no exceptions in that regard, and **paper 4** is perhaps the most speculative. However, the assumptions made are reasonable from a biological point of view in both papers. When clustered data is analyzed, the degree of speculation is reduced [28], and **Paper 2** is, perhaps, the most interesting from both a methodological and applied standpoint.

The contribution of **paper 1** is the modification of the Armitage-Doll model with random frailty, to account for an accelerated biological process in the host tissue of the cancer, at some point in time. One (at least implicit) criticism of the Armitage-Doll model as such, has been that it is not able to take into account an expanding host tissue [19]. **Paper 1** shows that this is the case even when frailty is accounted for. However, a consequence of the paper is that the Armitage-Doll model may easily be modified to take this into account, which was also mentioned in **paper 3**. The proposed model fits the incidence data of osteosarcoma and Ewing's sarcoma well, and another contribution of **paper 1**

is thus to act as supporting evidence for an underlying predisposition for these cancers. Possible extensions of this work would be to include data from several sources. Since obesity may influence both the timing of puberty and the velocity of growth in adolescence [53], it would be of great intresest to analyze data from countries at different stages of the so-called obesity epidemic. Furthermore, the model could possibly be improved by letting the timing of the onset and end of the period of enhancement of the biological process be distributed according to some distribution, rather than being fixed quantities.

A contribution of **paper 2** is the development of a hierarchical frailty model that takes familial correlation into account, for a disease that only males can develop, but that has a possible maternal mode of inheritance. Furthermore, the model allows for a correlation that are different for siblings than for parent-child pairs (as opposed to the study by Moger er al. [17]), as well as allowing for more children per family. The paper demonstrates how, once the parameters of the model has been estimated, the familial risk given virtually any combination of healthy/diseased family members may be found. Of special interest is the 21-fold increased risk of developing TGCT if two brothers have had the disease, and how this estimate is affected by additional, healthy brothers. Such estimates have, to our knowledge, not previously been published for TGCT. Although this study was the first analyzing data from a complete national registry using this type of methodology, it would be very interesting to be able to combine data from all (or some) of the Nordic countries. In this way, more precise estimates for the relative risks considering multiple affected family members could be obtained. Furthermore, it would be useful to have a ready-made computer program for estimating this type of models. Although estimation of the model used in **paper 2** was some

computer intensive, this could be feasible by implementing the program in a faster, lower-level programming language than `R`.

The contributions of **paper 3** are related to the illustrations of the existence of frailty effects in many aspects of biology and in the development of cancers and others diseases. The paper discusses how artefacts due to frailty, well known from survival analysis, may be of crucial importance for correctly interpreting fairly standard epidemiological measures. Thus, the paper contains both justifications of the importance of a frailty approach in analyzing data in various situations, as well as communicating to a non-statistical audience that frailty variation is important for correct interpretation.

The main contribution of **paper 4** is that it generates the hypothesis that the bimodal mortality rate after treatment of breast cancer patients is, in part, due to selection (i.e. frailty) effects. A univariate frailty model for capturing bimodal hazard rates is developed, and used to estimate the population hazard rate with respect to mortality in breast cancer patients receiving conservative treatment. The knowledge of the mechanisms that control what is referred to as tumor dormancy is limited, and the effects of the included covariates may thus be useful with regards to which patients that should be followed up most closely. With regards to further work, it would be interesting to apply the model to a even larger dataset, to obtain more precise results. Furthermore, a dataset (of breast cancer or other, relevant types of cancer) with longer follow-up could allow for the model to be expanded to include another basic hazard rate, possibly capturing another peak (if present). Also, it could be interesting to study the causes of death, and to apply a competing risks model where the cause-specific hazard rates could take on the form of the proposed model, with possibly correlated frailties.

References

1. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 1972; **34**(2):187–220.

2. Beard RE. Appendix: Note on some mathematical mortality models. In *Ciba Foundation Symposium-The Lifespan of Animals (Colloquia on Ageing)*, Wolstenholme GEW, O'Connor M, eds. John Wiley & Sons, Ltd, Chichester, UK, 1959; **5**:302–311. DOI: 10.1002/9780470715253.app1.

3. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 1979; **16**:439–454. DOI: 10.2307/2061224.

4. Aalen OO, Tretli S. Analyzing incidence of testis cancer by means of a frailty model. *Cancer Causes and Control* 1999; **10**:285–292. DOI: 10.1023/A:1008916718152.

5. Svensson E, Moger TA, Tretli S, Aalen OO, Grotmol T. Frailty modelling of colorectal cancer incidence in Norway: indications that individual heterogeneity in risk is related to birth cohort. *European Journal of Epidemiology* 2006; **21**:587–593. DOI: 10.1007/s10654-006-9043-8.

6. Haugen M, Bray F, Grotmol T, Tretli S, Aalen OO, Moger TA. Frailty modeling of bimodal age-incidence curves of nasopharyngeal carcinoma in low-risk populations. *Biostatistics* 2009; **10**:501–514. DOI: 10.1093/biostatistics/kxp007.

7. Grotmol T, Bray F, Holte H, Haugen M, Kunz L, Tretli S, Aalen OO, Moger TA. Frailty modeling of the bimodal age-incidence of Hodgkin lymphoma in

the Nordic countries. *Cancer Epidemiology, Biomarkers and Prevention* 2011; **20**(7):1350–1357. DOI: 10.1158/1055-9965.EPI-10-1014.

8. Svensson E, Rogvin M, Hultman CM, Reichborn-Kjennerud T, Sandin S, Moger TA. Schizophrenia susceptibility and age of diagnosis–a frailty approach. *Schizophrenia Research* 2013; **147**(1):140–146. DOI: 10.1016/j.schres.2013.03.004.

9. Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer* 1954; **8**:1–12. DOI: 10.1038/bjc.1954.1.

10. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978; **65**(1):141–151. DOI: 10.1093/biomet/65.1.141.

11. Hougaard P. A class of multivariate failure time distributions. *Biometrika* 1986; **73**(3):671–678. DOI: 10.1093/biomet/73.3.671.

12. Pickles A, Crouchley R, Simonoff E, Eaves L, Meyer J, Rutter M, Hewitt J, Silberg J. Survival models for developmental genetic data: age of onset of puberty and antisocial behavior in twins. *Genetic Epidemiology* 1994; **11**(2):155–170. DOI: 10.1002/gepi.1370110206.

13. Yashin AI, Vaupel JW, Iachine IA. Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies* 1995; **5**(2):145–159. DOI: 10.1080/08898489509525394.

14. Moger TA, Aalen OO, Heimdal K, Gjessing HK. Analysis of testicular cancer

data using a frailty model with familial dependence. *Statistics in medicine* 2004; **23**(4):617–632. DOI: 10.1002/sim.1614.

15. Moger TA, Aalen OO. A distribution for multivariate frailty based on the compound poisson distribution with random scale. *Lifetime Data Analysis* 2005; **11**(1):41–59. DOI: 10.1007/s10985-004-5639-z.

16. Moger TA, Aalen OO. Regression models for infant mortality data in norwegian siblings, using a compound poisson frailty distribution with random scale. *Biostatistics* 2008; **9**(3):577–591. DOI: 10.1093/biostatistics/kxn003.

17. Moger TA, Haugen M, Yip BH, Gjessing HK, Borgan Ø. A hierarchical frailty model applied to two-generation melanoma data. *Lifetime data analysis* 2011; **17**(3):445–460. DOI: 10.1007/s10985-010-9188-3.

18. Nordling C. A new theory on the cancer-inducing mechanism. *British journal of cancer* 1953; **7**(1):68–72.

19. Moolgavkar SH, Knudson AGJ. Mutation and cancer: a model for human carcinogenesis. *Journal of the National Cancer Institute* 1981; **66**(6):1037–1052. DOI: 10.1093/jnci/66.6.1037.

20. Luebeck EG, Moolgavkar SH. Multistage carcinogenesis and the incidence of colorectal cancer. *Proceedings of the National Academy of Sciences* 2002; **99**(23):15095–15100. DOI: 10.1073/pnas.222118199.

21. Meza R, Luebeck EG, Moolgavkar SH. Gestational mutations and carcinogenesis. *Mathematical Biosciences* 2005; **197**(2):188–210. DOI: 10.1016/j.mbs.2005.06.003.

22. Meza R, Jeon J, Moolgavkar SH, Luebeck EG. Age-specific incidence of cancer: Phases, transitions, and biological implications. *Proceedings of the National Academy of Sciences* 2008; **105**(42):16284–16289. DOI: 10.1073/pnas.0801151105.

23. Jeon J, Meza R, Moolgavkar SH, Luebeck EG. Evaluation of screening strategies for pre-malignant lesions using a biomathematical approach. *Mathematical Biosciences* 2008; **213**(1):56–70. DOI: 10.1016/j.mbs.2008.02.006.

24. Aalen OO. Phase type distributions in survival analysis. *Scandinavian Journal of Statistics* 1995; **22**:447–463.

25. Kopp-Schneider A. Carcinogenesis models for risk assessment. *Statistical methods in medical research* 1997; **6**(4):317–340.

26. Hougaard P. *Analysis of Multivariate Survival Data*. Springer-Verlag: New York, 2000.

27. Duchateau L, Janssen P. *The Frailty Model*. Springer: New York, 2008. DOI: 10.1007/978-0-387-72835-3.

28. Aalen OO, Borgan Ø, Gjessing HK. *Survival and Event History Analysis: A Process Point of View*. Springer: New York, 2008. DOI: 10.1007/978-0-387-68560-1.

29. Wienke A. *Frailty Models in Survival Analysis*. Chapman & Hall/CRC: Boca Raton, 2011. DOI: 10.1201/9781420073911-f.

30. Hanagal DD. *Modeling survival data using frailty models*. Chapman & Hall/CRC: Boca Raton, 2011.

31. Hougaard P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 1986; **73**(2):387–396. DOI: 10.1093/biomet/73.2.387.

32. Aalen OO. Modelling heterogeneity in survival analysis by the compound poisson distribution. *The Annals of Applied Probability* 1992; **2**(4):951–972. DOI: 10.1214/aoap/1177005583.

33. Gjessing HK, Aalen OO, Hjort NL. Frailty models based on Lévy processes. *Advances in Applied Probability* 2003; **35**(2):532–550. DOI: 10.1239/aap/1051201659.

34. Yashin AI, Iachine IA. How frailty models can be used for evaluating longevity limits: Taking advantage of an interdisciplinary approach. *Demography* 1997; **34**(1):31–48. DOI: 10.2307/2061658.

35. Iachine IA, Holm NV, Harris JR, Begun AZ, Iachina MK, Laitinen M, Kaprio J, Yashin AI. How heritable is individual susceptibility to death? the results of an analysis of survival data on danish, swedish and finnish twins. *Twin research* 1998; **1**(4):196–205. DOI: 10.1375/136905298320566168.

36. Yashin AI, Iachine IA. Dependent hazards in multivariate survival problems. *Journal of Multivariate Analysis* 1999; **71**(2):241–261. DOI: 10.1006/jmva.1999.1848.

37. Yashin AI, Iachine IA. What difference does the dependence between durations make? insights for population studies of aging. *Lifetime Data Analysis* 1999; **5**(1):5–22. DOI: 10.1023/A:1009622214567.

38. Wienke A, Holm NV, Christensen K, Skytthe A, Vaupel JW, Yashin AI. The heritability of cause-specific mortality: a correlated gamma-frailty model applied to mortality due to respiratory diseases in danish twins born 1870–1930. *Statistics in medicine* 2003; **22**(24):3873–3887. DOI: 10.1002/sim.1669.

39. Wienke A, Ripatti S, Palmgren J, Yashin A. A bivariate survival model with compound poisson frailty. *Statistics in Medicine* 2010; **29**(2):275–283. DOI: 10.1002/sim.3749.

40. Jonker M, Boomsma D. A frailty model for (interval) censored family survival data, applied to the age at onset of non-physical problems. *Lifetime data analysis* 2010; **16**(3):299–315. DOI: 10.1007/s10985-009-9141-5.

41. Korsgaard IR, Andersen AH. The additive genetic gamma frailty model. *Scandinavian Journal of Statistics* 1998; **25**(2):225–269. DOI: 10.1111/1467-9469.00102.

42. Petersen JH. An additive frailty model for correlated life times. *Biometrics* 1998; **54**(2):646–661. DOI: 10.2307/3109771.

43. Aalen OO. Modelling the influence of risk factors on familial aggregation of disease. *Biometrics* 1991; **45**(3):933–945. DOI: 10.2307/2532650.

44. Hopper JL, Carlin JB. Familial aggregation of a disease consequent upon correlation between relatives in a risk factor measured on a continuous scale. *American journal of epidemiology* 1992; **136**(9):1138–1147.

45. Heimdal K, Olsson H, Tretli S, Flodgren P, Børresen A, Fossa S. Familial testicular cancer in norway and southern sweden. *British journal of cancer* 1996; **73**(7):964–969. DOI: 10.1038/bjc.1996.173.

46. Nielsen M, Andersson C, Gerds TA, Andersen PK, Jensen TB, Køber L, Gislason G, Torp-Pedersen C. Familial clustering of myocardial infarction in first-degree relatives: a nationwide study. *European heart journal* 2013; **34**(16):1198–1203. DOI: 10.1093/eurheartj/ehs475.

47. Hemminki K, Vaittinen P, Dong C, Easton D. Sibling risks in cancer: clues to recessive or x-linked genes? *British journal of cancer* 2001; **84**(3):388. DOI: 10.1054/bjoc.2000.1585.

48. Sundquist K, Li X, Hemminki K. Familial risks of hospitalization for parkinson's disease in first-degree relatives: a nationwide follow-up study from sweden. *Neurogenetics* 2006; **7**(4):231–237. DOI: 10.1007/s10048-006-0055-z.

49. Sundquist J, Li X, Sundquist K, Hemminki K. Risks of subarachnoid hemorrhage in siblings: a nationwide epidemiological study from sweden. *Neuroepidemiology* 2007; **29**(3-4):178–184. DOI: 10.1159/00011158.

50. Demicheli R, Ardoino I, Boracchi P, Coradini D, Agresti R, Ferraris C, Gennaro M, Hrushesky W, Biganzoli E. Recurrence and mortality according to estrogen receptor status for breast cancer patients undergoing conservative surgery. ipsilateral breast tumour recurrence dynamics provides clues for tumour biology within the residual breast. *BMC cancer* 2010; **10**(1):656. DOI: 10.1186/1471-2407-10-656.

51. Demicheli R, Fornili M, Ambrogi F, Higgins K, Boyd JA, Biganzoli E, Kelsey CR. Recurrence dynamics for non–small-cell lung cancer: Effect of surgery on the development of metastases. *Journal of Thoracic Oncology* 2012; **7**(4):723–730. DOI: 10.1097/JTO.0b013e31824a9022.

52. Escedy J, Hunter D. The Origin of Cancer. In *Textbook of Cancer Epidemiology*, Adami HO, Hunter D, Trichopoulos D, eds. Oxford University Press Inc., 2008; 61–85.

53. Marcovecchio ML, Chiarelli F. Obesity and Growth during Childhood and Puberty. In *Nutrition and Growth*, Shamir R, Turck D, Phillip M, eds., vol. 106. World Review of Nutrition and Dietetics, Basel: Karger, 2013; 135–141. DOI: 10.1159/000342545.

1

2

3

December 16th, 2013, Revision

Understanding variation in disease risk:

The elusive concept of frailty

Odd O. Aalen[a,b,*], Morten Valberg[a], Tom Grotmol[b], Steinar Tretli[b]

[a]Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Norway

[b]Cancer Registry of Norway

*Corresponding author: Odd O. Aalen, Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, POB. 1122, Blindern, N-0317 Oslo, Norway. E-mail: o.o.aalen@medisin.uio.no. Tel.: +4722851236. Fax: +4722851313.

Abstract:

Background

The concept of frailty plays a major role in the statistical field of survival analysis. Frailty variation refers to differences in risk between individuals which goes beyond known or measured risk factors. In other words, frailty variation is unobserved heterogeneity. Although understanding frailty is of interest in its own right, the literature on survival analysis has demonstrated that frailty effects can lead to surprising artefacts in statistical estimation that are important to examine.

Methods

We present literature that demonstrates the presence and significance of frailty variation. We discuss the practical content of frailty, and show the link between frailty and biological concepts like (epi)genetics and heterogeneity in disease risk.

Results

Evidence is pointed out for the pervasive presence of heterogeneity between individuals. There are numerous suggestions in the literature that a good deal of this variation may be due to randomness. Heterogeneity often manifests itself as clustering of cases in families. We emphasize that apparently moderate familial relative risks can only be explained by strong familial variation. Finally, we highlight the potential impact of frailty on standard epidemiological measures.

Conclusions

Frailty variation is normally present and a good understanding of this phenomenon may be important in order to correctly interpret statistical analyses in epidemiological studies. Even moderate familial risk points to a high degree of variation between families and individuals. Statistical artefacts may arise as the result of frailty variation in many settings, and one should be cautious in the interpretation of hazard and incidence rates.

Keywords: Frailty, heterogeneity, random variation, epigenetics

Key messages:

- Variation in risk of disease often goes far beyond what is captured by measured risk factors.
- Much of the heterogeneity in risk may be established early in life, and stochastic variation in these processes could be a major contributor in this regard.
- Even moderate familial risk points to the existence of large variations in risk between families, and individuals, in the population.
- Failing to take into account the unobserved heterogeneity between individuals may lead to erroneous interpretations of standard epidemiological measures, like hazard ratios and age-incidence curves.

Introduction

In epidemiology and clinical science, it is often tacitly assumed that the risk of individuals with respect to a certain disease is similar in a population, apart from well documented differences due to risk factors or genetics. It is often presumed, for instance, that all individuals are vulnerable to the same risk factors. Differences between individuals tend to be ignored unless they can be expressed in terms of known risk factors or known genetic properties. However, the fact of the matter is that individuals are generally highly dissimilar also for a number of unknown, or just partly known, reasons. Indeed, the extent of this heterogeneity is probably not generally appreciated.

Heterogeneity which is unknown, or not represented in available data, is often referred to as frailty, a quantity that is varying between individuals. The term frailty comes from the statistical field of survival analysis, where there is a strong interest in this type of heterogeneity. Frailty is usually modeled by assuming that the hazard rate (baseline hazard) for an "average individual" is multiplied by a factor Z that renders the level for a specific individual:

$$\lambda(t) = Z\alpha(t)$$

When we integrate out the variation in Z to get the (observable) population hazard rate, the resultant function is quite different from the baseline hazard $\alpha(t)$. A number of various distributions exists for the frailty Z.[1, 2] Note that the baseline hazard may be a function of observed individual covariates, e.g. through a Cox model.

A number of diseases exhibit incidence rates that peak at young ages, including cancers like childhood leukemia and Hodgkin lymphoma, but also schizophrenia, which has recently been analyzed from a frailty point of view.[3] In several cases frailty variation is a reasonable explanation for an early peak in incidence, especially when the disease has a strong heritability, which is the case for diseases like schizophrenia and testicular cancer. The frailty approach has yielded particularly fruitful insights for testicular cancer.[4-7] Furthermore, the so-called frailty models form a basis for the analysis of familial association in cancer incidence.[6-8]

One goal of the present paper is to point out the ubiquity of heterogeneity, or varying frailty. We shall also emphasize the role of stochasticity. Furthermore, we will discuss how important indications of frailty variation may be deduced from data on familial association. Indeed, moderate familial association implies surprisingly strong variation in risk between individual families. Although understanding frailty variation is important in its own right, it may also be essential in order to correctly interpret statistical analyses in epidemiological studies. Finally, we will cover an issue that has been pointed out in the statistical literature: that unobserved frailty variation may lead to misleading comparisons of hazard rates and incidence rates, resulting in, among other things, an artificial cross-over effect.

Heterogeneity between individuals may be high

Individual variation in susceptibility

It is often obvious that disease risk is a fluid phenomenon, dependent on risk factors, genes, and the country of residence, among other things. For example, the risk of colon cancer varies widely across different countries; it has increased sharply (in fact more than tripled) in the last

few decades in many industrialized countries, and it varies substantially between different countries worldwide. This means that the risk of colon cancer is not a given quantity, but rather something that varies widely. It would follow that the individual risk of colon cancer is also a fluid phenomenon; that it varies considerably between individuals even when the outer circumstances are similar. Furthermore, there are large differences in risk across regions (Figure 1).[9] Therefore, it seems logical that the risk will not be homogeneous within regions (especially given the arbitrariness of many borders). In short, the variation in risk *between* regions strongly suggests a considerable variation *within* regions. This kind of variation has been clearly demonstrated for the United States with regard to the dependence on race of colorectal cancer incidence.[10] However, it is highly likely that there are other variations based on unknown risk factors.

Some of the biomedical literature has indicated that there is a high degree of variation in individual cancer susceptibility, thereby supporting the presence of frailty variation. An interesting paper is Balmain et al.[11] where they indicate a strong variation in susceptibility to breast cancer. They studied a population without high-risk individuals (those with BRCA1 and BRCA2 mutations), and still suggest a 40-fold difference in the risk of breast cancer between the top 20% and bottom 20% of the study population. Their model also suggests that more than 50% of cancers occur in the 12% of the population that is most susceptible. Peto and Mack[12] reached similar conclusions by studying the relatives of breast cancer patients. They observed a very strong familial risk in monozygotic twins which could only be explained by a large individual variation in risk. They make the following statement: "Our most surprising conclusion is that a high proportion of all breast cancers, and perhaps the majority, arise in women at very high risk." Also, the increased risk of a second breast

malignancy after ductal carcinoma in situ of the breast, even after adjustment for type of treatment, points to a large variation between individuals in susceptibility to this disease.[13]

Regarding colorectal cancer (CRC), Win et al.[14] suggest that "the risk of developing CRC varies approximately 20-fold between the people in the lowest quartile (average 1.25% lifetime risk of CRC) versus the highest quartile for familial risk profile (average 25% risk)". Another study of colorectal cancer in DNA mismatch repair gene mutation carriers showed a U-formed distribution in risk distinguishing a high-risk group from a moderate risk-group.[15]

It is important to be aware that even for cancers with strong attributable risk factors, frailty remains a large component. In lung cancer, for instance, there are strong indications that some people are much more vulnerable to the damage inflicted by smoking than others; with just 10-15% of smokers developing lung cancer.[16]

Susceptibility may be established early in life

The notion of "early life programming" has become popular. This idea was formulated in the Forsdahl-Barker hypothesis, which states that the risk of many diseases is strongly influenced by what happens very early in life, e.g. at or prior to birth. Forsdahl[17] showed a strong correlation between mortality rates for arteriosclerotic heart disease in people aged 40-69 years, and infant mortality in the same birth cohorts within Norwegian counties. Earlier works include Ravelli et al,[18] who observed babies born to women who were pregnant during the Dutch famine. They found that the weight of these children later in life depended on which trimester the famine affected. Kermack et al. found an association between childhood

and adult mortality for different birth cohorts.[19, 20] Barker et al. studied heart disease and found that areas in England that had the highest coronary heart disease mortality in the 1980s also had the highest child mortality rates 70 years earlier.[21, 22] Since then, many papers have been published indicating a relationship between early life conditions and disease risk later in life,[23] including the recent paper by Eriksson et al.[24] who assert in the title that "Boys live dangerously in the womb".

Epigenetics is used to describe heritable changes in gene expression that are not caused by alterations in the nucleotide sequence of the genome, and Forsdahl- Barker type effects have been tied to such epigenetic alterations.[25, 26] Painter et al showed that the children of those exposed to the Dutch famine *in utero* during World War II (WWII) were also at increased risk for ill health, which indicates that epigenetic effects in utero can even have transgenerational consequences.[27] Studies from the Netherlands and Scandinavia have shown a decreased risk of testicular cancer for birth cohorts born during WWII compared to those born before and after.[28, 29]

All these examples show that epigenetic alterations early in life may lead to a large degree of heterogeneity between adult individuals in a population.

Different types of variation

Figure 2 illustrates various ways in which risk can be distributed between individuals. Panel 1 indicates a risk that is quite similar across individuals, with some variation. Panel 2 shows a

situation where most individuals have a relatively similar risk, but there are some individuals who deviate quite a lot (the upper tail). This is expressed even more clearly in panel 3, where many individuals have a risk close to zero, while there are a number of individuals with high risk. An even stronger variation is illustrated in panel 4, where most individuals have risk close to zero, while some individuals have a very high risk. All these types of variation could actually occur. The examples given in this paper show that even the types of variation shown in panels 3 and 4 could be common. However, another issue is how risk develops over time. One view is that there is a rather small variation in risk at an early age, which increases over time as the result of the varying stresses of life. An alternative view argues that much of the variation in risk between individuals is determined very early in life, maybe even prior to birth.

Heterogeneity may be due to stochastic processes

Randomness

Heterogeneity, or varying frailty, between individuals may have a number of different explanations: environment, genetics or epigenetics; or it may be a purely stochastic phenomenon. There is a growing recognition in biology that both stochastic variation and chaotic variation are important[30].

Many studies point to large individual differences that do not have obvious explanations. A recent paper points out a strong association between the telomere length of finches at 25 days and later mortality over several years.[31] Kirkwood and Finch[32] show that even genetically identical (i.e. isogenic) worms have great variation in their lifetimes. They stress the random

and unpredictable nature of cell damage that occurs with ageing. Epigenetic factors are also likely contributors to these time-dependent processes.

Le and Cheng[33] studied the problem of why genetically identical cells in the body vary widely in their storage of fat, even when there is no difference in the expression of the particular genes that affect this storage. They found that the differences between cells were due to variation occurring in a cascade of events within an insulin-signaling pathway. These variations were very slight at the beginning of the cascade, but led to very different results at the end. This phenomenon of small variations in starting conditions yielding very big differences in the end product is well known in mathematics and is related to nonlinear equations; comprising essential elements of chaos theory. Nonlinearity is probably a common phenomenon in biology, and cascade phenomena would be expected to be nonlinear with complex feedback dynamics.

In an interesting Nature letter, Frank and Nowak[34] suggest a model where random mutations at a very young age can produce a developmental disposition to cancer. The idea is that during the gestational phase stem cells may mutate and then multiply randomly. If the mutation rate is high enough, this variation could be a dominant feature.[35, 36]

The examples given here show that great individual variation may simply be an accumulation of purely random variation combined with nonlinear dynamics.

Smith offered a fascinating discussion of the importance of randomness in epidemiology.[37] He asserts that epidemiology cannot capture the pervasive randomness which averages out at the population level. Our point here, though, is that when time is considered, there are telltale indications of random variation.

Genetic variation and rare variants

The lack of clear findings in many genome wide association (GWA) studies has led people to think that rare mutations might be responsible for many diseases. Rare gene combinations are difficult to discover in GWA studies, which may explain the apparent lack of genetic effects.[38, 39] Also, Fletcher and Houlston[40] explain how disease susceptibility may be an effect of common low-penetrance genes or rare gene combinations. Cirulli and Goldstein[41] suggest that "rare variants could be the primary drivers of common diseases" and state that e.g. rare copy number variants "are associated with an effect on the risk of disease that dramatically exceeds the effects of most common variants associated with a disease".

While a simple polygenic model, where the risk is a linear combination of a (possibly large) number of factors, will give a normally distributed risk like that in panel 1 of Figure 2, rare variants make it more likely that we will get variations of the type seen in panels 3 and 4.

Epigenetic stochasticity

During recent years, there has been growing recognition that environmental exposure affects cancer susceptibility through epigenetic changes, in addition to the traditional gene-environment interactions, which can promote mutations. This is particularly relevant in the developmental origins of health and disease hypothesis.[42] Some authors argue for a paradigm

shift, where the old view on the importance of DNA mutations is weighed down and supplemented by the modern view of epigenetic modifications.[43] There is, however, indication of an important stochastic component to these modifications, and it has even been speculated that the majority of important epigenetic changes may not be due to the environment, but to random events early in life.[43] This might explain the large variation that is often observed between genetically identical individuals.[44]

The competing explanations: frailty selection versus biological mechanism

Frailty explanations of observed incidence rates will typically attribute certain findings to statistical selection effects. A disease where frailty is likely to play a role is testicular cancer. The incidence of this disease is typical of cancer forms originating in early (fetal) life, reaching a peak at a rather young age (approx. 30 years) and then declining sharply. A reasonable explanation for this observation is that some men are susceptible to, and have an increased risk of acquiring testicular cancer, and do so relatively early. The declining incidence of testicular cancer with age is presumed to be due to high-risk individuals being selected out from the population after they acquire the disease. This in fact fits well with biological evidence suggesting that testicular cancer may be caused by cellular damage during fetal life, which has been used as a basis for a so-called frailty analysis of incidence.[4] In fact, the origin of testicular cancer is believed to be carcinoma in situ cells, the malignant transformation of which is initiated during early development from primordial germ cells, or gonocytes that either fail to end their proliferation or undergo proper differentiation.[45] Since the incidence rate of testicular cancer also has increased substantially during the last decades,

this damage appears to have become more prevalent over time. It should be noted that this kind of a statistical explanation typically competes with a biological mechanistic one. It has also been suggested that the decline in the risk of testicular cancer with age could be due to a declining testosterone level. Although the surge in testosterone level during puberty is important for the transformation of dormant carcinoma in situ cells to invasive testicular cancer, there is no evidence that individual testosterone level is a risk factor for testicular cancer.[46] Furthermore, the decline in testosterone is rather modest from the age of 30 years.

On the other hand, there are clearly cases where frailty is not the major cause of the decline in risk. One example is retinoblastoma, where there are almost no cases in individuals over 10 years of age. The likely explanation is that the retinoblasts are fully differentiated at the age of 10, and thus thereafter are not susceptible to malignant transformation.[47] However, in his seminal study on retinoblastoma, Knudson actually took varying frailty into account.[48] Long before the Rb1 gene was identified, he separated a very frail group (those with an inherited germ line mutation) from a less frail group (those who had the non-hereditary form), and used this to formulate his famous two-hit hypothesis. The case of retinoblastoma is thus an example of how the consideration of varying frailty combined with biological knowledge may provide valuable insights.

Competing frailty and biologic mechanistic explanations are often suggested, and it may not be obvious which one is correct. Part of the difficulty is that when frailty is estimated from single event data (e.g. the single occurrence of a specific type of cancer for an individual), there will necessarily be uncertainty. A much more precise assessment of frailty can be done

in a setting where there are repeated events (e.g. cancer in both breasts, or in the kidneys or testicles), or when studying cancer incidence in families, e.g. testicular cancer among brothers.[6, 7]

Familial cancer risk points to large individual heterogeneity

For many diseases there is a familial association in risk. A surprising and counterintuitive issue is that even a moderate familial association points towards a large variation in risk between families. Hence, the existence of a familial association is another argument for the presence of considerable individual heterogeneity in risk.

There is generally a familial association when it comes to cancer risk. For example, in breast cancer the genes BRCA1 and BRCA2 confer a very high familial risk. But even in the absence of such genes, sizeable familial association is still observed. Johns and Houlston[49] pointed out that having a first-degree relative with colorectal cancer more than doubles one's risk for the disease, while the risk is increased more than 4-fold when one has two first-degree relatives with CRC. The risk of testicular cancer for a brother of a case is increased about 6-fold.[6, 7] Tumors of the nervous-system also show a strong heritability (standardized incidence ratios around 2, but up to 27 for the rare multiplex families).[50]

Even familial risks that appear modest, like the relative risk of about 2 seen for relatives of breast or colon cancer patients, still imply a large variation in risk between individuals.[50-52] This has also been pointed out by Moger et al.,[7] and by Aalen[53] in a cardiovascular disease setting. In fact the variation in individual risk when even small familial risks are observed will

typically be of the type in panels 3 and 4 of Figure 2. An interesting quote from Hopper[51] stresses this surprising fact:

> "… even for a disease for which there is only what one might consider in epidemiological terms 'modest' familial aggregation (such as a two-fold increased risk for close relatives of affecteds), people of the same age and sex must differ greatly in their familial risks of disease (e.g. a 20-fold or more difference in risk between the quarter of the population at lowest familial risk and the quarter of the population at greatest familial risk). This familial risk gradient is in addition to differences due to 'non-familial' environmental or lifestyle factors that are specific to individuals. Finding the causes of even a modest proportion of familial aggregation of a disease could be a major step in understanding the causes of the disease itself."

Let us consider a very simple situation: Assume that the population is divided into two groups of equal size, and such that the probability of acquiring a specific disease is 1% in one group and 20% in the other. All the members of a given family belong to the same group, be it the high-risk or low-risk group. Consider that the familial relative risk is defined as follows: the conditional probability of developing the disease if a specific family member has acquired it, divided by the average risk of getting the disease. In our example the familial relative risk equals 1.82. Hence a relative risk of 20 at the individual level translates into a very modest familial risk just as suggested by Hopper.[51]

Since familial relationships are important for disease risk, it is useful to use study designs that to some extent control for such relationships. Within-pair twin studies are useful in this regard.[54]

Statistical models for familial risk

In order to get a deeper understanding one has to consider statistical models. The familial risk association depends on two conditions, namely the correlation between the risk factors within a family, and the variation in risk within the population associated with these factors. Assume that the risk depends exponentially on normally distributed risk factors with a correlation ρ, and that s denotes the relative risk associated with a change in the risk factor from mean – 2SD to mean + 2SD. The familial relative risk, r, associated with a diseased sibling is given by

$$r = \exp\{\rho(\ln s)^2/16\}, \quad (1)$$

a special case of a more general formula given by Aalen.[53] Assume for instance that $\rho = 0.5$ which is a very strong familial correlation. Then formula (1) as a function of s is plotted in Figure 3. One sees that even for $s = 10$ which represent a very strong effect of the risk factor, the value of r is still less than 1.2. Hence, for simple polygenetic inheritance at the risk factor level, the familial relative risk associated with even strong risk factors is very moderate.

In practice, familial association will have several sources, partly genetic, and partly a shared environment or culture, or attitude toward various risk behaviors. It can be shown that environmental influences contribute only very slightly to the observed familial risk association. However, measured risk factors could be poor surrogates for risk factors that are more strongly familial, and the effect could be somewhat prone to e.g. measurement error. [52]

Formula (1) presumes a normal distribution, which one would usually assume for simple polygenetic inheritance. Some skewness might be introduced, which might appear more realistic if some genes have a stronger effect than others, for example due to higher penetrance. We shall assume that two individuals have a common risk component which is gamma distributed with shape parameter δ. Following Aalen[53], the familial relative risk, r_F, is given by:

$$r_F = \left\{1 + \frac{\ln r}{\delta - 2(\delta \ln r)^{1/2}}\right\}^{\delta} \quad (2)$$

Note that when the shape parameter δ goes to infinity, this expression will converge to r (because an infinite δ implies a normal distribution for the common component). Plots of formula (2) as a function of δ and r are given in Figure 4. The major deviation occurs for $\delta = 1$ which corresponds to an exponential distribution of the common familial risk. This represents a high degree of skewness (Figure 5). It means that members of a minority of families have a much higher risk than others. Figure 5 also includes an illustration of an even more skewed gamma distribution.

Similar results are presented in Moger et al.[7] where a totally different mathematical model also indicated that even a very skewed familial frailty distribution will result in very moderate familial relative risks. The paper presents the following useful formula

$$R = CV^2 + 1$$

where CV is the coefficient of variation of the probability of being susceptible, as it varies between families, and R is the relative risk of another member of the family acquiring the disease if there is already a case in the family. From the above formula it is seen that

assuming e.g. $R = 2$ implies $CV = 1$. This means that the standard deviation equals the expectation, which again implies a strongly skewed distribution. If the distribution comes from the gamma family, then it has to be an exponential distribution. If CV is greater than 1, then the shape parameter of the gamma distribution is less than 1 which yields an extremely skewed distribution (Figure 5). In fact the cases discussed here correspond to panels 3 and 4 of Figure 2.

The conclusion from this brief review of familial association is that a familial relative risk of 2 or above is a strong indication for the existence of high risk groups of individuals.

Interpretation of epidemiological measures

Taking heterogeneity, or varying frailty, between individuals into account can be of crucial importance for the understanding of epidemiological features in a population. There is a natural tendency to assume that hazard rates and incidence rates can be taken at face value. Although these concepts appear to be simple, their interpretation can still be very difficult. The statistical interest in frailty stems in part from the fact that it can lead to curious statistical artefacts.

Cross-over effects

Consider two groups of individuals with hazard rates $\alpha(t)$ and $2\alpha(t)$, such that the hazard ratio is 2. In each of these groups there would necessarily be some unobserved heterogeneity between individuals. By introducing equally distributed frailty variables in the two groups, a decreasing hazard ratio over time may be obtained. Depending on the choice of frailty

distribution, the hazard ratio may even cross-over, and become lower than 1, such that the high-risk group appears to become the low-risk group (Figure 6). The decrease (and possible cross-over) of the hazard ratio over time is a frailty effect. Individuals in the high-risk group will on average experience events earlier than those in the low-risk group. This causes the proportion of highly susceptible individuals in the high-risk group to decrease faster than in the low-risk group, leaving an increasing proportion of less susceptible individuals. Thus, the hazard ratio will decrease. If, for instance, the population contains a non-susceptible subgroup, then the susceptible individuals in the high-risk group would be exhausted earlier than in the low-risk group, causing the relative risk to cross-over, and become lower than one, even if the hazard ratio stays constant on the individual level. This means that when frailty is not observed and cannot be accounted for, a wrong conclusion could be drawn regarding the true relationship between two groups. This is in fact a time-dependent version of Simpson's paradox, which means that the observed relationship (concerning e.g. risk of disease) between two groups is reversed at an aggregate level compared to what would be observed at a more detailed level if covariates could be conditioned on.

Another interesting effect of frailty occurs when discontinuing treatment in a clinical trial. Let us assume that the treatment group has hazard rate $\alpha(t)$ and the control group has hazard rate $2\alpha(t)$, presuming the treatment is effective. At the start of the study, the hazard ratio is two. Because the treatment is effective, patients in the control group will on average have events earlier than in the treatment group, and the hazard ratio will decrease with time. At some point the difference between the hazard rates is so small that it is decided treatment is no longer effective, and it is stopped. A possible consequence of this decision is that the hazard ratio drops below one, and it appears protective to be a member of the control group (Figure

7). In the control group the most frail individuals would already have had an event at this point, and at the time of discontinuing the treatment, there would be a higher proportion of less frail individuals in the control group. In the treatment group, frail individuals that would already have had an event if they had not been treated, have a very high risk immediately after the treatment is stopped. Not being aware of a possible frailty effect may lead to a wrong impression of the effect of discontinuing a treatment for an individual.

False protectivity

In a competing risks framework, two (or more) events compete in determining the failure of an individual. The failure rate of each cause is expressed in terms of a cause-specific hazard rate. As in the example above, the hazard rates may be influenced by frailties. If these frailties are correlated, then one may observe a *false protectivity*.[55] If a covariate has a detrimental effect on one of the two competing risks, it may, at the population level, appear to be protective in the other cause-specific hazard rate.

Frailty and models of carcinogenesis

The famous multi-stage model of Armitage and Doll set the stage for a mathematical approach to understanding cancer incidence, and it continues to play a fundamental role in our understanding of the carcinogenic process. This was exemplified by the re-publication of the original article in the *International Journal of Epidemiology* at its 50 year anniversary in 2004. Since it was first suggested, however, more sophisticated models have been published. Moolgavkar and Knudson proposed a two-hit model (combined with clonal expansion of initiated cells) that took heterogeneity of the carcinogenic process itself into account, and explained peaking incidence rates of certain cancers by the varying (decreasing) number of

stem cells susceptible to mutation. Their model has later been expanded to allow for a cell to undergo several transitions before going into the clonal expansion phase,[56] as well as other further developments of the model.[57, 58] All these models were created to facilitate the understanding of cancer development on an individual level. Meza et al. studied the effect of gestational mutations on cancer risk, and stated that "Even with identical gestational mutation rates in all individuals in a population, at birth individuals are at different risk because of random variation in the number of mutated cells at birth".[35] Taking varying frailty into account (i.e. heterogeneity in risk between individuals), a Weibull hazard rate, as suggested by the Armitage-Doll model, makes sense. This is the case, even when assuming clonal expansion at some intermediate step. A mathematical formulation is that, on the individual level, the hazard rate of an event is given as a product of the Weibull hazard and an individual frailty factor. As opposed to the exhaustion of susceptible stem cells within the individual, the model considers the exhaustion of initially highly susceptible individuals as an explanation of a peaking incidence curve. This approach may also be modified in several ways, including taking into account an expanding host tissue during e.g. puberty.[59] An important element is thus to combine models of carcinogenesis with a realistic understanding of individual differences,[5, 59, 60] to better understand features of population age-incidence rates.

Interpretation of incidence rates

It turns out that also changes in epidemiological incidence rates over calendar time can be wrongly interpreted if one does not take into consideration the possible heterogeneity between individuals. Consider the simple Armitage-Doll multistage model of carcinogenesis, which states that a cell has to go through a certain number of transitions to reach malignancy. As an example, consider the simple version of a multi-stage model as shown in Figure 8. Assume

that the transition rates are not the same for all individuals, but that there is a strong variation in susceptibility. Consider for instance a population where only a small subgroup is susceptible to the cancer in question, and the majority has a zero rate of cancer initiation (transition from a normal cell to an intermediate cell). If the initiation rate increases abruptly at a given point in (calendar) time, the incidence rate may increase to a peak, then drop and stabilize on a higher level. This is illustrated in Figure 9a, for the simple multistage model in Figure 8, with only 1% of the population being susceptible. The same point, with 90% being susceptible, is illustrated in Figure 9b. Although a simplification, the abrupt increase in the initiation rate could be the result of a risk factor that becomes more pronounced in the population at a given time.

The above example is simple, but illustrate that changes in the prevalence of risk factors may have an impact on observed incidence rates, even a long time after the change occurred. While the real biological change here was an abrupt increase in the prevalence of a risk factor, the observed incidence rates gave the impression of a risk that first increased and then decreased. It is of course more likely that the presence of risk factors changes gradually over time, and this will have a similar effect on observed incidence rates as in the above example. The observed incidence rate will continue to change after the prevalence of the risk factor has stabilized. If a cell requires more events to become malignant, changes in prevalence of different risk factors may affect the transition rates to various states. This could possibly also lead to multimodal shapes of hazard rates.

The point we are making, is that changes in the incidence rate may not be a simple reflection of what is happening at a biological level. It is well known that underlying effects will be smoothed out in the observed incidence. But in addition to this, frailty may produce incidence rates with aspects that are unrepresentative of the underlying changes. Care should be taken before drawing conclusions on an individual level based on observations in a population.

Conclusion

We have pointed out a number of findings that indicate the presence of a considerable individual variation in the risk of cancer and other diseases that goes beyond what is due to measured risk factors. Varying frailty may create artefacts when studying incidence rates and other epidemiological measures, such as a decline in incidence due to the frailest individuals experiencing the event early.

Familial associations that appear moderate may cover a large underlying variation in risk between individuals. This and other aspects of individual variation point towards caution in interpretation. The presence of individual heterogeneity cannot be ignored. It may be necessary to perform mathematical modeling to get a proper understanding of the nature and magnitude of the phenomenon of frailty in any given study population.

Funding

This work was partially supported by a grant from the Norwegian Research Council (191460/V50), and by the Norwegian Cancer Society, project/grant number 171851.

References

1. Hougaard P. *Analysis of Multivariate Survival Data*. New York, NY: Springer-Verlag, 2000.
2. Aalen OO, Borgan Ø, Gjessing HK. *Survival and Event History Analysis: A process Point of View*. New York, NY: Springer, 2008.
3. Svensson E, Rogvin M, Hultman CM, Reichborn-Kjennerud T, Sandin S, Moger TA. Schizophrenia susceptibility and age of diagnosis - A frailty approach. *Schizophr Res* 2013; **147:** 140-6.
4. Aalen OO, Tretli S. Analyzing incidence of testis cancer by means of a frailty model. *Cancer Causes Control* 1999; **10:** 285-92.
5. Moger TA, Aalen OO, Halvorsen TO, Storm HH, Tretli S. Frailty modelling of testicular cancer incidence using Scandinavian data. *Biostatistics* 2004; **5:** 1-14.
6. Valberg M, Grotmol T, Tretli S, Veierod MB, Moger TA, Aalen OO. A Hierarchical Frailty Model for Familial Testicular Germ-Cell Tumors. *Am J Epidemiol* 2013. [Epub ahead of print].
7. Moger TA, Aalen OO, Heimdal K, Gjessing HK. Analysis of testicular cancer data using a frailty model with familial dependence. *Stat Med* 2004; **23:** 617-32.
8. Moger TA, Haugen M, Yip BH, Gjessing HK, Borgan Ø. A hierarchical frailty model applied to two-generation melanoma data. *Lifetime Data Anal* 2011; **17:** 445-460.
9. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. *GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10*. [Internet]. Lyon, France: International Agency for Research on Cancer, 2010. http://globocan.iarc.fr (18 June 2013, date last accessed).
10. Ollberding NJ, Nomura AM, Wilkens LR, Henderson BE, Kolonel LN. Racial/ethnic differences in colorectal cancer risk: the multiethnic cohort study. *Int J Cancer* 2011; **129:** 1899-906.
11. Balmain A, Gray J, Ponder B. The genetics and genomics of cancer. *Nat Genet* 2003; **33 Suppl:** 238-44.
12. Peto J, Mack TM. High constant incidence in twins and other relatives of women with breast cancer. *Nat Genet* 2000; **26:** 411-4.
13. Falk RS, Hofvind S, Skaane P, Haldorsen T. Second events following ductal carcinoma in situ of the breast: a register-based cohort study. *Breast Cancer Res Treat* 2011; **129:** 929-38.
14. Win AK, Macinnis RJ, Hopper JL, Jenkins MA. Risk prediction models for colorectal cancer: a review. *Cancer Epidemiol Biomarkers Prev* 2012; **21:** 398-410.
15. Dowty JG, Win AK, Buchanan DD*, et al.* Cancer risks for MLH1 and MSH2 mutation carriers. *Hum Mutat* 2013; **34:** 490-7.
16. Landvik NE, Zienolddiny S, Skaug V, Haugen A. Gene variants and lung cancer risk. *BMC Proceedings* 2010; **4(Suppl 2):**.
17. Forsdahl A. Are poor living conditions in childhood and adolescence an important risk factor for arteriosclerotic heart disease? *Br J Prev Soc Med* 1977; **31:** 91-5.
18. Ravelli GP, Stein ZA, Susser MW. Obesity in young men after famine exposure in utero and early infancy. *N Engl J Med* 1976; **295:** 349-53.
19. Kermack WO, McKendrick AG, McKinlay PL. Death-rates in Great Britain and Sweden: Expression of Specific Mortality Rates as Products of Two Factors, and some Consequences thereof. *J Hyg (Lond)* 1934; **34:** 433-57.

20. Smith GD, Kuh D. Commentary: William Ogilvy Kermack and the childhood origins of adult health and disease. *Int J Epidemiol* 2001; **30:** 696-703.
21. Barker DJ, Winter PD, Osmond C, Margetts B, Simmonds SJ. Weight in infancy and death from ischaemic heart disease. *Lancet* 1989; **2:** 577-80.
22. Barker DJ, Osmond C. Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. *Lancet* 1986; **1:** 1077-81.
23. van der Pols JC, Bain C, Gunnell D, Smith GD, Frobisher C, Martin RM. Childhood dairy intake and adult cancer risk: 65-y follow-up of the Boyd Orr cohort. *Am J Clin Nutr* 2007; **86:** 1722-9.
24. Eriksson JG, Kajantie E, Osmond C, Thornburg K, Barker DJ. Boys live dangerously in the womb. *Am J Hum Biol* 2010; **22:** 330-5.
25. Thompson RF, Einstein FH. Epigenetic basis for fetal origins of age-related disease. *J Womens Health (Larchmt)* 2010; **19:** 581-7.
26. Pearson H. Study of a lifetime. *Nature* 2011; **471:** 20-4.
27. Painter RC, Osmond C, Gluckman P, Hanson M, Phillips DI, Roseboom TJ. Transgenerational effects of prenatal exposure to the Dutch famine on neonatal adiposity and health in later life. *BJOG* 2008; **115:** 1243-9.
28. Verhoeven R, Houterman S, Kiemeney B, Koldewijn E, Coebergh JW. Testicular cancer: marked birth cohort effects on incidence and a decline in mortality in southern Netherlands since 1970. *Int J Cancer* 2008; **122:** 639-42.
29. Bergström R, Adami HO, Mohner M*, et al.* Increase in testicular cancer incidence in six European countries: a birth cohort phenomenon. *J Natl Cancer Inst* 1996; **88:** 727-33.
30. Strogatz SH. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering (Studies in Nonlinearity)*. Cambridge, MA: Perseus Books Group, 2001.
31. Heidinger BJ, Blount JD, Boner W, Griffiths K, Metcalfe NB, Monaghan P. Telomere length in early life predicts lifespan. *Proc Natl Acad Sci U S A* 2012; **109:** 1743-8.
32. Kirkwood TB, Finch CE. The old worm turns more slowly. *Nature* 2002; **419:** 794-95.
33. Le TT, Cheng JX. Single-cell profiling reveals the origin of phenotypic variability in adipogenesis. *PLoS One* 2009; **4:** e5189.
34. Frank SA, Nowak MA. Developmental predisposition to cancer. *Nature* 2003; **422:** 494.
35. Meza R, Luebeck EG, Moolgavkar SH. Gestational mutations and carcinogenesis. *Math Biosci* 2005; **197:** 188-210.
36. Frank SA. Evolution in health and medicine Sackler colloquium: Somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. *Proc Natl Acad Sci USA* 2010; **107 Suppl 1:** 1725-30.
37. Smith GD. Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. *Int J Epidemiol* 2011; **40:** 537-62.
38. Maher B. Hiding place for missing heritability uncovered. *Nature* 2010. doi: 10.1038/news.2010.33
39. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol* 2010; **8:** e1000294.
40. Fletcher O, Houlston RS. Architecture of inherited susceptibility to common cancer. *Nat Rev Cancer* 2010; **10:** 353-61.
41. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010; **11:** 415-25.
42. Walker CL, Ho SM. Developmental reprogramming of cancer susceptibility. *Nat Rev Cancer* 2012; **12:** 479-86.
43. Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* 2010; **465:** 721-7.
44. Finch CE, Kirkwood TBL. *Chance, Development, and Aging*. New York, NY: Oxford University Press, 2000.

45. McGlynn KA, Cook MB. Etiologic factors in testicular germ-cell tumors. *Future Oncol* 2009; **5:** 1389-402.
46. Swerdlow AJ, Huttly SR, Smith PG. Testis cancer: post-natal hormonal factors, sexual behaviour and fertility. *Int J Cancer* 1989; **43:** 549-53.
47. Chial H. Tumor Suppressor (TS) Genes and the Two-Hit Hypothesis. *Nature Education* 2008; **1:** 177.
48. Knudson AG. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA* 1971; **68:** 820-3.
49. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. *Am J Gastroenterol* 2001; **96:** 2992-3003.
50. Hemminki K, Tretli S, Sundquist J, Johannesen TB, Granström C. Familial risks in nervous-system tumours: a histology-specific analysis from Sweden and Norway. *Lancet Oncol* 2009; **10:** 481-8.
51. Hopper JL. Disease-specific prospective family study cohorts enriched for familial risk. *Epidemiol Perspect Innov* 2011; **8:** 2.
52. 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.
53. Aalen OO. Modelling the influence of risk-factors on familial aggregation of disease. *Biometrics* 1991; **47:** 933-45.
54. Stone J, Dite GS, Giles GG, Cawson J, English DR, Hopper JL. Inference about causation from examination of familial confounding: application to longitudinal twin data on mammographic density measures that predict breast cancer risk. *Cancer Epidemiol Biomarkers Prev* 2012; **21:** 1149-55.
55. Di Serio C. The protective impact of a covariate on competing failures with an example from a bone marrow transplantation study. *Lifetime Data Anal* 1997; **3:** 99-122.
56. Luebeck EG, Moolgavkar SH. Multistage carcinogenesis and the incidence of colorectal cancer. *Proc Natl Acad Sci U S A* 2002; **99:** 15095-100.
57. Jeon J, Meza R, Moolgavkar SH, Luebeck EG. Evaluation of screening strategies for pre-malignant lesions using a biomathematical approach. *Math Biosci* 2008; **213:** 56-70.
58. Meza R, Jeon J, Moolgavkar SH, Luebeck EG. Age-specific incidence of cancer: Phases, transitions, and biological implications. *Proc Natl Acad Sci USA* 2008; **105:** 16284-9.
59. Valberg M, Grotmol T, Tretli S, Veierød MB, Devesa SS, Aalen OO. Frailty modeling of age-incidence curves of osteosarcoma and Ewing sarcoma among individuals younger than 40 years. *Stat Med* 2012; **31:** 3731-47.
60. Haugen M, Bray F, Grotmol T, Tretli S, Aalen OO, Moger TA. Frailty modeling of bimodal age-incidence curves of nasopharyngeal carcinoma in low-risk populations. *Biostatistics* 2009; **10:** 501-14.

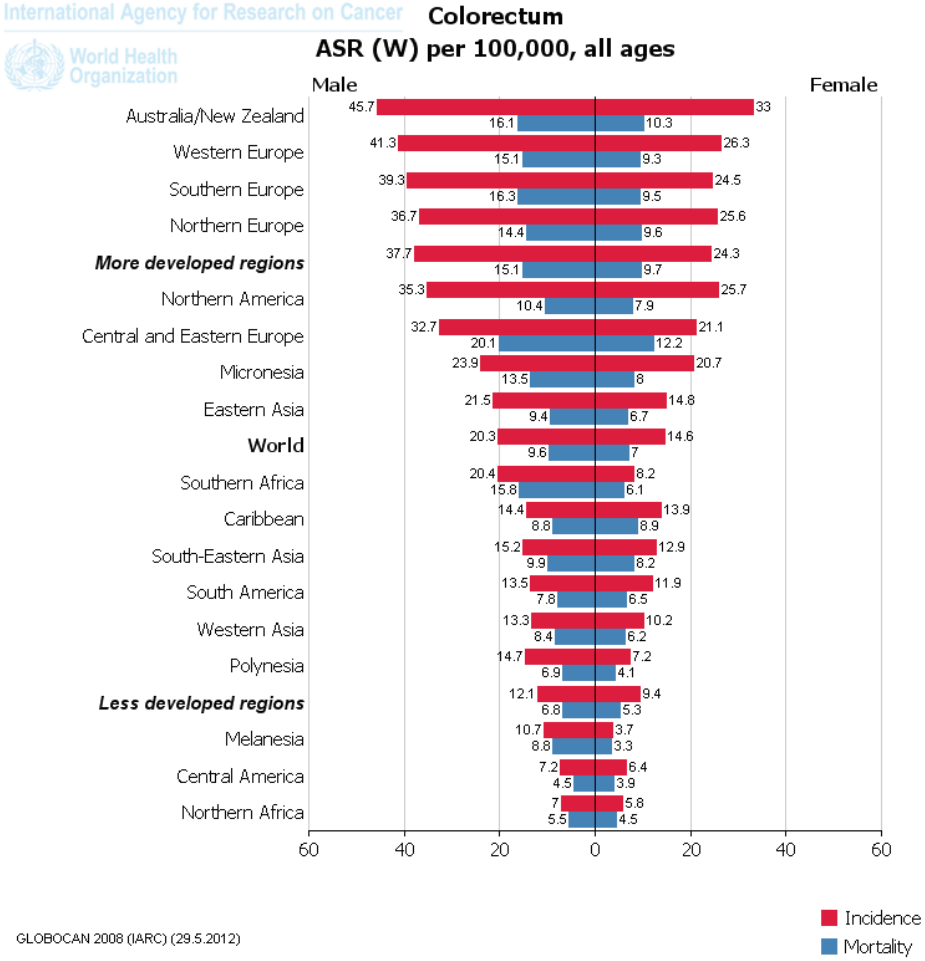

Figure 1: Colorectal cancer incidence in various regions. Picture constructed by Globocan.[9]

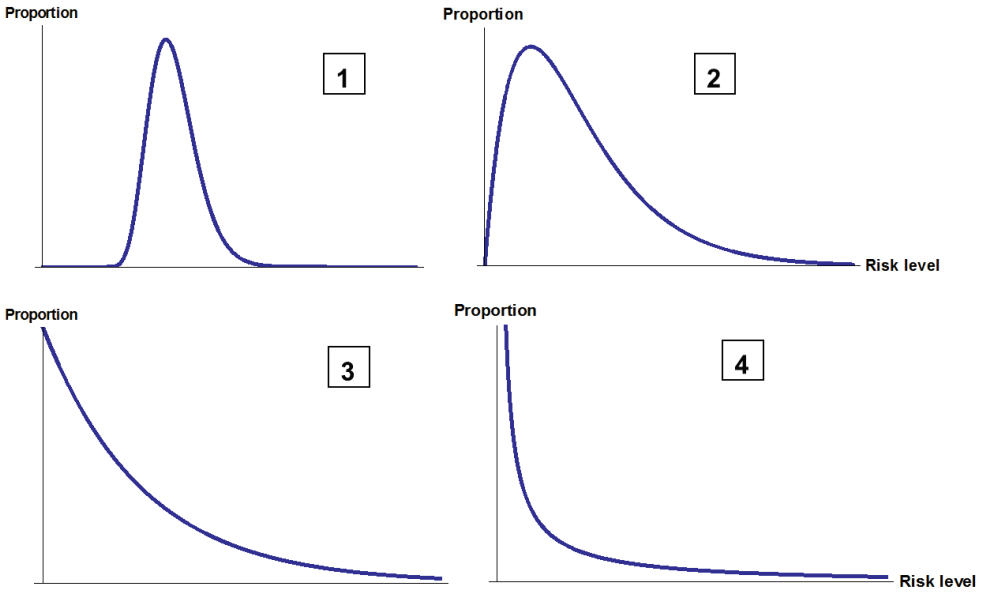

Figure 2: Various types of distributions of risk (frailty) at an early age: (1) small variation in risk between individuals, (2) large group at moderate risk, and a smaller group of individuals at high risk, (3) very skewed: many individuals at low risk and a small group at high risk, (4) most individuals at close to zero risk and a few individuals at a high risk.

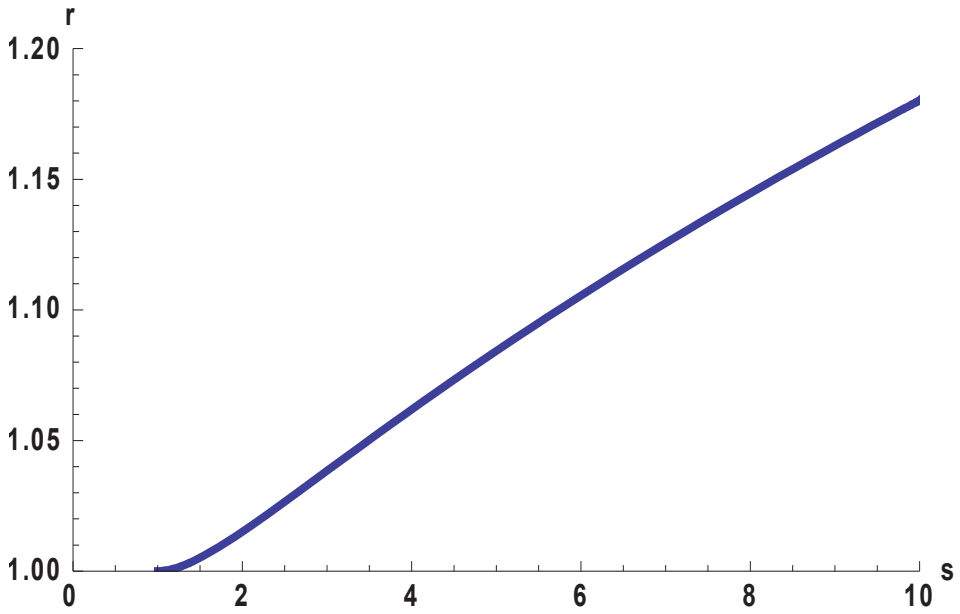

Figure 3: The familial relative risk, *r*, associated with a diseased sibling as a function of *s* according to formula (1) in the text, where *s* denotes the relative risk associated with a change in a risk factor from mean minus two standard deviations to mean plus two standard deviations. Based on normally distributed variation in risk.

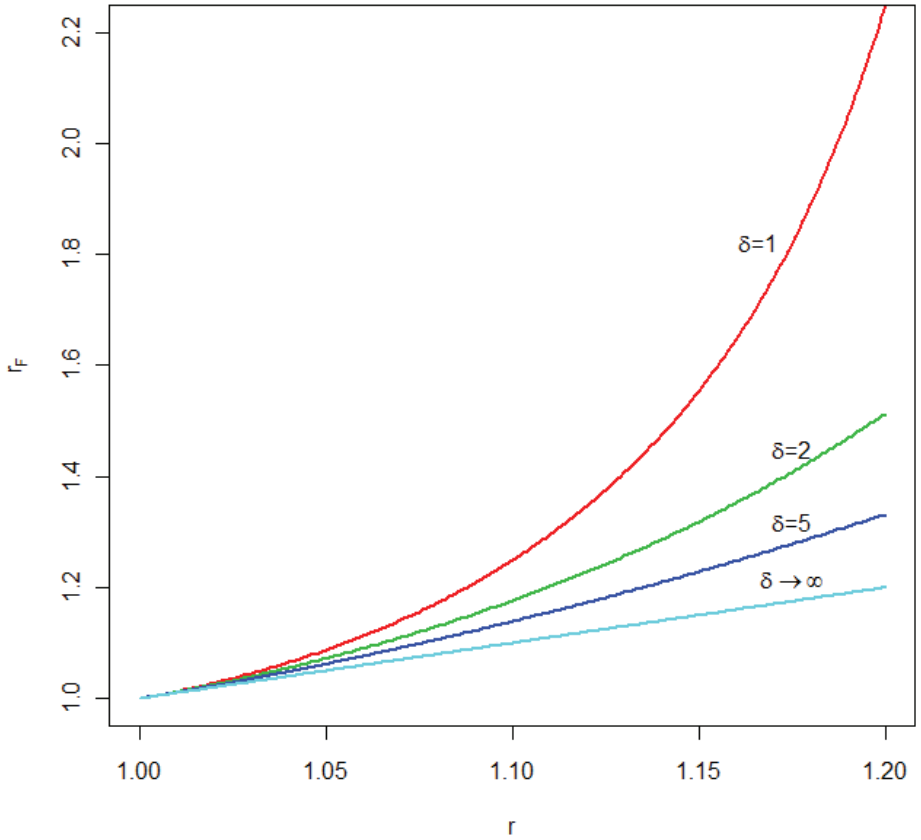

Figure 4: The familial relative risk, r_F, associated with a diseased sibling as a function of *r* according to formula (2) in the text, where the two individuals in the family have a common risk component that is gamma distributed with shape parameter δ. Here *r* is the familial relative risk from formula (1), that is the familial relative risk without the skewness introduced by the common, gamma distributed component. r_F is plotted for given values of δ. Note that $\delta \to \infty$ implies $r_F = r$.

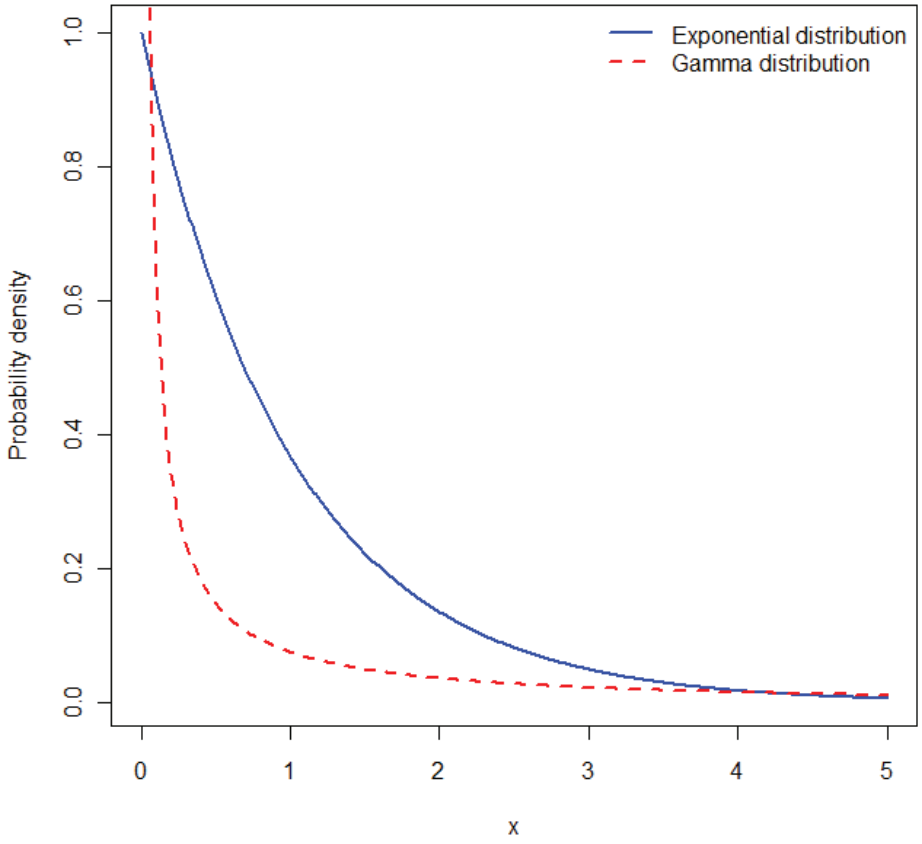

Figure 5: Probability densities for the exponenetial distribution (solid line) and the gamma distribution with shape parameter 0.5 (dashed line).

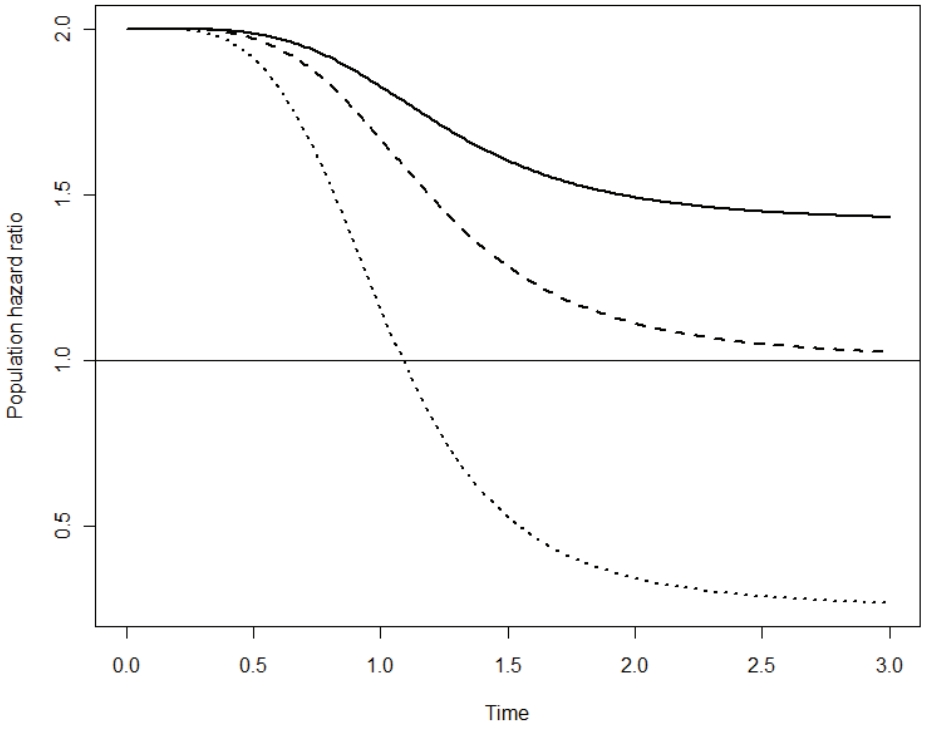

Figure 6: Assume that the hazard rates in two risk groups are $\alpha(t)$ and $2\alpha(t)$ respectively. When frailty variables are introduced, the observed relative risk declines over time as shown in the figure. Three frailty distributions are used; one leads to a crossover of the hazard ratio. This case corresponds to a frailty distribution with a positive probability of zero frailty (i.e. a non-susceptible group). See Aalen et al.[2], Chapter 6, for technical details.

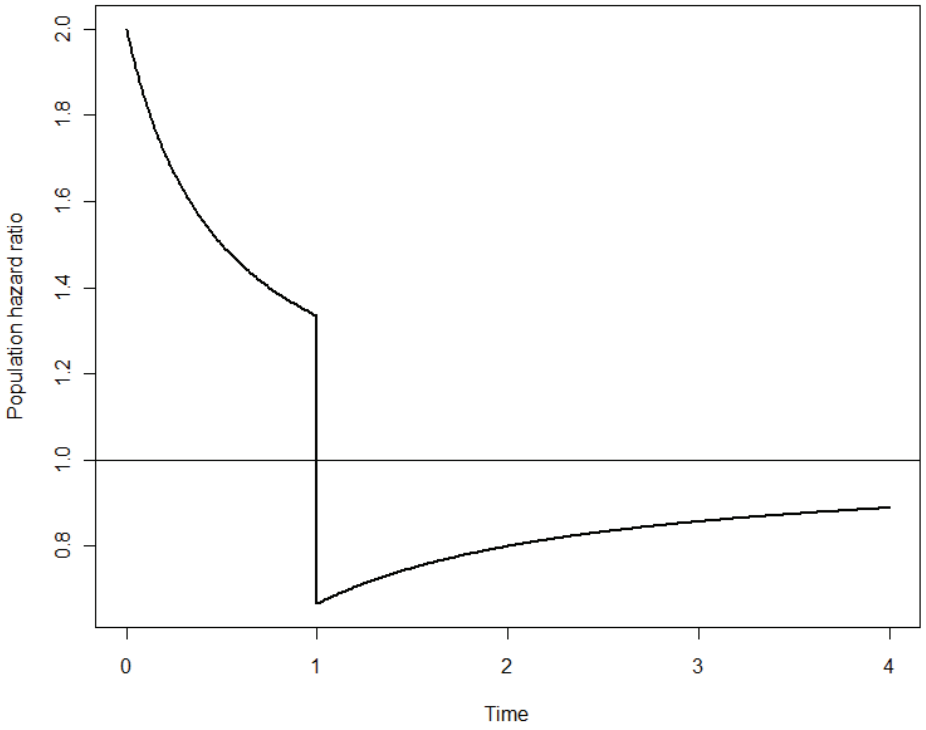

Figure 7: Effect of discontinuing treatment. A control group with hazard rate $2\alpha(t)$ is compared with a treatment group with hazard rate $\alpha(t)$. Treatment is discontinued at time point 1.

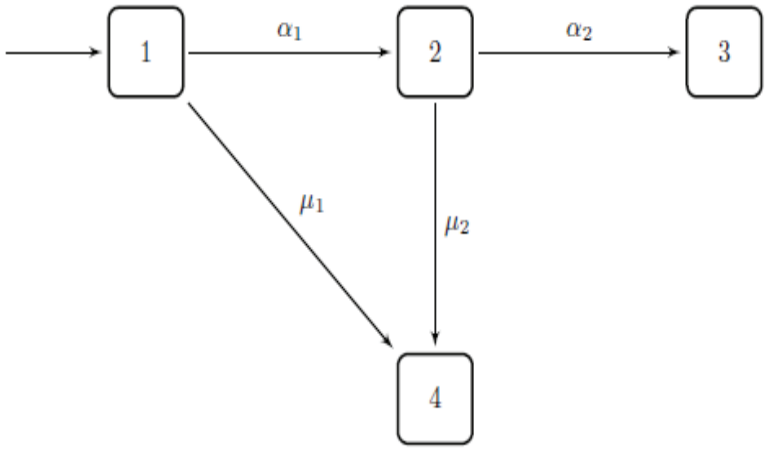

Figure 8: Simple illustration of an Armitage-Doll multi-stage model of carcinogenesis. The states represent the stages of the carcinogenic process. State one is the healthy state, state two is an intermediate state, and in state three a malignant cell has developed. State four is a censored state. The αs and μs are transition rates.

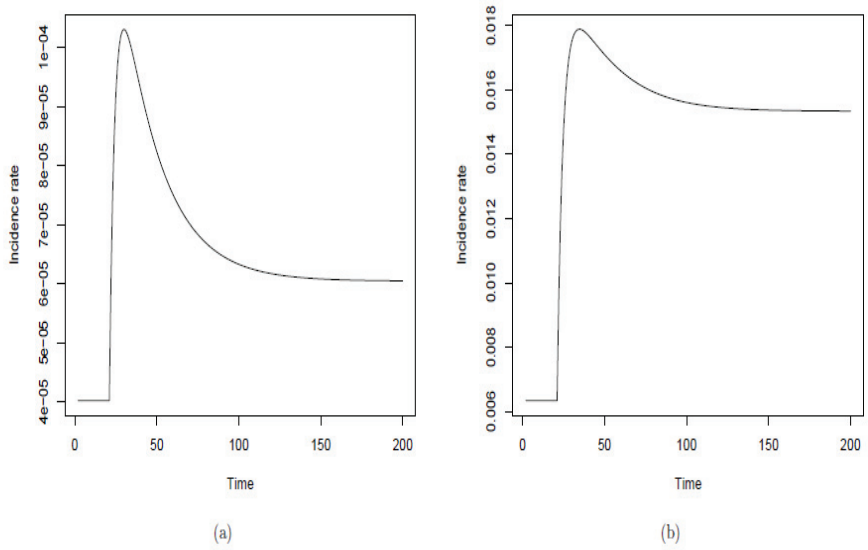

Figure 9: Incidence rates for the model in Figure 8. Assume that 10,000 individuals enter state one per time unit. The transition rates are $\alpha_1 = 0.01$ for time < 20, and $\alpha_1 = 0.03$ for time ≥ 20. Also, $\alpha_2 = 0.02$, $\mu_1 = 0.01$ and $\mu_2 = 0.05$ a) 1% of the population is susceptible, i.e. having $\alpha_1 \neq 0$. b) 90% of the population is susceptible, i.e. having $\alpha_1 \neq 0$.