The Population Incidence of Cancer

Christopher Hornsby

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of the
University of London.

Department of Computer Science
University College London

January 9, 2009

I, **Christopher Hornsby**, confirm that the work presented in this thesis is my own. Where information has be derived from other sources, I confirm that this has been indicated in the thesis.

Dedicated to Rebecca

Abstract

In this thesis stochastic techniques are used in attempts to understand cancer risk, its relationship to patient age and genotype, as well as its distribution in human populations. The starting point for the thesis is the general observation that cancer incidence grows in approximate proportion to an integer power of age. Quasi-mechanistic mathematical models of cancer incidence have suggested that the integer power in a given case is related to the number of crucial cellular events that must occur for a malignant tumour to evolve from a healthy tissue. This idea and its limitations are explored. Further applications of cancer incidence models are then evaluated and developed. Specifically, a critical examination is presented of the notion that increases in risk associated with a particular predisposing germline gene mutation, can provide information about the disease-associated activity of that gene. Finally, there is a discussion of heterogeneity in liability to cancer. Methods for quantifying this heterogeneity and its effect on incidence patterns are investigated.

Acknowledgements

Thank-you to the Medical Research Council for generously extending my funding.

Contents

1 Introduction **25**

1.1 The mechanistic basis of cancer 28

1.1.1 Genetic instability 31

1.2 Structure of this thesis 31

2 Multistage Theory **33**

2.1 Armitage and Doll 33

2.1.1 Derivation of Armitage and Doll's formula 35

2.2 The Hazard Function 37

2.3 The two stage clonal expansion model 39

2.3.1 Derivation of the TSCE model 42

2.4 Likelihood Constructs 51

2.4.1 Population based age-specific incidence 51

2.4.2 Non-population based incidence 52

2.5 Applications of Multistage Modelling 53

2.5.1 Breast Cancer and Clemmesen's Hook 54

2.5.2 Declining incidence in old age 55

2.5.3 Smoking and Lung Cancer 56

2.5.4 Prostate cancer and acceleration 57

2.6 Discussion 58

3 How many mutations are in a cancer? **61**

3.1 Original Multistage Model 62

3.2 Logistic clonal expansion 65

3.3 Evidence from cancer genome projects 71

3.4 Rate-Limiting Events 71
3.4.1 When is a mutation rate-limiting? 73
3.5 Discussion 76

4 Comparative studies of risk in inherited and sporadic tumours 78
4.1 Estimating the rate of *APC* mutation 79
4.1.1 Previous estimates of somatic gene mutation rates 79
4.1.2 A comparative method for estimating the *APC* mutation rate . . 83
4.1.3 Results 89
4.1.4 Discussion 90
4.2 HNPCC and sporadic colorectal cancer 95
4.2.1 HNPCC 95
4.2.2 Defining HNPCC 96
4.2.3 Calculating age-specific risk of colorectal cancer in HNPCC patients 97
4.2.4 Heterogeneity and acceleration matching 103
4.2.5 Heterogeneity and acceleration matching in the case of HNPCC 107
4.2.6 Theoretical ΔLLA patterns 108
4.2.7 Sporadic MSI+ colorectal cancer 116
4.3 Discussion 120

5 Population variance in cancer liability 123
5.1 Introduction 123
5.2 Genetic liability to cancer 124
5.2.1 Sibling risk owing to a rare dominant single gene syndrome . . 124
5.2.2 Offspring risk under multiplicative polygenic susceptibility . . . 127
5.3 Environmental liability to cancer 131
5.4 Temporal environmental variance in cancer liability 132
5.5 Estimating total population variance 139
5.6 Using incidence of second primary cancers to estimate liability variance 142
5.6.1 Incidence of second primary colon cancers 146
5.6.2 Incidence of second primary breast cancers 149
5.7 Discussion 150

6 Discussion **153**

6.1 Directions for further work 160

Bibliography **162**

Appendix A **180**

Appendix B **182**

Glossary of terms and abbreviations **184**

List of Figures

1.1 Hanahan and Weinburg identified 6 common hallmarks of cancer cells (redrawn from [HW00]). 30

1.2 Schematic picture of selected signalling pathways regulating cell fate and proliferation (redrawn from [HW00]). These pathways may contribute to a malignant phenotype when hyperactive (purple pathway) or retarded (blue pathways). 30

2.1 Coverage of the cancer registries by region (per cent of total population). The map includes all registries that were members of the International Association of Cancer Registries in 2006. Recreated from Parkin [Par06]. 34

2.2 Incidence measured in annual primary cases per 100,000 population with 95% CIs (left) and log-log plot of the same (right). For many types of cancer the incidence rate seems to follow a power law, increasing in proportion to (age)^n where n depends on the particular cancer being considered. We say, in these cases that the incidence is 'log-log linear' because it appears as a straight line on double logarithmic axes. Leukaemias and sarcomas additionally show small peaks in early childhood and adolescence respectively. These peaks could reflect periods of intense proliferation among the cancer target cells. Gradients were calculated using a least squares method. All data taken from Cancer Research UK. CancerStats - http://info.cancerresearchuk.org/cancerstats - year of diagnosis 2003 (accessed Sept 20, 2007). 35

2.3 Schematic picture of a single stem cell lineage in Armitage and Doll's multistage model. The lineage mutates between states x_i at rate μ per annum. The waiting time between each mutation is exponentially distributed. State x_0 represents a healthy stem cell lineage and state x_n is the malignant state. 36

2.4 (a) Fit of Armitage and Doll's original multistage model to colon cancer incidence with 95% CIs shown for observed data. $100,000 \times h(t)$ is plotted alongside incidence rates recorded among Finnish females between 1959 and 1961 as published in Doll [DPW66]. A maximum likelihood method was used, with likelihood function constructed according to Luebeck and Moolgavkar [LM02], to optimize the hazard function given in equation (2.5). (b) The same plot on log-log axes. . . . 39

2.5 (a) In the original multistage theory a healthy cell lineage becomes transformed through multiple hereditary cellular changes / (epi) mutations. Each follows sequentially from the previous at rates μ_1 to μ_6. Although the mutations are assumed to happen in a defined order only the final step produces a phenotypic effect. (b) In the Two Stage Clonal expansion model (TCSE) an initial mutation (μ_1) causes a clone to grow. Any among the cells in the clone can then become malignant through a second mutation (μ_2). Although most, if not all cancers, contain more than two mutations, TSCE was designed to reflect the major rate-limiting steps identified in chemical tumourigenesis:- initiation (first mutation), promotion (clonal growth) and progression (final mutation) 40

2.6 The two stage clonal expansion model. $S(t)$ is the number of susceptible stem cells at age t. Although this can be modelled by a stochastic process, often $S(t)$ is taken to be a deterministic function of time ($S(t) \equiv$ 'a constant' for example). The initiated 'I' cells have acquired the first heritable change. They divide symmetrically at rate α per annum and die at rate β per annum. One further change , at rate μ_2 per cell per annum is required to make an intermediate cell malignant. . . . 43

2.7 Breast cancer incidence is piecewise log-log linear. The first piece has a steeper gradient and gives way to the second phase in the late 40s around the time of the menopause. Data taken from the SEER database - http://.seer.cancer.gov - year of diagnosis 1993-1997 (accessed Sept 20 2007) 54

2.8 All fits performed using a maximum likelihood method. 95% CIs shown for observed data. Different breast stem cell growth patterns in TSCE model of breast cancer. (Left column) Model based on assumption that size of stem cell pool remains constant. (Middle column) Model based on assumption that size of stem cell pool changes to account for development of breast during puberty. (Right column) Size of stem cell pool changes to account for both development of breast during puberty and involution of breast in old age. Observed breast data taken from Moolgavkar et al.[MDS80]. 55

2.9 The incidence of lung cancer in the British Doctors smoking cohort [DPBS04] was modelled by two separate groups, both using TSCE with smoking-dose-sensitive parameters but assuming different models of dose response. The dose responses implied by both studies are shown. μ_s andμ_n represent the initiation rate in smokers and non-smokers respectively. d_s, d_n and v_s, v_n are the rates of promotion and progression in smokers and non-smokers. The left column shows the ratio $\frac{\mu_s}{\mu_n}$ at different smoking doses, implied by Hazelton and co-workers[HCM05] (black line) model of dose response. The ratio derived by Schollnberger and colleagues [SMB$^+$06] (red line) is also shown. The centre and right columns show the ratios for promotion and progression. 58

2.10 (Left) Age-specific acceleration is the slope of the log-log age-incidence plot. Prostate cancer acceleration is very high during the forties and decreases rapidly thereafter. The other common carcinomas hover around an acceleration of five or six. In general, clonal growth and sequences of multiple mutations create acceleration. Depletion in target cell numbers or reductions in clonal growth rates can reduce acceleration. Lung incidence data taken from the CPS2 cohort [SMB$^+$06], all other incidence data taken from the SEER database - http://seer.cancer.gov - year of diagnosis1993-1997 (accessed Sept 20 2007). Acceleration was calculated by interpolating the incidence data, and then taking the log-log slope of the interpolating function. (Right) Changes in the prostate cancer incidence curve with calendar time in Connecticut. Trend lines are five-year moving averages. Data from SEER nine registry, Nov 2007 submission (accessed June 5 2008). . . . 59

3.1 Probability (posterior density) of different numbers of mutations implied by observed bowel cancer incidence data. Armitage and Doll's original model strongly implies six mutations. By contrast, a model in which stem cell lineages undergo a slow clonal expansion after receiving a certain number of hits (see text), suggests only three or four mutations. In both fits, mutation rates were constrained to fall in the range $10^{-8} - 10^{-2}$ per cell per year and the total number of mutations must be eight or less. The initial number of healthy target cells was set at 10^8. We used uniform priors for the mutation rates and the mutation numbers. 64

3.2 In the logistic clonal expansion model, a cell lineage starts to divide symmetrically on receiving n_{int} initial hits. The resulting clone grows logistically. Any cell in the clone can become malignant by receiving n_f further hits. 65

3.3 The deterministic growth profile of clones in the logistic clonal expansion model. 70

3.4 A multistage model to test the feasibility of determining 'fast' cellular events from incidence data. A cell lineage becomes malignant following n_μ slow steps (at rate μ), followed by n_ν fast steps at rate ν. When the ν is large enough, the quality of fit becomes insensitive to n_ν. 72

3.5 Relative quality of fit achieved by adding in one to five extra stages to an optimized multistage model of bowel cancer. The optimized model assumes 10^8 stem cells at risk and that cancer occurs after six mutations each occurring with probability 7×10^{-4} per cell per year. Quality of fit is measured as a posterior density on the number of extra stages. A uniform prior on 0 - 5 extra stages was assumed. Fast steps, expected to occur in less than six months (i.e. with a mutation rate, ν, larger than 2 per cell per year) have a small effect on incidence and so the quality of fit does not decline substantially when these are added. Bowel cancer data taken from [DPW66] (year of diagnosis 1960-1962). 74

4.1 In Luria-Delbruck fluctuation analysis (left), the number of mutant colonies arising in plated clones can be used to estimate in vitro mutation rates. By analogy, in vivo, the number of tumours arising in individuals can be used to estimate mutation rates. 80

4.2 The pathways of FAP and sporadic bowel cancer are separated by only a single, truncating *APC* hit. In the case of FAP the first *APC* hit already exists in the germline. In a sporadic patient, a given cell lineage takes T_ν years to acquire a truncating APC mutation in one allele and then a further $T_{\text{FAP}}^{\text{lin}}$ years to become malignant. A FAP lineage only takes $T_{\text{FAP}}^{\text{lin}}$ years to become malignant. 82

4.3 The epithelial sheet is replenished, initially, by a fixed number of independent stem cell lineages. These are shown in grey and their differentiating progeny are shown in white. Black cells represent stem cell descendants that have at least one mutant APC allele. The assumptions of the model do not preclude expansion through symmetric division, in the black cell lineages 84

4.4 Posterior distribution of the *APC* mutation rate, ν, measured in mutations per allele per year, calculated using a single parameter likelihood function. Sporadic data colon cancer incidence taken from Doll [DPW66], cases diagnosed between 1960 and 1962. 90

4.5 $P[T_{\text{FAP}} \leq t]$, the cumulative risk of FAP at age t, was calculated separately for males (left column) and females (right column) by interpolating observed FAP data. $P[T_{\text{spor}} \leq t]$, the cumulative risk of sporadic bowel cancer at age t, was constructed from $P[T_{\text{FAP}} \leq t]$, according to assumptions about the relationship between FAP and sporadic bowel cancer described in the text. Using the optimum ν, provides an adequate approximation to observed sporadic data. 91

4.6 Posterior distribution of the *APC* mutation rate ν, measured in mutations per allele per year, calculated using a three parameter likelihood function. Data on sporadic colon cancer incidence taken from Doll [DPW66], cases diagnosed between 1960 and 1962. 92

4.7 $P[T_{\text{FAP}} \leq t]$, the cumulative risk of FAP at age t, is represented by a smooth function, parametrized by μ and n. $P[T_{\text{spor}} \leq t]$, the cumulative risk of sporadic bowel cancer at age t, was again constructed separately for males and females from $P[T_{\text{FAP}} \leq t]$, as described in the text. Using the optimum parameter vector $(\hat{\mu}, \hat{n}, \hat{\nu})$, which maximizes the likelihood function, a good fit to the sporadic data can be made. . . . 93

4.8 Penetrance estimates for bowel cancer in HNPCC have been lowered in light of concerns over ascertainment bias. The original studies, for example Aarnio et al. [AMA$^+$95], that used family history to identify HNPCC kindreds, were enriched for multiple case families and so are thought to have overestimated risk. More recently, Bayesian statistical methods have been used by Quehenberger et al. [QVvH05] to correct for ascertainment bias by conditioning on sample phenotype. This has resulted in a markedly reduced penetrance estimate (figure redrawn from Aarnio et al. and Quehenberger et al. [AMA$^+$95, QVvH05]. . . . 97

4.9 Penetrance functions estimated in three studies that have attempted to mitigate, as far as possible, ascertainment bias. An unexpected but common feature, is the shallow gradient after age 50. 99

4.10 Incidence of CRC in MMR mutation carriers derived from the penetrance functions given in figure 4.9. Declining incidence, after age 50, could be indicative of a combination of factors. For example a severe decline in susceptible target cells, heterogeneity in liability, age-related cell behaviour and study design issues, e.g. ascertainment bias. 100

4.11 Penetrance / incidence estimates from log-log linear simulated patient data, using the inference methods of Jenkins et al. and Dunlop et al. Three different functions were used to simulate the patient data with cumulative risks to age 70 of 0.73, 0.54 and 0.24. In each case the number of patients in the simulated samples were matched to the real sample sizes used by Jenkins et al. and Dunlop et al. 102

4.12 Dots - penetrance of breast cancer in *BRCA1* and *BRCA2* female mutation carriers as calculated by Struewing et al. [SHW$^+$97]. Line - fit of Struewing et al.s estimates. The incidence shown was calculated from the logistic function (equation (4.9)). 103

4.13 Breast cancer rates for females who carry a mutation in BRCA1 or BRCA2, shown as solid lines, versus those females who do not have a mutation shown as dashed lines. The circles in (a) and (c) mark the estimated fraction of females in each class that have not yet developed tumors, taken from figure 1B of Struewing et al. [SHW$^+$97]. In (b) and (d), the observed fraction tumorless, S_{obs}, is transformed to the 'real' fraction tumorless, S_r, via $S_r = \frac{max - (1 - S_{obs})}{max}$, where *max* is the fraction of carriers who have fully elevated risk. Panels (a) and (b) used the smooth.spline function of the R computing language (R Development Core Team 2004) to fit a smooth curve to the logarithms of the observed points, with smoothing parameter set to 0.5; (c) and (d) force a stiffer, less curved fit with a smoothing parameter of 0.6. The second row shows incidence on a $\log_{10}$ scale, obtained from -dln(S)/dt, where S, is the fraction tumorless in the curves of the top row. The bottom row shows ΔLLA, the difference in the log-log slopes of incidence in the second row of plots (Redrawn from Frank [Fra07]). 105

4.14 Panels (a) to (c): survival rates for male FAP patients who carry a mutation in *APC*. The circles mark the estimated probability of being tumorless at various ages, taken from figure 4.7. Panels (d) through (f) show incidence for carriers and non-carriers (dashed line) on a $\log_{10}$ scale. Non-carrier incidence relates to British males diagnosed in 1961 and is taken from [DPW66]. Panels (g) through (i) show ΔLLA, the difference in the log-log slopes of incidence in the second row of plots. . 106

4.15 Panels (a) to (c): survival rates for male colorectal cancer patients who carry a mutation in *MLH1* or *MSH2*. The circles mark the estimated probability of being tumorless at various ages, taken from table 4 of Quehenberger et al. [QVvH05]. Panels (d) through (f) show incidence for carriers and non-carriers (dashed line) on a $\log_{10}$ scale. Non-carrier incidence is taken from table 3 of Quehenberger et al. Panels (g) through (i) show ΔLLA, the difference in the log-log slopes of incidence in the second row of plots. 107

4.16 ΔLLA arising from a germline mutation which abrogates one of n stages in progression under the Armitage and Doll hazard (equation (2.5)) with number of cell lineages $N = 10^8$, $n = 6, 10, 14$ and μ chosen in each case so that the penetrance at age 80 in the healthy genotype (blue line) is equal to 5%. While the effect on penetrance of the mutant gene (red line) is diminished as the rate of transition between stages increases (with increasing n), ΔLLA remains roughly equal to one. 109

4.17 ΔLLA arising from a germline mutation which increases the rates of transitions under the Armitage and Doll hazard (equation (2.5)) with $N = 10^8$, $n = 6, 10, 14$ and μ chosen in each case so that the penetrance at age 80 in the healthy genotype (blue line) is equal to 5%. All of the age axes are logarithmic to base 10. Incidence also is plotted on a $\log_{10}$ scale. The increased transition rate, ν, even at only 4 times the original transition rate, causes a strong increase in incidence but only a small deceleration in mutation carriers (red line). Incidence of CRC observed in HNPCC is never higher than 10^{-1} 110

4.18 ΔLLA arising from a germline mutation which increases the rates of q out of n transitions by a factor of 4 under the Armitage and Doll hazard (equation (2.5)) with $N = 10^8$, $n = 6$, $q = 1, 3, 5, 6$ and μ chosen so that the penetrance at age 80 in the healthy genotype (blue line) is equal to 5%. 111

4.19 ΔLLA arising from a germline mutation which slows the rates of 3 out of n transitions while quickening all other transitions. Incidence in the healthy genotype (blue line) is modelled by equation (2.5) with $N = 10^8$, $n = 10$ and μ chosen so that the penetrance at age 80 is equal to 5%. The heterogeneity in the syndrome associated transition rates produces a plateauing incidence and rising ΔLLA (red line). . . . 111

4.20 Penetrance, log-incidence and ΔLLA assuming a six-step pathway with one clonal expansion in healthy patients (blue line) and the same pathway with one step deleted in mutation carrying patients (red line). The top four panels are calculated assuming a clonal expansion in the final stage with capacity of $K_5 = 10^6$ cell lineages and initial growth rate of 0.4. The bottom four panels assume a faster growing and larger clone. In either case the mutation rate per lineage is chosen so that cumulative risk to age 80 is 5% in healthy patients. The effect of the more aggressive clone is only to shift the kink in ΔLLA to earlier ages. The departure from a constant ΔLLA of one remains small in either case. . . 115

4.21 Penetrance, log-incidence and ΔLLA assuming a six step pathway with one clonal expansion in healthy patients (blue line) and the same pathway but with two clonal expansions in mutation carriers (red line). All clones have capacity $K = 10^6$ and initial growth rate $r = 0.4$. The mutation rate per lineage is chosen so that cumulative risk to age 80 is 5% in healthy patients. The mutation carriers out-accelerate the healthy patients throughout mid-life causing a negative ΔLLA. 116

4.22 Number of cases of MSS and MSI+ CRC occurring at 9 regional hospitals in southeast Finland over a four-year period. 118

4.23 Estimated age-structure of the population served by nine regional hospitals in southeast Finland. (a) Frequency by 5-year age group, averaged over the years 1994-1998, taken from 'Statistics Finland' for 8 geographical regions of Southeast-Finland. (b) Frequency of population by age in 8 healthcare regions, taken from the Finnish cancer registry. (c) Population age-structure of the combined regions in (a) and in (b). . 119

4.24 (Green line): average annual cases of CRC per unit population during 1994-1998, estimated from nine hospitals in southeast Finland, assuming a 300,000 catchment population. (Orange line): nationwide cancer incidence over the period 1993-1997, taken from the Finnish Cancer Registry [PWF$^+$97]. (Blue lines): MSI+ incidence estimated from the nine hospitals data. 120

4.25 (Left): The red circles are log of incidence of MSI+ CRC, estimated from the data of Salovarra et al. [SLK+00] and Aaltonen et al. [ASK+98]. The circles are fit with the smooth.spline function of the R computing language, with smoothing parameter set to 0.5. (Right): LLA calculated from the smoothing spline opposite. 121

4.26 (Top row): the red circles show the fraction of MMR mutation carriers who are tumourless on a log scale, estimated in three studies of HNPCC penetrance. The data are fit with smoothing splines as in figure 4.13. (Middle row): the solid lines show incidence derived from the smoothing splines above. The dashed lines show incidence of MSI+ sporadic CRC as estimated above (see figure 4.24). (Bottom row): ΔLLA calculated as the difference in gradient between the solid and dashed lines from the middle row. 121

5.1 Simplistic liability distribution for CRC. Population frequency of FAP taken as 1:10000 [BFB+94] and population frequency of germline MMR mutation taken as 1:3000 [DFN+00]. For a review of further hereditary CRC syndromes that are rarer still see Lynch and de la Chapelle [LdlC03] 124

5.2 Given that one offspring is a confirmed case of cancer (sibling (a) shown in black), the probability that the other will develop cancer within a lifetime depends on, among other things, the existence of predisposing allelic variants in the population. 125

5.3 Relative sibling risk as a function of deleterious allele frequency for various genotype relative risk values. Calculated from equation (5.4) which assumes random mating. The biphasic nature of this graph can be explained as follows: for very low allele frequencies an affected sibling is only in rare cases likely to carry the deleterious allele, and hence the sibling risk approaches the population risk as the allele frequency tends to zero. For very high allele frequencies, an affected sibling will likely carry the allele but then so will most of the population so the sibling risk approaches the population risk also as the allele frequency tends to one. The data point highlighted in red in table 5.1 is shown. . . 127

5.4 A nuclear family with undetermined number of offspring. Given that one parent is a confirmed case of cancer, the probability that a given offspring will develop cancer within a lifetime again depends on the existence of predisposing allelic variants in the population. 128

5.5 Relative offspring risk based on 1,2, ... ,8 susceptibility loci plotted against allele frequency. 130

5.6 Notional liability distribution for "sporadic" CRC based on a multiplicative polygenic model with 8 risk loci, p=0.1 and R=2. The baseline risk, s, is fixed so that mean lifetime risk is 0.05. The distribution arising from the product of many independent positive valued random variables tends to be lognormal. 130

5.7 Under MFT, the heritability h (equation (5.8)) implied by a fixed ratio in twin relative risk, between monozygotic or dizygotic twins, increases with increasing population prevalence, K. $\sigma_g^2+\sigma_c^2+\sigma_e^2$ was normalized to one in the calculations of twin relative risk and σ_c^2 was fixed at 5%. . 133

5.8 Top: age-specific female breast cancer incidence in Connecticut. Plotted for the years 1973 - 2003, giving a rough picture of dispersion in the disease counts at each age. Bottom: a rising temporal trend for 40, 60 and 80 year olds can be seen following the initiation of mammography screening in the early 1980s [AJD06]. This trend contributes to the count dispersion. 134

5.9 Top: colorectal cancer incidence in Connecticut for the years 1973 - 2003. Bottom: downwards temporal trends in incidence for 40, 60 and 80 year olds are significant from the 80s onwards. Increased use of sigmoidoscopy and fecal occult blood tests (triggering colonoscopy) beginning in the 70s seems to have precipitated the early detection and removal of precancerous legions (e.g. adenomas) eventually impacting on incidence in the following decade [CTC+94]. Lifestyle changes may also play a role in the continuing steady reduction in colorectal cancer incidence. 135

5.10 Top: age-specific prostate cancer incidence in Connecticut (1973 - 2003). Bottom: strong temporal incidence trends following the introduction of prostate specific antigen (PSA) screening in the late 1980s [KFFM00]. PSA testing is highly sensitive. Its use has meant cancers are registered at earlier stage / age and has also led to the detection of some cases that would never have become clinically apparent over the lifetime of the patient in the absence of PSA testing. 136

5.11 Posterior densities for Σ_E^2 in the case of prostate, breast and colorectal cancer. 138

5.12 Lognormal distribution with mean set to one, and variance given by $\frac{2^n}{100}$ for $n = 0, 1, \ldots, 6$. 139

5.13 (a) Distribution of lifetime risk implied by the study of Locatelli et al. (b) Increasing liability variance causes a decoupling of expected population incidence and the baseline hazard rate. (c) A larger proportion of cases arise in a smaller minority of the population as the lognormal variance is increased. Under a variance of 45, 80% of cases occur in the 20% of the population at highest risk. 143

5.14 Locatelli et al. [LRLY07] used a Gompertz hazard to model age of breast cancer onset in Swedish twins born between 1886 and 1967. A Gompertz hazard is a questionable model for Breast cancer incidence. . 144

5.15 Relative hazard for second primary cancer given one primary diagnosis accumulated by age 40. 146

5.16 Theoretical relationship between relative risk and liability variance based on the studies of Hoar et al. and Harvey and Brinton. 148

5.17 Relationship between fraction of the population at highest risk and the fraction of cases occurring in that high risk subset for breast and colon cancer - estimated from the data of Harvey and Brinton and Hoar et al. respectively. 151

5.18 Colon cancer shows a rising risk of second primary with time since initial diagnosis, as predicted by the lognormal relative hazards model. Breast cancer, by contrast, shows a stable incidence with time since initial diagnosis. 152

6.1 (a) Red line: background population penetrance of 5% ($c = 10^{-12}$, $d = 5$). Purple line: 30% of the population have complete penetrance by age 80 ($a = 10^{-8}$, $b = 4$. Green line: observed penetrance in the mixed population. (b) Incidence corresponding to the population penetrance. 157

6.2 (a) Green line: higher penetrance in some lineages disposed to a pathway with greater acceleration of 7. Red line: lower penetrance in cells disposed to a pathway with a lower acceleration of 5. Blue line: the composite penetrance has a modulating acceleration which starts at 5 and rises to 7. 158

6.3 (b) Penetrance arising from two different distributions of mutation rate in target lineages within a tissue. The dashed red line is calculated assuming every lineage has a mutation rate of 0.001 per annum. The blue line is calculated assuming a lognormal distribution of mutation rate as shown in (a). 159

List of Tables

2.1 Example incidence data in the format produced by PBCRs 38

3.1 Expected time lapse in years before one, two or three specific mutations occur in any of a clone of target cells. Clone size is measured in cells. Hits refers to the number of specific gene mutations that are to occur in any one cell of the clone. The mutation rate is quoted per cell per cell generation, assuming 100 generations per year. Since a continuous model of mutation is assumed, mutations can occur at any time and are not limited to fixed points in the cell cycle. This explains why the expected time lapse is less than one cell generation time in some cases . 75

4.1 FAP data (males only) from [Ash69], the patients treated during each age interval are removed from the study, and no longer form part of the analysis. This is reflected in the 'Patients for Analysis' column, which contains the total number of patients who began the study and have not yet received treatment. 85

5.1 Sibling relative risk, λ_s, as a function of disease allele frequency, p, and genotype relative risk, R. 126

5.2 Expected verses observed incidence of second primary colon cancer in Connecticut. Data from Hoar et al. Relative risk is calculated as the ratio of observed to expected cases. Incidence is the ratio of observed cases to person years in interval. 147

5.3 Expected verses observed incidence of second primary breast cancer in Connecticut. Data from Harvey and Brinton. Relative risk is calculated as the ratio of observed to expected cases. Incidence is the ratio of observed cases to person years in interval. 149

Chapter 1

Introduction

Most cancers occur with the same characteristic pattern of incidence [PWFS05]. The simplicity of this pattern is in contrast to the perceived complexity of carcinogenesis. Age-onset statistics therefore represent a seductive set of data and have provoked many bold but often misguided conclusions concerning the physiopathological mechanisms of cancer. Half a century has passed since the original 'multistage theory' of Armitage and Doll [AD54]. Although their basic idea of a healthy cell becoming malignant in several 'rate-limiting' steps is still accepted, prevailing wisdom about the nature and number of these steps has never settled into a consensus [Arm85]. Meanwhile, many quantitative attempts to learn about cancer aetiology from incidence statistics have floundered in the face of too many unknowns and too few data [HPT07]. Indeed, a lack of specificity in recorded incidence rates continues to pose a problem. Take bowel cancer as an example, it is easy to come by statistics on the age-distribution of bowel cancer. High quality bowel cancer incidence data are available for many different populations and at many points in calendar time over the last 50 years [PWFS05]. Turn your interest to bowel cancer with micro-satellite instability, however, and the data pool shrinks spectacularly. Go a step further to cancer of the bowel, with micro-satellite instability and arising in the context of a rare hereditary syndrome and the 'pool' may dry up altogether. Gathering such bespoke statistics is not usually something that a single person or research group can feasibly undertake or commission. Population based studies require a well orchestrated collaboration between physicians, hospitals and diagnostic departments [Par06]. So, in the main, good must be made of the data collected by established registries. The ideal registry would subject every cancer within its catchment area to genotyping and extensive laboratory analysis, and curate an exten-

sive database with genomic information for each patient's neoplastic and normal tissue, in addition to the usual items; gender, date at onset and physical location of the tumour etc.. Enriching as this would be for the study of incidence patterns, and their underlying aetiologies, such a situation is clearly prohibitively expensive, ethically untenable and unlikely to materialize in the near future. Mathematicians studying cancer incidence must work with what is available. Should their research lead to a theory of great interest but which requires validation through further observation, usually they will be powerless to collect the necessary data. This is by no means a unique position, there are many branches of science in which the process of observation is not carried out directly at the behest of the relevant theoreticians. Nevertheless, it is mentioned here because, lacking a quick, clean cycle of observation, analysis and hypothesis, is an issue particularly relevant to epidemiology. It informs the complexity of mathematical incidence model one can reasonably expect to validate [Fra05], while placing a heavy premium on extant data and the creative use thereof. In the context of these considerations, which really amount to nothing more than a complaint about scarcity of data, the task of modelling incidence statistics can be usefully contrasted with other quantitatively focused branches of cancer research that concern themselves with different aspects of the disease and are motivated by different objectives. One example of interest, quite distinct from the work contained in this thesis, is the mathematical description of tumour spheroid growth [AM04]. An in vitro tumour can be cultivated by planting cancer cells in a culture medium. Quantitative models of the resulting spheroid typically centre around the responses of its cells to oxygen and nutrient diffusing into the tumour from the medium. The cells may consume nutrient, proliferate, move around, enter growth arrest or commit suicide depending on their exposure to nutrient and oxygen. Many of the parameters relevant to the dynamics of an in vitro tumour spheroid are directly measurable [KSKK98], for instance, proliferation rates of cultured cells in different chemical environments or the viscosity of the growth medium. Furthermore, it is possible to generate a large number of falsifiable hypotheses such as "we predict that the concentration of glucose and oxygen will fall in the middle of the tumour", or, "we predict that if the concentration of glucose in the growth medium is increased by a factor of X, then the limiting size of the avascular tumour will increase by a factor of Y." Experimental validation of these hypotheses is readily at hand. With control of the

object under scrutiny comes the power to falsify. In such a case as this, setting out to build a fairly detailed model, with many parameters relating to measurable quantities seems justified. Even still, you must know when to stop. In biological problems, the potential for adding detail and complexity is almost unlimited. Sensible decisions must be made; should the cell be treated as a black box? Or should explicit details of the cell's metabolism be included? If a model acquires too many unmeasurable parameters, then it can become nothing more than a data fitting machine with little predictive power [BA98]. Quantitative work on in vitro tumour spheroids, among other things, is motivated by a desire to advance the art of mathematical modelling, to demonstrate its power or discover its limitations. In addition, it can provide a valuable framework for analysing and reasoning about an interesting experimental system. There is always the hope that a very successful model will emerge than can be used to guide or optimize an important enterprise like drug design or testing [BBK04] in much the same way as computational fluid mechanics has guided the evolution of aeroplane wing shapes for example. Another interesting but contrasting application of mathematics in cancer research, is related to genetic counselling. Specifically, the problem of calculating a person's risk of carrying a susceptibility allele, given their family history of cancer and perhaps information relating to other known risk factors [Fou08]. This problem is attached to a concrete short term payoff. Namely, the ability to save money and reduce anxiety by targeting genetic testing to those patients who will most likely prove positive. The mathematical model used to solve this problem need only be as detailed or sophisticated as is required to increase its predictive accuracy. Given a set of inputs (the details of a patients family history), it must produce the correct output (the probability that the patient is a carrier of the gene of interest). An involved representation of the chain of causality between inputs and outputs is not necessarily required. This is because the model will always be used to answer the same question. It need not generalize in the same way that an in-silico model of a tumour would have to, were it ever to be of any use to the pharmaceutical industry. For example, take the successful BOADICEA model of genetic susceptibility to breast and ovarian cancer [ACP$^+$08]. BOADICEA incorporates hypothetical, undiscovered susceptibility alleles because doing so has a positive effect on its performance. However, the unknown alleles are assumed simply to cause a log normal distribution of liability in the population. There is no requirement

for explicit terms representing each of the alleles, their population frequencies and the risks they produce when occurring together in various combinations in the same individual. Mathematical models of human cancer incidence (the subject of this thesis), fall somewhere between the two examples given above. In terms of manipulability of the object under study and the power to falsify theoretical model predictions, they are certainly at a disadvantage compared with in vitro tumour models, or any model of a malleable experimental system. Hence, complexity ought to be scaled down accordingly. However, since the objective is to build a somewhat generalizable theory of incidence suitable for more than just regression, we can expect a greater level of complexity and incorporation of more biological detail than found in a susceptibility predictor. Striking the optimum balance is difficult, but an important determiner of success.

1.1 The mechanistic basis of cancer

To clarify the above it should be helpful to sketch an outline of the cancer disease process and to highlight its main points of contact with the models developed throughout this thesis. The regenerative tissues of the body can be viewed as tightly regulated multicellular communities. A healthy compartment of proliferative cells will maintain its architecture through a controlled balance of cell birth, differentiation and cell death [PA07]. Strict rules, ratified into the circuitry of each participating cell, dictate social cell-cell and cell-extracellular matrix interactions [and08]. Proliferation is prohibited except in appropriate circumstances and enforced senescence or suicide follows aberrant behaviour at the intercellular or intracellular level. A hard limit to the maximum number of divisions each cell lineage can undertake provides further protection from the unrestrained proliferation which characterises cancer [Hay65]. However, as a cell lineage ages and accumulates various kinds of genetic and epigenetic damage, the genomic information encoding social responses to internal and external stimuli gradually looses fidelity. As a result a cell may begin to behave in an aberrant fashion, dividing constituently in defiance of the laws of tissue homeostasis. If a clone of such rebellious cells is established it may overturn the healthy functioning of the organ in which it arises or in an organ it has spread to. So it can be seen that cancer is closely linked to the aging process: a process of accumulating cellular damage leading to macroscopic

loss of integrity and consequent system failure. Cancer mortality is distinguished from intrinsic mortality or 'death from old age' however, by the characteristics and details of the system failure itself. To make a popular analogy with the automobile, death from old age would correspond to an old rusted car which slowly grinds to a halt, or simply fails to start one morning, the causative damage being widespread and hence the final trigger for system failure difficult to determine. Cancer, by contrast, better corresponds to an unfortunate but more specific combination of subcomponent failures leading to a dramatic catastrophe such as an explosion whose cause may more readily be audited. Cancer is hence seldom a disease of the extreme elderly as it must harness some of the natural vitality of the body to propagate itself [HPLW08].

Much of the study of the molecular basis of cancer has focussed on the specific gene alterations which encourage cancerous cell behaviour. However, when viewed in terms of these mutational signatures cancers appear very heterogeneous with different patients of the same tumour type expressing different mutations. The vast number of alternative mutational combinations that can lead to the same cancerous end-point has been affirmed by the results of recent cancer genome projects [SJW$^+$06, GSS$^+$07] and led to the suggestion that cancer should perhaps be viewed not just in terms of the affected genes that drive it but the pathways to which these genes belong. The signalling pathway, as a unit of explanation, may more readily elucidate the commonalities between tumours and may also prove a more useful target for drug intervention [Jon08]. A unifying pathway model of cancer was famously articulated by Hanahan and Weinburg before the aforementioned cancer genome projects came to fruition [HW00]. Hanahan and Weinburg identified 6 generic acquired properties which define cancerous cell behaviour and hypothesised that a small number of common core signalling pathways must be altered (albeit it through various potential gene targets) to achieve these (figure 1.1).

It is common practice in multistage modeling for the genetic alterations that disrupt tissue homeostasis to be described as stochastic events, obeying point processes in time.

To illustrate this with a specific example (see figure 1.2), biallelic mutation of *SMAD4* may result in its failure to transduce growth repressive TGFβ signalling, thus conferring insensitivity to anti-growth signals. These biallelic mutations would be

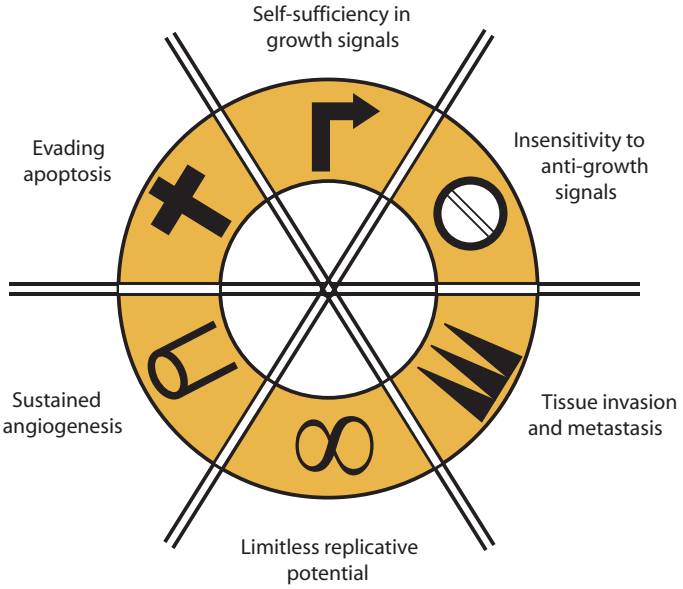

Figure 1.1: Hanahan and Weinburg identified 6 common hallmarks of cancer cells (redrawn from [HW00]).

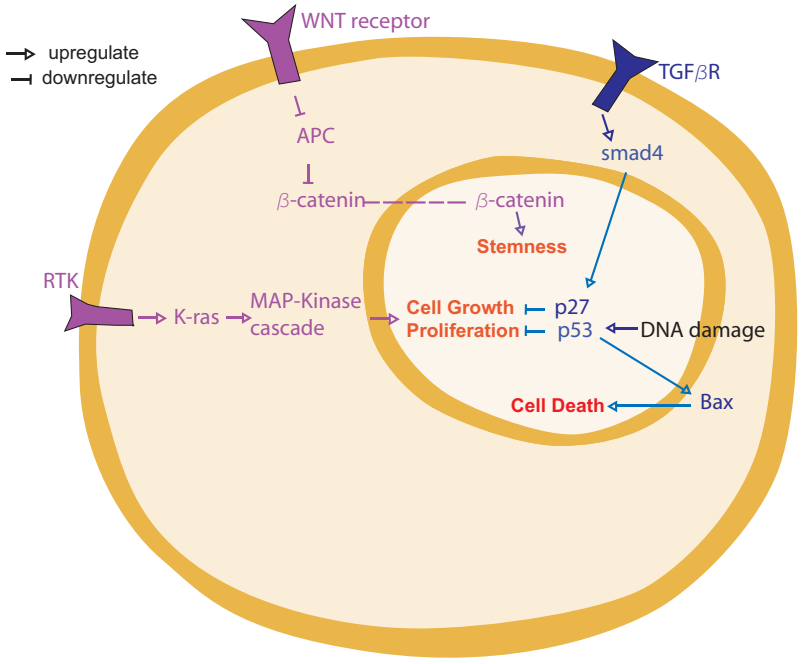

Figure 1.2: Schematic picture of selected signalling pathways regulating cell fate and proliferation (redrawn from [HW00]). These pathways may contribute to a malignant phenotype when hyperactive (purple pathway) or retarded (blue pathways).

modelled as two separate random events. Alternatively, mutation of the *K-RAS* oncogene may constituently activate mitogenic signal transduction pathways creating self-sufficiency in growth signalling. This would be modelled as a single stochastic event. Finally, biallelic *p53* mutation could inactivate the cells apoptotic response to DNA damage, thus allowing the cell to evade apoptosis and continue proliferating in an aneuploid state.

1.1.1 Genetic instability

Aneuploidy is often observed in tumour cells. Other types of widespread genetic damage are also typical [LKV98]. This has led to the suggestion that a key property of cancer cells may be an acquired genetic instability. Cells with a retarded capacity to maintain genomic integrity ought to age more quickly and more readily acquire the various mutations required for malignancy. The existence of hereditary cancer syndromes caused by germline mutations that target genes involved in DNA maintenance certainly suggests that an elevated rate of DNA mutation is carcinogenic in some circumstances [dlC04]. However, the wider perspective on genetic instability and the extent to which it precipitates or is rather a consequence of cancer in general is uncertain [SHT03, RNVL03]. The onset of genetic instability can be represented cleanly in multistage models as a simple increase to a mutation rate parameter. For example, in chapter four, a model with a variable mutation rate is used to simulate the accumulation of mutations in miss-match repair deficient tumors.

1.2 Structure of this thesis

Chapter two is a critical review of multistage modelling. Much of the mathematical machinery required for the remainder of the thesis is also developed here. Chapter three is a quantitative discussion of inherent difficulties in analysing age-distributions. Specifically, the problem of distinguishing which kinds of aetiological events can be investigated with incidence statistics. Armed with the lessons of chapters two and three, chapter four presents two attempts to learn cancer aetiology from age-onset patterns. Both concern hereditary bowel cancer syndromes, and the disease-related activity of the genes which underlie these syndromes. Chapter five is a discussion of liability to cancer and an attempt to quantify its variance in human populations. Chapter six gives

a summary of the central results and conclusions of the thesis and examines directions for further work.

Chapter 2

Multistage Theory

2.1 Armitage and Doll

Cancer incidence refers to the rate at which the disease arises. Measured in cases per 100,000 people per year, accurate accounts of incidence have only been possible since the first half of the twentieth century. The advent of the population based cancer registry (PBCR) led to the first reliable statistics on rates of cancer by age at diagnosis and site. The PBCR achieves these data by recording every new case of cancer in a defined population - usually those persons living within a specified geographical area. Beginning in Europe in 1927 and North America in 1940, this has evolved into a global activity. The International Association of Cancer Registries currently has 449 members worldwide covering over 20% of the world's population (figure 2.1).

The rise of population based cancer registration was motivated by a wish to compare prevalence between different places and over time [DPW66]. Such comparisons have uncovered potential carcinogens through the identification of environmental factors that modify cancer risk. Based on the observation that migrants often assume the cancer rates of their new country [Hae61], it was concluded in the early eighties that large disparities in cancer burden between England / USA and other countries were attributable to differences in diet, smoking, reproductive behaviour, sexual behaviour, infection and occupational exposures. The existence and extent of these associations have been confirmed in subsequent epidemiological studies [Col06]. Meanwhile accumulated registry data have been put to use in many aspects of cancer control, from planning to the evaluation of screening and treatment programmes. For a review see Parkin [Par06].

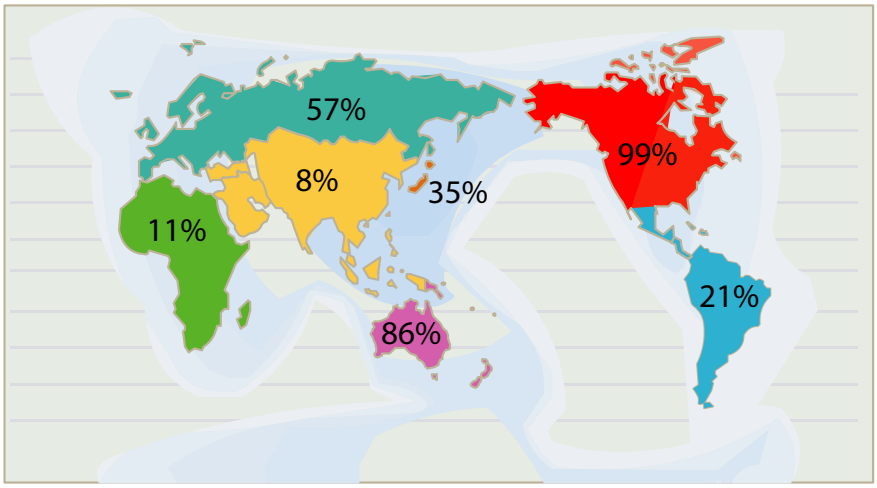

Figure 2.1: Coverage of the cancer registries by region (per cent of total population). The map includes all registries that were members of the International Association of Cancer Registries in 2006. Recreated from Parkin [Par06].

An alternative branch of cancer epidemiology developed in parallel with the aforementioned descriptive studies of incidence. In 1954, Armitage and Doll published a landmark study of the age-distribution of cancer [AD54]. Mortality statistics (taken as a good indicator of incidence) recorded in several developed countries [Nor53] had revealed an intriguing dependence of cancer on age. The number of deaths in a specified age group, observed over one year, was roughly proportional to the nth power of age, with n around five or six for many cancers including the common carcinomas. We now know this to be true of incidence also (figure 2.2). Armitage and Doll proposed a 'multistage theory' to explain this observation. They showed that if six or seven rare cellular changes led to cancer (figure 2.5a), then its age-distribution would have approximately the correct shape (figure 2.4). Their proposed 'cellular changes' can be equated with gene (epi) mutations. The key to Armitage and Doll's formulation was to assume that cancer arises in a susceptible target of asymmetrically dividing cells (which can now be thought of as stem cells). Each such stem cell and its lineal descendants could then be considered as a single entity - a stem cell lineage. Under this simplification, the probability that an organ is afflicted with cancer before a given age has a straight-forward interpretation. It is the probability that at least one of the susceptible stem lineages comprising the organ has acquired the necessary number of mutations by the age given. A crude expression for this probability can be written in

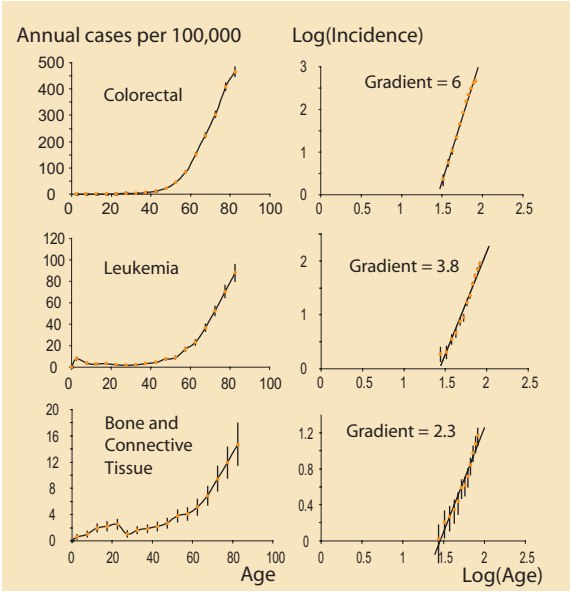

Figure 2.2: Incidence measured in annual primary cases per 100,000 population with 95% CIs (left) and log-log plot of the same (right). For many types of cancer the incidence rate seems to follow a power law, increasing in proportion to (age)^n where n depends on the particular cancer being considered. We say, in these cases that the incidence is 'log-log linear' because it appears as a straight line on double logarithmic axes. Leukaemias and sarcomas additionally show small peaks in early childhood and adolescence respectively. These peaks could reflect periods of intense proliferation among the cancer target cells. Gradients were calculated using a least squares method. All data taken from Cancer Research UK. CancerStats - http://info.cancerresearchuk.org/cancerstats - year of diagnosis 2003 (accessed Sept 20, 2007).

terms of the number of lineages at risk, N, the number of mutations required, n, and also the probability of mutation per year at each locus, μ.

2.1.1 Derivation of Armitage and Doll's formula

First of all, a cell lineage and its random acquisition of mutations are modelled by a continuous time Markov chain. The states of the Markov chain represent the different genotypes created by successive mutations (figure 2.3). The notation x_i is used to

represent the ith state, created by i mutations. x_n is the malignant state, requiring n mutations altogether. The waiting time between each mutation is exponentially distributed with mean $\frac{1}{\mu}$ years. In other words, the rate of mutation is assumed to be μ per annum at each locus. The notation $x(t)$ is used to represent the trajectory of a particular lineage across the chain of states. $x(t)$ takes values $x_1, x_2, \ldots$ The notation $X_i(t)$ is used to represent the probability that the lineage is in the state x_i at age t. This means that $X_i(t) = P[x(t) = x_i]$. Using this notation, the Kolmogorov forward equations for a single cell lineage are:

$$\begin{aligned}
\frac{d}{dt}[X_0(t)] &= -\mu X_0(t) \\
\frac{d}{dt}[X_1(t)] &= \mu(X_0(t) - X_1(t)) \\
&\vdots \\
\frac{d}{dt}[X_{n-1}(t)] &= \mu(X_{n-2}(t) - X_{n-1}(t)) \\
\frac{d}{dt}[X_n(t)] &= \mu X_{n-1}(t).
\end{aligned}$$

Figure 2.3: Schematic picture of a single stem cell lineage in Armitage and Doll's multistage model. The lineage mutates between states x_i at rate μ per annum. The waiting time between each mutation is exponentially distributed. State x_0 represents a healthy stem cell lineage and state x_n is the malignant state.

The initial condition for this system of linear ODEs is:

$$\begin{pmatrix} X_0(0) \\ X_1(0) \\ \vdots \\ X_n(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Solving this system gives

$$\begin{aligned} X_0(t) &= \exp^{-\mu t} \\ X_i(t) &= \frac{\mu^i t^i}{i!} \exp^{-\mu t}, \quad i = 1, \ldots, n-1 \\ X_n(t) &= 1 - \sum_{i=0}^{n-1} \frac{\mu^i t^i}{i!} \exp^{-\mu t}. \end{aligned} \tag{2.1}$$

$X_n(t)$ is the probability that a given stem cell lineage is malignant at age t. It can be used to calculate the probability that one or more of a collection of stem lineages is malignant at age t. If there are a total of N stem cell lineages mutating independently of one another, then the probability that none of them is malignant by age t is $(1 - X_n(t))^N$. The probability that one or more of them is malignant by age t, is one minus the probability that none of them are. So if T represents the age at which the first stem lineage becomes malignant then

$$P[T \leq t] = 1 - (1 - X_n(t))^N. \tag{2.2}$$

Substituting in from equation (2.1) gives

$$P[T \leq t] = 1 - \left(\sum_{i=0}^{n-1} \frac{\mu^i t^i}{i!} \exp^{-\mu t} \right)^N. \tag{2.3}$$

2.2 The Hazard Function

In addition to $P[T \leq t]$, it is also useful to calculate the hazard function, $h(t)$. The hazard function is sometimes referred to as the incidence function. It gives the instantaneous rate of occurrence of cancers in a collection of non-malignant lineages at age t. The formal definition of the hazard function is

$$h(t) = \lim_{\Delta t \to 0} \left(\frac{1}{\Delta t} P[t < T \leq t + \Delta t | T > t] \right).$$

$h(t)$ is related to $P[T \leq t]$ by the formula:

$$P[T \leq t] = 1 - \exp\left[-\int_0^t h(s)\,ds\right],$$

since by definition:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0}\left(\frac{1}{\Delta t} P[t < T \leq t + \Delta t | T > t]\right) \\ &= \frac{1}{1 - P[T \leq t]} \lim_{\Delta t \to 0}\left(\frac{1}{\Delta t} P[T \leq t + \Delta t] - P[T \leq t]\right) \\ &= \frac{\frac{d}{dt}\left(P[T \leq t]\right)}{1 - P[T \leq t]} = -\frac{d}{dt}\ln(1 - P[T \leq t]). \end{aligned} \tag{2.4}$$

The hazard function is useful for data fitting and also as a continuous approximation to observed 'age-specific incidence' . PBCRs provide incidence in annual primary cases per 100,000 by age group. Usually the age groups are 5 years wide, so that the data is presented as:

Age last birthday	Annual primary cases per 100, 000
0 - 4	0
5 - 9	0
10-14	2
15 - 19	5

Table 2.1: Example incidence data in the format produced by PBCRs

This 'age-specific incidence' is well approximated by $100,000 \times h(t)$. For example, the expected 'age-specific incidence' observed in the age interval $[t, t + 5)$ is roughly $100,000 \times h(\hat{t})$ where $t \leq \hat{t} \leq t + 5$. A simple way to visualize the quality of an

incidence model is to plot observed age-specific incidence against the hazard function. For the Armitage Doll model (2.3), the hazard can be expressed in terms of elementary functions using equation (2.4):

$$h(t) = \frac{N\mu^n t^{n-1}}{(n-1)! \sum_{i=0}^{n-1} \frac{(\mu t)^i}{i!}}. \tag{2.5}$$

As an illustrative test of the Armitage and Doll hazard, if it is assumed that there are $N = 10^8$ stem cell lineages in the average colon [PBH03], fitting to colon cancer incidence (figure 2.4) implies, $n = 6$ and $\mu = 8 \cdot 10^{-4}$. That the estimate for μ is high compared with estimates made in human cell cultures [SKT+87], may reflect an absence in the model of mechanisms, such as selection and clonal growth, which can accelerate the multistage process despite low rates of gene mutation.

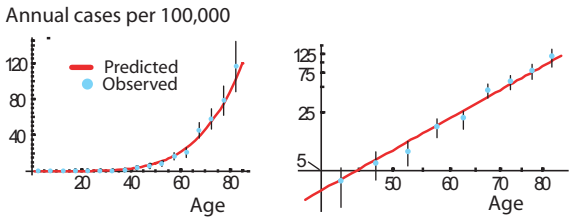

Figure 2.4: (a) Fit of Armitage and Doll's original multistage model to colon cancer incidence with 95% CIs shown for observed data. $100,000 \times h(t)$ is plotted alongside incidence rates recorded among Finnish females between 1959 and 1961 as published in Doll [DPW66]. A maximum likelihood method was used, with likelihood function constructed according to Luebeck and Moolgavkar [LM02], to optimize the hazard function given in equation (2.5). (b) The same plot on log-log axes.

2.3 The two stage clonal expansion model

It had been suggested that cancer might arise through mutation in the hereditary material of a somatic cell since as early as 1930 [MM30]. Despite this, when the multistage theory was first published, ideas about the causes of cancer were still dominated by those of the great 19th century German pathologists. One popular such theory was that cancer arose from embryonic cells that had failed to differentiate and persisted in

adult tissues. Even as late as 1960 there was still significant doubt regarding mutational theories [Bru60]. In this context, the pertinent insight of Armitage and Doll was that steep increases of cancer with age are indicative of random, heritable, multiple and rare causal cellular events. For a more exact understanding of the age dependence of a particular cancer, the simplifying assumptions of their original theory are insufficient.

Every cancer deviates to some extent from log-log linearity. Moreover, a cell centric model, that considers only a simple sequence of mutations without any benign growth before malignancy, does not adequately represent our understanding of cancer as somatic evolution. Revised multistage theories partially address these issues by incorporating clonal expansion and other mechanistic details. Among such models, the most widely adopted are the two stage clonal expansion model (TSCE) [MDV88] due to Moolgavkar and colleagues and its derivatives [LM02, LW03].

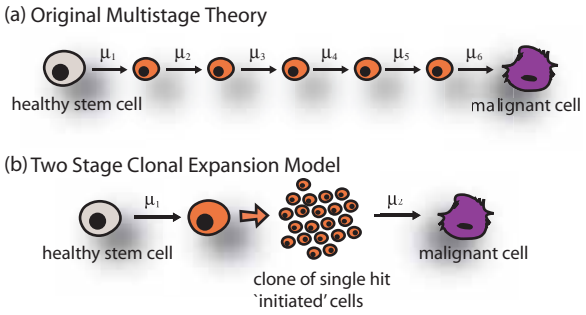

Figure 2.5: (a) In the original multistage theory a healthy cell lineage becomes transformed through multiple hereditary cellular changes / (epi) mutations. Each follows sequentially from the previous at rates μ_1 to μ_6. Although the mutations are assumed to happen in a defined order only the final step produces a phenotypic effect. (b) In the Two Stage Clonal expansion model (TCSE) an initial mutation (μ_1) causes a clone to grow. Any among the cells in the clone can then become malignant through a second mutation (μ_2). Although most, if not all cancers, contain more than two mutations, TSCE was designed to reflect the major rate-limiting steps identified in chemical tumourigenesis:- initiation (first mutation), promotion (clonal growth) and progression (final mutation)

TSCE (figure 2.5b) shares the common basic assumptions of the original multistage theory. A population of cell lineages is at risk for a given cancer. Mutations can afflict any of these lineages with a certain probability per cell generation, and cancer arises when the first of these lineages has acquired enough mutations. TSCE differs in that the first mutation is assumed to cause a benign growth. Specifically, when a target lineage receives its first mutational hit, it divides, giving birth to a clone of identical 'initiated' one-hit lineages. Any member of this clone is then at risk of becoming cancerous through only one further mutation. Steep increases in cancer with age under TSCE are caused by the growing number of one hit lineages populating benign precursor lesions as a patient ages.

TSCE has proved a versatile theory, able to synthesize a variety of incidence patterns, both log-log linear and otherwise. From a technical perspective, its biggest triumph lies in its stochastic representation of clonal growth. Similar but inferior modifications to multistage theory treat the expansion of an initiated lineage as inevitable once the lineage has arrived at a certain genotype. In TSCE, mutant clones may become extinct through random cell death while they are still young. If they survive, their growth profiles are exponential on average but fluctuate randomly about this trend. An obvious limitation of TSCE is that it only allows for two rate-limiting stages. Elegant generalizations of TSCE have removed this restriction and can account for more than one initiating mutation [LM02] as well as multiple and sequential rounds of clonal expansion at different growth rates [Lit96]. An unresolved limitation of TSCE is the restriction to exponential clone growth. In reality a variety of growth profiles are to be expected, depending on the phenotype of a clone's constituent cells, and also the environment in which they are growing. For example, if an outgrowing clone is competing with its parent clone for resources, the growth rate of the outgrowing clone should affect that of the parent. In all generalizations of TSCE, each clone grows independently and exponentially and so this type of competitive behaviour is not considered.

2.3.1 Derivation of the TSCE model

There are many incarnations of TSCE and various approaches to deriving the hazard functions associated with each. The formulation given here represents the core modelling techniques required to use TSCE or its derivatives. It will be drawn upon in section 3.2 to build a novel clonal expansion model. The approach combines the work of Little [Lit96] and Moolgavkar [LM02] and has been designed to be intuitive and generalizable.

In the basic TSCE model, the number of susceptible stem cell lineages is represented by a deterministic function of age $S(t)$. Sometimes this is assumed to

be constant, for example $S(t) \equiv 10^8$. Alternatively $S(t)$ can be chosen to reflect tissue growth as will be discussed for breast cancer models below. Initiated cells are created through mutation at rate $\mu_1(t)$ per cell per year. So, in a small time interval Δt an initiated cell will arise through mutation of a normal cell with probability $\mu_1(t)S(t)\Delta t + o(\Delta t)$. The probability that more than one initiated cell will arise in the interval is $o(\Delta t)$. Equivalently, the number of initiated cells arising in the time interval $[t, t+\tau)$ follows a Poisson distribution with mean $\int_t^{t+\tau} \mu_1(s)S(s)\,ds$.
The age-dependence of μ_1 can be exploited to model periods of exposure to a more carcinogenic environment, or other factors causing age-related variation in mutation rate.

Initiated cells, denoted by 'I' in figure 2.6, once created, grow into a clone and become malignant according to the following rules. Between times t and $t + \Delta t$ each initiated cell can either:

1. divide symmetrically with probability $\alpha(t)\Delta t + o(\Delta t)$, i.e. split into two initiated cells, thus increasing the number of initiated cells by 1,

2. die or differentiate with probability $\beta(t)\Delta t + o(\Delta t)$, thus decreasing the number of initiated cells by 1,

3. divide into one initiated cell and one malignant cell with probability $\mu_2(t)\Delta t + o(\Delta t)$

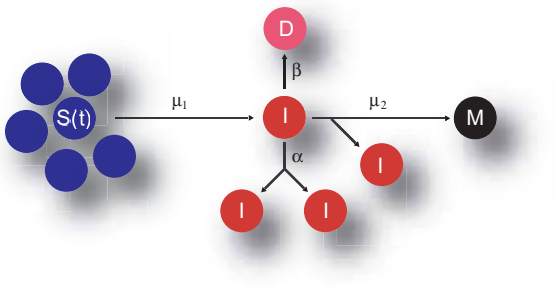

Figure 2.6: The two stage clonal expansion model. $S(t)$ is the number of susceptible stem cells at age t. Although this can be modelled by a stochastic process, often $S(t)$ is taken to be a deterministic function of time ($S(t) \equiv$ 'a constant' for example). The initiated 'I' cells have acquired the first heritable change. They divide symmetrically at rate α per annum and die at rate β per annum. One further change , at rate μ_2 per cell per annum is required to make an intermediate cell malignant.

or,

4. do nothing, with probability $1 - \Delta t(\alpha(t) + \beta(t) + \mu_2(t)) + o(\Delta t)$.

The chance of any other event, for example two of the above occurring together, is vanishingly small ($o(\Delta t)$).

To calculate the hazard function for the basic TSCE process described above, two probability generating functions are used. The main generating function for the process is:

$$\psi[y_1, y_2; t] = \sum_{i_1 \geq 0, i_2 \geq 0} y_1^{i_1} y_2^{i_2} P[\underline{Y}(t) = (i_1, i_2)], \tag{2.6}$$

where $\underline{Y}(t) = (Y_1(t), Y_2(t))$. $Y_1(t)$ is the number of initiated cells at age t and $Y_2(t)$ is the number of malignant cells. Both are created under rules 1 - 4 above, assuming that there are only healthy cells at age zero. i.e. $Y_1(0) = Y_2(0) = 0$. Note that:

$$\psi[1,0;t] = \sum_{i_1=0}^{\infty} P[\underline{Y}(t) = (i_1, 0)] = 1 - P[T \leq t], \tag{2.7}$$

where T is the age at which the first malignant cell occurs. The hazard, $h(t)$, is related to $P[T \leq t]$ by equation (2.4). So by equations (2.4) and (2.7), $h(t)$ can be expressed in terms of ψ:

$$h(t) = -\frac{d}{dt} \ln[\psi[1,0;t]]. \tag{2.8}$$

$\psi[y_1, y_2; t]$ itself can be expressed in terms of a second generating function:

$$\phi[y_1, y_2; t, s] = \sum_{i_1, i_2} y_1^{i_1} y_2^{i_2} P[\underline{Y}(t, s) = (i_1, i_2)].$$

$\phi[y_1, y_2; t, s]$ is the generating function for a process which begins at time s with only a single cell (an initiated cell). Hence, $\underline{Y}(t, s) = (Y_1(t, s), Y_2(t, s))$ is a random vector representing the number of initiated ($Y_1(t, s)$) and malignant cells ($Y_2(t, s)$) at time t, arising from the initial state $(1, 0)$ at time s.

To express the main generating function, ψ, in terms of the subsidiary ϕ, $P[\underline{Y}(t) = (i_1, i_2)]$ can be rewritten as:

$$P[\underline{Y}(t) = (i_1, i_2)] = \sum_k P[c(t) = k] P[\underline{Y}(t) = (i_1, i_2) | c(t) = k], \tag{2.9}$$

where $c(t)$ is the number of initiated cells that have arisen from the healthy cell compartment by age t. Substituting (2.9) into (2.6):

$$\psi[y_1, y_2; t] = \sum_{i_1, i_2} y_1^{i_1} y_2^{i_2} \sum_k P[c(t) = k] P[\underline{Y}(t) = (i_1, i_2) | c(t) = k]. \tag{2.10}$$

$c(t)$ is a Poisson count-process with intensity $\mu_1(t) \cdot S(t)$ so:

$$P[c(t) = k] = \frac{\left(\int_0^t \mu_1(s) S(s)\, ds \right)^k}{k!} \exp\left[- \int_0^t \mu_1(s) S(s)\, ds \right]. \tag{2.11}$$

To complete the expression on the RHS of (2.10) it remains to calculate $P[\underline{Y}(t) = (i_1, i_2) \,|c(t) = k]$:

$$P[\underline{Y}(t) = (i_1, i_2)|c(t) = k] = \sum_{\substack{\underline{N}_1, \underline{N}_2, \ldots, \underline{N}_k \\ \underline{N}_1 + \cdots + \underline{N}_k = (i_1, i_2)}} \prod_{i=1}^{k} P[\underline{Y}(t, s_i) = \underline{N}_i]. \tag{2.12}$$

Here $\underline{Y}(t, s_i) = (Y_1(t, s_i), Y_2(t, s_i))$ is the number of initiated and malignant cells seeded by the ith of the k initiated cells created in the Poisson process. The ith such initiated cell is created at time s_i. The $\underline{N}_i$'s are pairs of integers (n_1, n_2) representing numbers of initiated and malignant cells. They are constrained so that $\underline{N}_1 + .. + \underline{N}_k = (i_1, i_2)$. In summary, formula (2.12) expresses the probability of having i_1 initiated cells and i_2 malignant cells by age t, in terms of the behaviour of the k initiated cells generated from the healthy cell pool. To arrive at the state (i_1, i_2) by age t, the cells arising from the k initiated cells must sum together to make i_1 initiated cells and i_2 malignant ones. In formula (2.12) all combinations in which this is the case are considered in the sum.

When k events occur in the interval $[0, t]$ under an inhomogeneous Poisson process with intensity $\mu(s)S(s)$ the arrival times of these events are independent and identically distributed. The common probability density for the arrival times at $s < t$ is:

$$\frac{\mu_1(s)S(s)}{\int_0^t \mu_1(r)S(r)\,dr}.$$

This PDF can be used to make an expression for $P[\underline{Y}(t, s_i) = \underline{N}_i]$:

$$P[\underline{Y}(t, s_i) = \underline{N}_i] = \frac{\int_0^t \mu_1(s)S(s)P[\underline{Y}(t, s) = \underline{N}_i]\,ds}{\int_0^t \mu_1(r)S(r)\,dr}. \tag{2.13}$$

Substituting (2.13), (2.11) and (2.12) into (2.10) gives:

$$\psi[y_1, y_2; t] = \sum_{i_1, i_2} y_1^{i_1} y_2^{i_2} \sum_k \frac{\left(\int_0^t \mu_1(s) S(s)\, ds\right)^k}{k!} \exp\left[-\int_0^t \mu_1(s) S(s)\, ds\right]$$
$$\sum_{\substack{\underline{N}_1, \underline{N}_2, \ldots, \underline{N}_k \\ \underline{N}_1 + \cdots + \underline{N}_k = (i_1, i_2)}} \prod_{i=1}^{k} \frac{\int_0^t \mu_1(s) S(s) P[\underline{Y}(t,s) = \underline{N}_i]\, ds}{\int_0^t \mu_1(r) S(r)\, dr},$$

swapping the order of summation over the i_js and k gives:

$$\psi[y_1, y_2; t] = \sum_k \frac{\left(\int_0^t \mu_1(s) S(s)\, ds\right)^k}{k!} \exp\left[-\int_0^t \mu_1(s) S(s)\, ds\right] \tag{2.14}$$
$$\sum_{i_1, i_2} y_1^{i_1} y_2^{i_2} \sum_{\substack{\underline{N}_1, \underline{N}_2, \ldots, \underline{N}_k \\ \underline{N}_1 + \cdots + \underline{N}_k = (i_1, i_2)}} \prod_{i=1}^{k} \frac{\int_0^t \mu_1(s) S(s) P[\underline{Y}(t,s) = \underline{N}_i]\, ds}{\int_0^t \mu_1(r) S(r)\, dr}. \tag{2.15}$$

To simplify this expression the following result is used:

Lemma

If y, $x_1, x_2, \ldots$ and x_n all map from a vector space, V, into the real numbers and y is such that:

$$y(a + b) = y(a) \cdot y(b), \forall a, b \in V,$$

then

$$\sum_r y(r) \sum_{\substack{r_1, r_2, \ldots, r_n \\ r_1 + r_2 + \cdots + r_n = r}} \prod_{i=1}^{n} x_i(r_i) = \prod_{i=1}^{n} \left[\sum_r y(r) x_i(r)\right]. \tag{2.16}$$

A proof is given in appendix A. Using the lemma, equation (2.15), can be reduced to:

$$\psi[y_1, y_2; t] = \sum_k \frac{\left(\int_0^t \mu_1(s)S(s)\,ds\right)^k}{k!} \exp\left[-\int_0^t \mu_1(s)S(s)\,ds\right] \prod_{i=1}^{k} \sum_{i_1,i_2} y_1^{i_1} y_2^{i_2} \frac{\int_0^t \mu_1(s)S(s)P[\underline{Y}(t,s) = (i_1,i_2)]\,ds}{\int_0^t \mu_1(r)S(r)\,dr}$$

$$= \sum_k \frac{\left(\int_0^t \mu_1(s)S(s)\,ds\right)^k}{k!} \exp\left[-\int_0^t \mu_1(s)S(s)\,ds\right] \left[\sum_{i_1,i_2} y_1^{i_1} y_2^{i_2} \frac{\int_0^t \mu_1(s)S(s)P[\underline{Y}(t,s) = (i_1,i_2)]\,ds}{\int_0^t \mu_1(r)S(r)\,dr}\right]^k.$$

This expression for $\psi[y_1, y_2; t]$ can be manipulated further by taking the sum over i_1 and i_2 inside the integral of $P[\underline{Y}(t,s) = (i_1,i_2)]$ against the Poisson density:

$$\psi[y_1, y_2; t] = \sum_k \frac{\left(\int_0^t \mu_1(s)S(s)\,ds\right)^k}{k!} \exp\left[-\int_0^t \mu_1(s)S(s)\,ds\right] \left[\frac{\int_0^t \mu_1(s)S(s)\sum_{i_1,i_2} y_1^{i_1} y_2^{i_2} P[\underline{Y}(t,s) = (i_1,i_2)]\,ds}{\int_0^t \mu_1(r)S(r)\,dr}\right]^k$$

$$= \sum_k \frac{\left(\int_0^t \mu_1(s)S(s)\,ds\right)^k}{k!} \exp\left[-\int_0^t \mu_1(s)S(s)\,ds\right] \left[\frac{\int_0^t \mu_1(s)S(s)\phi[y_1, y_2; t, s]\,ds}{\int_0^t \mu_1(r)S(r)\,dr}\right]^k$$

$$= \exp\left[-\int_0^t \mu_1(s)S(s)\,ds\right] \sum_k \frac{\left[\int_0^t \mu_1(s)S(s)\phi[y_1, y_2; t, s]\,ds\right]^k}{k!}.$$

Ultimately

$$\psi[y_1, y_2; t] = \exp\left[\int_0^t \mu_1(s)S(s)[\phi[y_1, y_2; t, s] - 1]\, ds\right]. \tag{2.17}$$

Substituting (2.17) into the expression for the hazard function given in (2.8):

$$h(t) = -\int_0^t \mu_1(s)S(s)\frac{\partial \phi}{\partial t}[1, 0; t, s]\, ds. \tag{2.18}$$

Before the hazard can be computed, it remains to obtain an expression for $\frac{\partial \phi}{\partial t}[1, 0, t; s]$. This can be done via the Kolmogorov forward and backward equations for ϕ. To begin with, the backward equation can be derived as follows:

$$\frac{\partial}{\partial s}\phi[y_1, y_2; t, s] = \sum_{\underline{N}} y_1^{i_1} y_2^{i_2} \frac{\partial}{\partial s} P[\underline{Y}(t, s) = \underline{N}]. \tag{2.19}$$

Here, $\frac{\partial}{\partial s}P[\underline{Y}(t, s) = \underline{N}]$, is calculated as the following limit,

$$\frac{\partial}{\partial s}P[\underline{Y}(t, s) = \underline{N}] = \lim_{h\to 0}\left[\frac{P[\underline{Y}(t, s) = \underline{N}] - P[\underline{Y}(t, s - h) = \underline{N}]}{h}\right].$$

Hence the name 'backward equation'; the partial derivative with respect to s is found by extending the time interval, $t - s$, backward in time. By contrast, in a 'forward equation' the derivative with respect to t is found by incrementing t forward in time.
$P[\underline{Y}(t, s) = \underline{N}] - P[\underline{Y}(t, s - h) = \underline{N}]$ can be expressed as:

$$\Delta(h) = P[\underline{Y}(t, s) = \underline{N}] - \sum_{\underline{N}_T} P[\underline{Y}(s, s - h) = \underline{N}_T] P_{\underline{N}_T}[\underline{Y}(t, s) = \underline{N}],$$

where $\underline{N}_T = (i_1^T, i_2^T)$ is a transitional state between $(1, 0)$ and $\underline{N} = (i_1, i_2)$. Also $P_{\underline{N}_T}[\underline{Y}(t, s) = \underline{N}]$ is the probability of being in state $\underline{N}$ at age t having started in state $\underline{N}_T$ at age s.
$\Delta(h)$ can be rearranged:

$$\Delta(h) = P[\underline{Y}(t, s) = \underline{N}](1 - P[\underline{Y}(s, s - h) = (1, 0)])$$

$$- \sum_{\underline{N}_T \neq (1,0)} P[\underline{Y}(s, s-h) = \underline{N}_T] P_{\underline{N}_T}[\underline{Y}(t,s) = \underline{N}],$$

dividing by h and taking the limit as $h \to 0$ gives:

$$\begin{aligned}
\lim_{h\to 0} \frac{\Delta(h)}{h} &= \frac{\partial}{\partial s} P[\underline{Y}(t,s) = \underline{N}] \\
&= (\alpha(s) + \beta(s) + \mu_2(s)) P[\underline{Y}(t,s) = \underline{N}] - \alpha(s) P_{(2,0)}[\underline{Y}(t,s) = \underline{N}] \\
&\quad -\mu_2(s) P_{(1,1)}[\underline{Y}(t,s) = \underline{N}] - \beta(s) P_{(0,0)}[\underline{Y}(t,s) = \underline{N}]. \qquad (2.20)
\end{aligned}$$

Substituting (2.20) into (2.19) leads to:

$$\begin{aligned}
\frac{\partial}{\partial s}\phi[y_1, y_2; t, s] &= \sum_{\underline{N}} y_1^{i_1} y_2^{i_2} (\alpha(s) + \beta(s) + \mu_2(s)) P[\underline{Y}(t,s) = \underline{N}] \qquad (1) \\
&\quad - \sum_{\underline{N}} y_1^{i_1} y_2^{i_2} \alpha(s) P_{(2,0)}[\underline{Y}(t,s) = \underline{N}] \qquad (2) \\
&\quad - \sum_{\underline{N}} y_1^{i_1} y_2^{i_2} \mu_2(s) P_{(1,1)}[\underline{Y}(t,s) = \underline{N}] \qquad (3) \\
&\quad - \sum_{\underline{N}} y_1^{i_1} y_2^{i_2} \beta(s) P_{(0,0)}[\underline{Y}(t,s) = \underline{N}]. \qquad (4)
\end{aligned}$$

Note that the first component of $\frac{\partial}{\partial s}\phi[y_1, y_2; t, s]$, denoted (1), can be written as:

$$\sum_{\underline{N}} y_1^{i_1} y_2^{i_2} (\alpha(s) + \beta(s) + \mu_2(s)) P[\underline{Y}(t,s) = \underline{N}] = (\alpha(s) + \beta(s) + \mu_2(s))\phi[y_1, y_2; t, s],$$

component (2) can be written as:

$$\begin{aligned}
\sum_{\underline{N}} y_1^{i_1} y_2^{i_2} \alpha(s) P_{(2,0)}[\underline{Y}(t,s) = \underline{N}] &= \alpha(s) \sum_{\underline{N}} y_1^{i_1} y_2^{i_2} \sum_{\underline{N}_H} P[\underline{Y}(t,s) = \underline{N}_H] P[\underline{Y}(t,s) = \underline{N} - \underline{N}_H] \\
&= \alpha(s) \sum_{\underline{N}_H} P[\underline{Y}(t,s) = \underline{N}_H] \sum_{\underline{N}} y_1^{i_1} y_2^{i_2} P[\underline{Y}(t,s) = \underline{N} - \underline{N}_H] \\
&= \alpha(s) \sum_{\underline{N}_H} y_1^{i_1^H} y_2^{i_2^H} P[\underline{Y}(t,s) = \underline{N}_H] \\
&\qquad \sum_{\underline{N}} y_1^{i_1 - i_1^H} y_2^{i_2 - i_2^H} P[\underline{Y}(t,s) = \underline{N} - \underline{N}_H] \\
&= \alpha(s)\phi[y_1, y_2; t, s]^2,
\end{aligned}$$

where $\underline{N}_H$ represents the state achieved by one of the two original initiated cells, making up one half of the target state $\underline{N}$. Components (3) and (4) can be expressed similarly. The resulting PDE for $\frac{\partial}{\partial s}\phi[t,s]$ (suppressing the dependence on y_1 and y_2) is:

$$\frac{\partial}{\partial s}\phi[t,s] = (\alpha(s)+\beta(s)+\mu_2(s)(1-y_2))\phi[t,s] - \alpha(s)\phi[t,s]^2 - \beta(s).$$

Differentiating with respect to t gives:

$$\frac{\partial}{\partial s}\left(\frac{\partial}{\partial t}\phi[t,s]\right) = (\alpha(s)+\beta(s)+\mu_2(s)(1-y_2))\frac{\partial}{\partial t}\phi[t,s] - 2\alpha(s)\phi[t,s]\frac{\partial}{\partial t}\phi[t,s]. \tag{2.21}$$

Fixing t and setting $(y_1, y_2) = (1, 0)$, (2.21) becomes an ODE w.r.t. s and can be solved numerically for $\frac{\partial}{\partial t}\phi[1,0;t,s]$. However, a boundary condition is required. This can be obtained via the Kolmogorov forward equation for $\phi[y_1, y_2; t, s]$ which is derived as follows:

$$\begin{aligned}
\frac{\partial}{\partial t}\phi[y_1,y_2,t;s] &= \sum_{i_1,i_2} y_1^{i_1}y_2^{i_2}\frac{\partial}{\partial t}P[\underline{Y}(t,s)=(i_1,i_2)] \\
&= \sum_{i_1,i_2} y_1^{i_1}y_2^{i_2}\alpha(t)(i_1-1)P[\underline{Y}(t,s)=(i_1-1,i_2)] \qquad (\sharp) \\
&+ \sum_{i_1,i_2} y_1^{i_1}y_2^{i_2}\beta(t)(i_1+1)P[\underline{Y}(t,s)=(i_1+1,i_2)] \\
&+ \sum_{i_1,i_2} y_1^{i_1}y_2^{i_2}\mu_2(t)i_1P[\underline{Y}(t,s)=(i_1,i_2-1)] \\
&- \sum_{i_1,i_2} y_1^{i_1}y_2^{i_2}(\alpha(t)+\beta(t)+\mu_2(t))i_1P[\underline{Y}(t,s)=(i_1,i_2)].
\end{aligned}$$

The component of $\frac{\partial}{\partial t}\phi[y_1,y_2,t;s]$ denoted by $(\sharp)$ can be written as:

$$\sum_{i_1,i_2} y_1^{i_1}y_2^{i_2}\alpha(t)(i_1-1)P[\underline{Y}(t,s)=(i_1-1,i_2)]$$

$$
\begin{aligned}
&= \alpha(t)y_1^2 \sum_{i_1,i_2} (i_1-1) y_1^{i_1-2} y_2^{i_2} P[\underline{Y}(t,s) = (i_1-1, i_2)] \\
&= \alpha(t)y_1^2 \frac{\partial}{\partial y_1} \left[\sum_{i_1,i_2} y_1^{i_1-1} y_2^{i_2} P[\underline{Y}(t,s) = (i_1-1, i_2)] \right] \\
&= \alpha(t)y_1^2 \frac{\partial}{\partial y_1} \phi[y_1, y_2, t; s].
\end{aligned}
$$

The other components can be expressed in the same way to make the Kolmogorov forward equation for $\phi[t,s]$ (suppressing the dependence on y_1 and y_2):

$$
\frac{\partial}{\partial t}\phi[t,s] = \left[\alpha(t)y_1^2 + \mu_2(t)y_1y_2 + \beta(t) - (\alpha(t) + \beta(t) + \mu_2(t))y_1\right] \frac{\partial}{\partial y_1}\phi[t,s].
$$

A boundary condition can now be derived by setting $y_1 = 1$, $y_2 = 0$ and $s = t$:

$$
\frac{\partial}{\partial t}\phi[y_1, y_2; t, s]|_{y_1=1, y_2=0, s=t} = -\mu_2(t), \tag{2.22}
$$

since $\dfrac{\partial}{\partial y_1}\phi[t,s]|_{y_1=1,y_2=0,s=t} = 1$. Numerical integration of (2.21) using the boundary condition (2.22) and substitution into (2.18) yields an expression for the hazard function.

2.4 Likelihood Constructs

In much of the literature on mathematical modelling of incidence, models are fit to observed data using maximum likelihood. In chapters three and four likelihood techniques will be used to estimate how many mutations cause a cancer and also the rates of these mutations. There are standard methods for constructing the requisite likelihood functions. Two such methods are presented here.

2.4.1 Population based age-specific incidence

Population based incidence data are often recorded by age group (table 2.1). So, for example, over a particular year, a cancer registry may have an average of Pop_1 patients within its catchment area that fall into the youngest age group. This group could be all those patients aged between 0 and 4 at last birthday for example. The same registry may reside over Pop_2 patients who fall into the next youngest age group, say patients aged 5 to 9 at last birthday and Pop_3 patients in the third age group etc.. During a year of observation there may be D_i primary cancers of a certain type observed

among the Pop_i patients in the ith age group. What is the probability of observing the data $\{D_1, D_2, D_3, \dots\}$ given the population sizes are $Pop_1, Pop_2, Pop_3, ..$ and given a model M of the cancer in question? Suppose the hazard in an individual under model M is $h(t)$. Make the simplifying assumption that the Pop_i are constant over the observation year. Also make the assumption that cancers occur in the ith age group at rate $Pop_i \cdot h_i$ throughout the year, with $h_i = h(t_i), L_i \leq t_i \leq U_i$ where L_i and U_i are the upper and lower bounds of the age group respectively.

Under these assumptions the observed number of cases, D_i, in the ith age group follows a Poisson distribution with mean $\int_0^1 Pop_i \cdot h_i \; dt$ so:

$$P[D_i] \simeq \frac{(Pop_i \cdot h_i)^{D_i}}{D_i!} \exp[-Pop_i \cdot h_i].$$

Since the members of the different Pop_i's are independent of each other, the likelihood of $\{D_1, D_2, D_3, \dots\}$ given the model, M is:

$$L[\{D_1, D_2, .\}|M] \simeq \prod_i \frac{(Pop_i \cdot h_i)^{D_i}}{D_i!} \exp[-Pop_i \cdot h_i]. \tag{2.23}$$

2.4.2 Non-population based incidence

Data for rare cancers, for example those arising in the context of an inherited syndrome, are usually based on a small sample (< 1000) of patients. Typically, the ages at presentation of individuals (who all eventually get cancer) are available, but the frequency with which the disease occurs in the general population is not. Calabrese and colleagues [CTS04] have derived a likelihood function for such a series of ages at presentation $\{t_1, t_2, ..., t_n\}$. The t_i's are assumed to be independent. The probability that a patient (who eventually gets cancer) presents between ages t_i and $t_i + 1$ can be represented as a conditional probability:

$$\begin{aligned} \omega[t_i] &= P[t_i \leq T < t_i + 1 | T < T_d] \\ &= \frac{P[t_i \leq T < t_i + 1, T < T_d]}{P[T < T_d]}, \end{aligned}$$

here T_d is a random variable representing the age at death of the average patient. $P[t_i]$ is conditional on the age at cancer being less than the age at death. The event $[T < T_d]$

is equivalent to the event that the patient eventually gets cancer. Consider the random vector (T, T_d). If the cancer under consideration is a relatively minor cause of death, it can be assumed that T and T_d are independent [CTS04]. The density at an arbitrary point (t, t_d) is then the product $f_T(t)f_{T_d}(t_d)$ where f_T and f_{T_d} are the density functions for T and T_d respectively. $P[T < T_d]$ can be expressed in terms of these densities:

$$\begin{aligned} P[T < T_d] &= \int_{t=0}^{\infty} \int_{t_d=t}^{\infty} f_T(t) f_{T_d}(t_d)\, dt_d dt \\ &= \int_{t=0}^{\infty} \int_{t_d=t}^{\infty} f_{T_d}(t_d)\, dt_d f_T(t)\, dt \\ &= \int_{t=0}^{\infty} s(t) f_T(t)\, dt. \end{aligned}$$

Here $s(t) = \int_{t_d=t}^{\infty} f_{T_d}(t_d)\, dt_d$ is the survival function giving the probability that $T_d > t$. $s(t)$ can be approximated from known data. Using this notation:

$$\omega[t_i] = \frac{\int_{t=t_i}^{t_i+1} s(t) f_T(t)\, dt}{\int_{t=0}^{\infty} s(t) f_T(t)\, dt},$$

and the likelihood, L, of the data $\{t_1, t_2, t_3, \dots\}$ is:

$$L = \prod_{i=1}^{n} \omega(t_i) = \prod_{i=1}^{n} \left(\frac{\int_{t=t_i}^{t_i+1} s(t) f_T(t)\, dt}{\int_{t=0}^{\infty} s(t) f_T(t)\, dt} \right). \tag{2.24}$$

2.5 Applications of Multistage Modelling

Typical derivatives of TSCE or original multistage theory can be obtained by allowing certain model parameters to vary with time. As mentioned previously, a mutation rate may change with age to reflect changing influences of the tissue micro-environment. Alternatively the number of healthy cell lineages at risk may increase with age to account for tissue growth between conception and adulthood. Another common modification is to incorporate multiple rounds of clonal expansion at different growth rates.

More subtle phenotypic effects like genome destabilization have also been quantified [LW03]. It is most instructive to consider these and other various incarnations of multistage theory in the context of their applications.

2.5.1 Breast Cancer and Clemmesen's Hook

Breast carcinoma, and its dependence on age, is complicated by the temporal sequence of reproductive events beginning with menarche and ending with menopause. The result is a curious looking incidence profile referred to as Clemmesen's hook, so named because of its appearance on log-log paper (figure 2.7). Its shape reflects a rapid increase in incidence starting in the 20s followed by a gentler rise beginning in the late 40s and continuing into old age

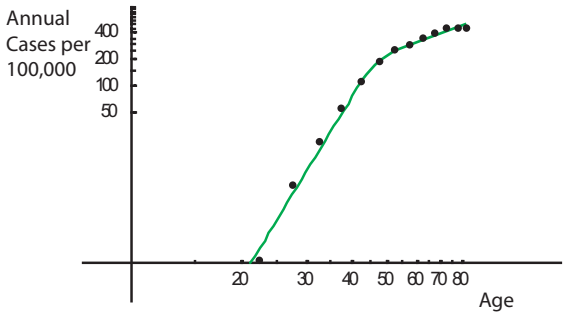

Figure 2.7: Breast cancer incidence is piecewise log-log linear. The first piece has a steeper gradient and gives way to the second phase in the late 40s around the time of the menopause. Data taken from the SEER database - http://.seer.cancer.gov - year of diagnosis 1993-1997 (accessed Sept 20 2007)

.

A derivative of TSCE has been used in a quantitative attempt to explain breast cancer incidence [MDS80]. The basic TSCE model, with a target stem cell pool of constant size, provides a poor fit to data (figure 2.8 - left column), but improvements are made with suitable modifications. First of all, predicted risk of cancer at young ages can be improved by assuming that the susceptible target cells (breast stem cells) grow in number to reflect the development of the breast during puberty (middle column). The excess risk observed to associate with an early menarche follows in this context because the stem cells of the mature breast start to accumulate mutations from an earlier age. To

refine the fit further, reductions in the susceptible cell pool and growth rate of initiated clones can be added to reflect involution of the breast in old age (right column).

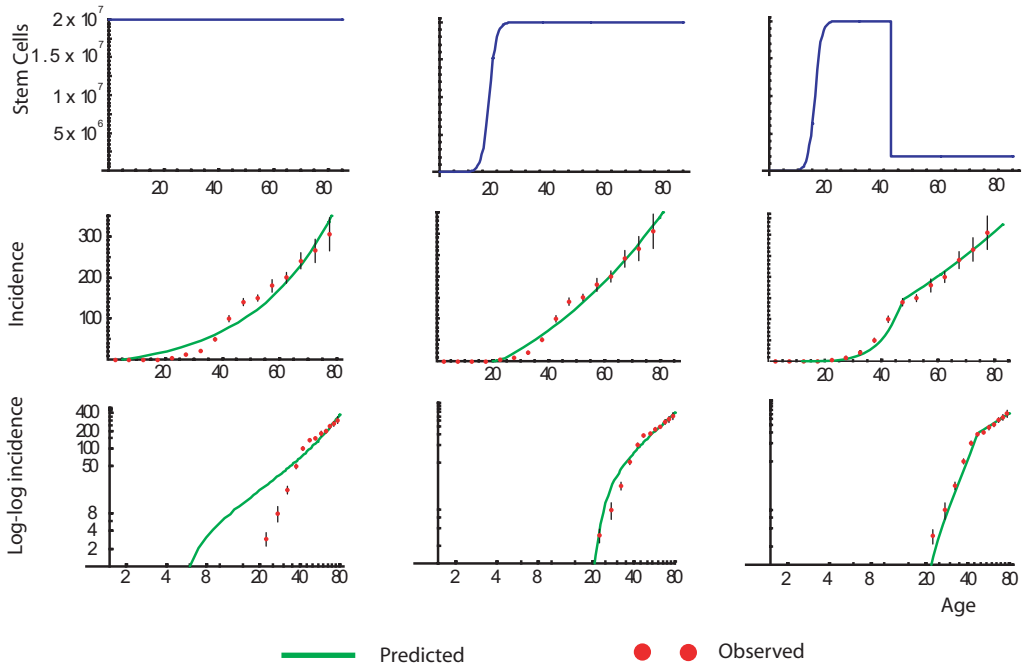

Figure 2.8: All fits performed using a maximum likelihood method. 95% CIs shown for observed data. Different breast stem cell growth patterns in TSCE model of breast cancer. (Left column) Model based on assumption that size of stem cell pool remains constant. (Middle column) Model based on assumption that size of stem cell pool changes to account for development of breast during puberty. (Right column) Size of stem cell pool changes to account for both development of breast during puberty and involution of breast in old age. Observed breast data taken from Moolgavkar et al.[MDS80].

2.5.2 Declining incidence in old age

Although the mellowing of breast cancer risk associated with Clemmesen's hook is unusual, cancer incidence in general begins to plateau above 65 years. Its subsequent decline in some cases is incompatible with an idealized log-log linear cancer onset. This fact was initially dismissed as an error of diagnosis and/or reporting in the elderly. Quantitative arguments have now shown that simple assumptions of population heterogeneity have substantial impact on expected risk and could entirely account for

its decline in old age [HJTMF$^+$00, SMT$^+$06]. Specifically, if a subpopulation is living with genetic susceptibility to a given cancer, among a background population at relatively low risk, then mortality will peak and begin to decline as this subpopulation gradually dies out.

The idea of a subpopulation with elevated cancer risk is compatible with the concept of polygenic susceptibility. For many cancers, the genetic contribution to familial aggregation is only partially accounted for by highly penetrant single gene defects[CLH02]. A portion of the remaining susceptibility may be inherited through several genes. For instance, aggregation of breast cancer is consistent with a polygenic model in which 50% of breast cancers occur in the 12% of the population with the greatest predisposition [PAB$^+$02]. Alternative theories for the peak in cancer incidence among the elderly focus on increases in apoptosis and cell senescence that occur in old age [PW01] or temporal changes in cancer risk [AUAY05].

2.5.3 Smoking and Lung Cancer

Bronchial carcinoma arises in a classic log-log linear fashion in non-smokers and is dramatically more prevalent among the smoking population. Multistage models have been used in attempts to explain excess risk in smoking cohorts, the goal being an understanding of the mechanism through which tobacco smoke exerts its carcinogenic effect. Typically a model of lung cancer is posed for non-smokers. This is then adapted for a given smoking cohort by perturbing parameters (mutation rates for example) from their basal level during the years for which the subjects of the cohort have used cigarettes. The magnitudes of these perturbations are chosen to depend upon the smoking level of the cohorts members, measured in cigarettes per day. Various attempts to elucidate smoking risk in this manner have relied on TSCE to model the underlying incidence in non-smokers. The problem has then been to decipher which of the three phases of the TSCE model, initiation, promotion or progression, is most significantly affected in the smoking cohort. Unfortunately, different studies using slightly different methodology or datasets, have yielded different conclusions. For example, Hazelton et al. [HCM05] and Schollnberger et al. [SMB$^+$06] both used TSCE, with dose responsive parameters, to model incidence of lung cancer among the British Doctors smoking cohort. Hazelton

et al. assumed that kinetic rates among smokers departed from those of non-smokers according to a general power law. So, for example, a smoker of d cigarettes per day, has a mutation rate, μ_s, related to that of a non-smoker, μ_n, by $\mu_s = \mu_n(1 + a \cdot d^b)$ where a and b are free parameters, inferred from data, used to calibrate the model. Schollnberger et al. instead assume $\mu_s = \mu_n \cdot (1 + f(d; a, b))$ where $f(d; a, b) = b \cdot (1 - exp[-(a/b) \cdot d])$. The dose responses of TSCE parameters, as predicted in the two studies, are shown in figure 2.9. Despite very similar methodology, Hazelton et al. emphasize the effect on smoking of initiation and promotion while Schollnberger et al.'s method downplays the relative contribution of promotion. Of interest, however, is Hazelton et al.'s ability to predict risk in ex-smokers. It has long been suggested that the lack of an abrupt fall in risk post quitting, indicates that the final event triggering clonal expansion of a fully malignant bronchial cell is unaffected by smoking [Arm85, Pet01]. The absence of a smoking effect on Hazelton et al.'s progression rate lends some support to this hypothesis.

2.5.4 Prostate cancer and acceleration

The original multistage theory claimed that cancer incidence at age t was proportional to t^{n-1} where n is the number of stages a cell must pass through to become malignant. As has been discussed above, this is an idealization and many cancers stray from the log-log linear relationship. Such cancers have an incidence with a changing, rather than constant, gradient on log-log paper. The gradient at any particular age is referred to as 'age-specific cancer acceleration.' For instance, an idealized log-log linear cancer has a constant acceleration with age. By contrast, breast cancer, in the paradigm of Clemmesens Hook, has a roughly constant acceleration until the late forties and a lower but roughly constant acceleration thereafter. In general, depletion in the number of healthy target cells will result in a reduced acceleration. This was exploited by Moolgavkar et al. in their model of breast cancer incidence [MDS80]. Prostate cancer incidence is distinguished from that of other common epithelial cancers by a dramatic decrease with age (figure 2.10). Frank has used multistage arguments to show how the size and position of an early acceleration peak might depend qualitatively on the number, size and speed of clonal expansions leading to the disease [Fra04b]. Prostate cancer incidence has, however, exhibited strong temporal trends over the past two decades as a result

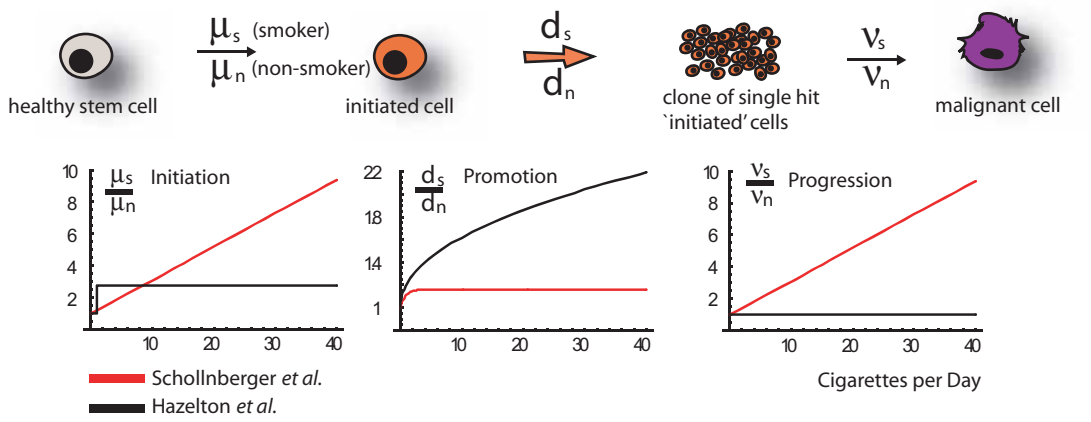

Figure 2.9: The incidence of lung cancer in the British Doctors smoking cohort [DPBS04] was modelled by two separate groups, both using TSCE with smoking-dose-sensitive parameters but assuming different models of dose response. The dose responses implied by both studies are shown. μ_s andμ_n represent the initiation rate in smokers and non-smokers respectively. d_s, d_n and v_s, v_n are the rates of promotion and progression in smokers and non-smokers. The left column shows the ratio $\frac{\mu_s}{\mu_n}$ at different smoking doses, implied by Hazelton and co-workers[HCM05] (black line) model of dose response. The ratio derived by Schollnberger and colleagues [SMB$^+$06] (red line) is also shown. The centre and right columns show the ratios for promotion and progression.

of changing screening practices [KFFM00]. These have dramatically altered the shape of the observed prostate cancer age-distribution (figure 2.10), and are likely to have a significant confounding effect on inferences made from the derivative of this curve.

2.6 Discussion

Quasi-mechanistic modelling of cancer incidence is now a 50 year old discipline. Since its beginning, ideas about cancer aetiology have changed dramatically and, naturally, these changes have been driven by the molecular biology revolution. Multistage models have had to undergo adaptation and revision as the number of alterations in the 'cancer genome' has been shown to be progressively larger than the initially hypothesized 'two hits'. During this process of adaptation and revision, the difficulty of inferring aetiological details from incidence data alone has become apparent. A minor

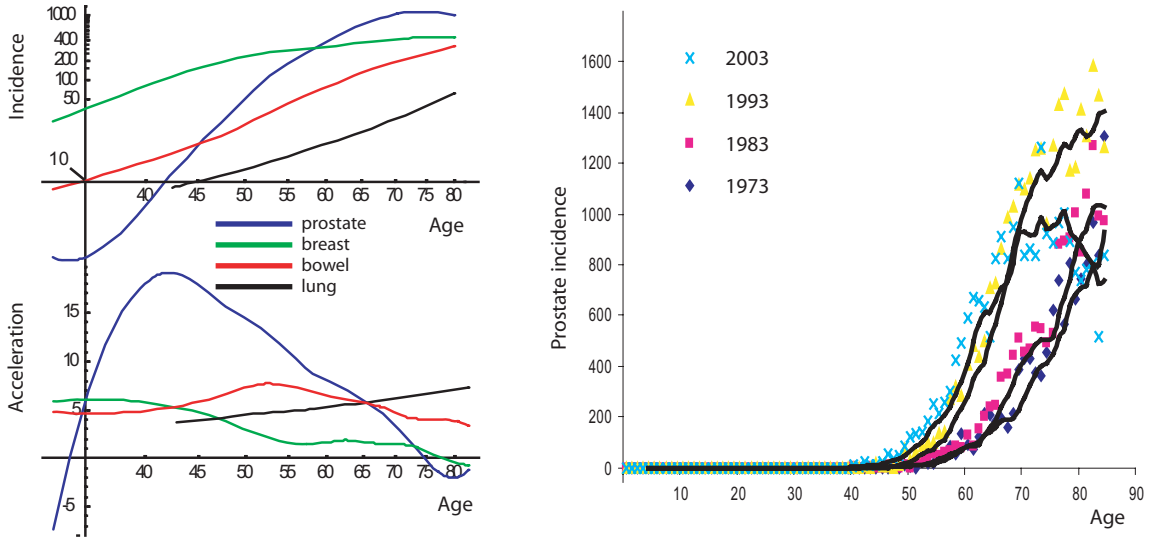

Figure 2.10: (Left) Age-specific acceleration is the slope of the log-log age-incidence plot. Prostate cancer acceleration is very high during the forties and decreases rapidly thereafter. The other common carcinomas hover around an acceleration of five or six. In general, clonal growth and sequences of multiple mutations create acceleration. Depletion in target cell numbers or reductions in clonal growth rates can reduce acceleration. Lung incidence data taken from the CPS2 cohort [SMB+06], all other incidence data taken from the SEER database - http://seer.cancer.gov - year of diagnosis1993-1997 (accessed Sept 20 2007). Acceleration was calculated by interpolating the incidence data, and then taking the log-log slope of the interpolating function. (Right) Changes in the prostate cancer incidence curve with calendar time in Connecticut. Trend lines are five-year moving averages. Data from SEER nine registry, Nov 2007 submission (accessed June 5 2008).

problem is that temporal trends in non-aetiological factors, like screening, can create red herrings in recorded incidence rates, as shown for prostate cancer. Such temporal trends can be resolved to an extent however. The greater difficulty is that several distinct, but equally plausible models, can often be fit to the same incidence pattern, but with different conclusions. This was demonstrated for smoking and lung cancer where conclusions regarding the carcinogenic behaviour of cigarette smoke were sensitive to minor changes in dose-response specification. Quantitative analysis can certainly play a useful role in generating plausible hypotheses for qualitatively interesting features

of age-onset, the aforementioned examples being breast cancer and the general cancer burden in very old age. However, the predictive power of incidence modelling is more questionable. The next chapter focuses on the predictability of the most fundamental multistage parameter of all, the number of mutations or transformations a cell lineage must undergo to become malignant.

Chapter 3

How many mutations are in a cancer?

A tractable quantitative theory of cancer cannot account for every mechanism that might contribute to the disease. In practice, choices must be made about which features are important so that negligible details can be excluded in the name of simplification. For example, while it is clear that extra-cellular factors are crucial in determining the clonal evolution of a tumour, models used to interpret incidence data have tended to concentrate on heritable changes at the genomic level as drivers of this process. Micro-environmental selection parameters that control the relationship between genotype and phenotype are typically accounted for only through fixed clonal growth profiles assumed to associate with a given combination of mutations. It is of crucial importance to understand how such simplifications impact inferences made with the resulting models. Below it is demonstrated that estimates of the number of mutational stages that lead to cancer are sensitive to the assumptions about clonal growth on which they are made. This is illustrated by using two contrasting models to estimate the number of mutations in bowel cancer. Armitage and Doll's model, which assumes no clonal growth, is tested against a model that incorporates a logistically growing precursor lesion. Including a precursor lesion is shown to result in a lower estimate of mutational stages. The sequences of consecutive mutations described in Armitage and Doll's model, and the idea of a clonally expanding precursor lesion, are the two most widely adopted quantitative explanations for a rising cancer incidence with age. Inclusion of a precursor lesion results in a lower estimate of mutational stages, because clonal expansion of the precursor lesion accounts for some of the rise in incidence with age. A shorter sequence of mutations is then required to produce the remainder of the rise in risk. Paradoxically, however, large clonal expansions raise the possibility of mutations

happening very quickly, Such mutations may have a negligible effect on the incidence pattern of a cancer and be undetectable via incidence modelling. Therefore, incorporating clonal expansion can lead to lower estimates of mutation numbers but also throws open the possibility that these are severe underestimates. Those cellular events which are necessary to produce a cancer but that negligibly effect the time taken for the cancer to develop are referred to as non-'rate-limiting'. In the second part of this chapter, a quantitative definition for 'rate-limiting' is outlined and used to determine in what contexts gene mutations are rate-limiting, based on various notional mutation rates and clone sizes of precursor lesions.

3.1 Original Multistage Model

Armitage and Doll originally suggested that late-onset epithelial cancers contained about six mutations [AD54]. This estimate was formed by qualitative comparisons between the predicted incidence of the multistage model and observation. In place of a qualitative inference procedure, Bayesian methods can be used to calculate the probability that the number of mutations, n, is 2, 3, 4, 5, 6, etc. based on the original multistage model and the observed colon cancer data. In this way an idea of the relative quality of fit provided by each value of n can be obtained. The version of Armitage and Doll's model given in chapter 2 (equation 2.5) has only three parameters, n - the number of mutations, μ - the annual mutation rate per cell at all loci and N - the total number of susceptible stem cells. With N fixed at 10^8, a posterior density, $P[\mu, n|\underline{D}]$, can be constructed for N and μ, where $\underline{D} = \{\{D_1, D_2, ..\}, \{Pop_1, Pop_2, ...\}\}$ is the observed data. D_i represents the annual primary disease counts in the ith age group. Pop_i is the size of the ith age group. $N = 10^8$ is justified on the basis that the colon contains of the order of 10^7 crypts [CTS04] and each crypt contains on the order of 10 stem cells [YTS01]. The posterior density at the point (μ, n) can be expressed as:

$$P[\mu, n|\underline{D}] = \frac{L[\underline{D}\,|\mu, n]\pi[\mu, n]}{\sum_i \int_u L[\underline{D}\,|u, i]\pi[u, i]du}, \tag{3.1}$$

where $\pi[\mu, n]$ is the prior density and $L[\underline{D}\,|\mu, n]$ is the likelihood function.

If a uniform prior is taken i.e:

$$\pi[\mu, n] = \begin{cases} \dfrac{1}{(\mu_{\max} - \mu_{\min})(n_{\max} + 1 - n_{\min})} & \mu_{\min} \leq \mu \leq \mu_{\max},\ n_{\min} \leq n \leq n_{\max} \\ 0 & \text{otherwise} \end{cases}$$

then equation 3.1 becomes

$$P[\mu, n|\underline{D}] = \frac{L[\underline{D}\,|\mu, n]}{\sum_{i=n_{\min}}^{n_{\max}} \int_{u=\mu_{\min}}^{\mu_{\max}} L[\underline{D}\,|u, i]du}.$$

By integrating out μ as a nuisance parameter [Siv96], the marginal posterior density for n, $P[n|\underline{D}]$ is:

$$P[n|\underline{D}] = \frac{\int_{u=\mu_{\min}}^{\mu_{\max}} L[\underline{D}\,|u, n]du}{\sum_{i=n_{\min}}^{n_{\max}} \int_{u=\mu_{\min}}^{\mu_{\max}} L[D|u, i]du}. \tag{3.2}$$

An expression for the likelihood function, $L[\underline{D}|\mu, i]$, was previously derived, see equation (2.23). Using (2.23), (3.2) becomes:

$$P[n|\underline{D}] = \frac{\int_{u=\mu_{\min}}^{\mu_{\max}} \prod_j \frac{(Pop_j h(t_j, u, n))^{D_j}}{D_j!} \exp[-Pop_j h(t_j, u, n)]\, du}{\sum_{i=n_{\min}}^{n_{\max}} \int_{u=\mu_{\min}}^{\mu_{\max}} \prod_j \frac{(Pop_j h(t_j, u, i))^{D_j}}{D_j!} \exp[-Pop_j h(t_j, u, i)]\, du},$$

where $h(t, \mu, n) = \dfrac{N\mu^n t^{n-1}}{(n-1)! \sum_{i=0}^{n-1} \frac{(\mu t)^i}{i!}}$ is as given in equation (2.5).

Figure 3.1 shows a graph of $P[n|\underline{D}]$ using incidence data for bowel cancer (ICD 153 - neoplasm of the large intestine, excluding rectum, (eighth revision of the International Classification of Diseases)) in UK males (four regions) taken from Doll [DPW66] (year of diagnosis 1960-1962). To calculate $P[n|\underline{D}]$, the following values were assumed:

- N was fixed at 10^8.

- $\mu_{min} = 10^{-8}$ and $\mu_{max} = 10^{-2}$

- $n_{min} = 2$ and $n_{max} = 8$

Under these assumptions, six mutations is the overwhelming favourite. The sharp peak in the posterior distribution for n arises because $h(t) \simeq \prod \mu_i t^{n-1}$ for small mutation rates μ_i so that the shape of the predicted incidence curve and also the quality of fit is dictated primarily by the exponent $n - 1$ where n is the number of stages.

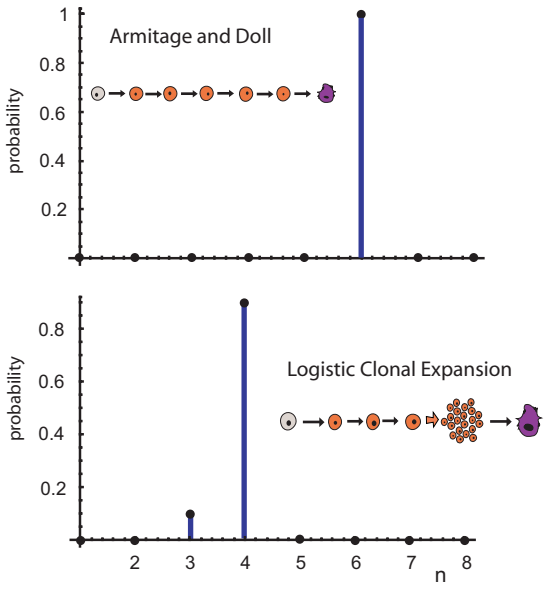

Figure 3.1: Probability (posterior density) of different numbers of mutations implied by observed bowel cancer incidence data. Armitage and Doll's original model strongly implies six mutations. By contrast, a model in which stem cell lineages undergo a slow clonal expansion after receiving a certain number of hits (see text), suggests only three or four mutations. In both fits, mutation rates were constrained to fall in the range $10^{-8} - 10^{-2}$ per cell per year and the total number of mutations must be eight or less. The initial number of healthy target cells was set at 10^8. We used uniform priors for the mutation rates and the mutation numbers.

3.2 Logistic clonal expansion

Suppose the original multistage model is modified slightly and the inference procedure is repeated. Will the estimate of mutation numbers change? Consider a revised multistage model which includes a clonal expansion (figure 3.2). Each cell lineage, upon acquiring an initial number of hits, n_{int}, begins to grow into a clone of lineages (as in the case of the TSCE model). A further n_f hits then convert any cell in the clone into a malignant cell. The total number of hits, n, required to produce a cancer is $n = n_{\text{int}} + n_f$. Rather than using the standard TSCE model, where the clone grows exponentially, a novel model will be presented here in which the clone grows to a limiting capacity in a logistic fashion. This may be a more suitable representation of the behaviour of adenomas in the large intestine [TB95].

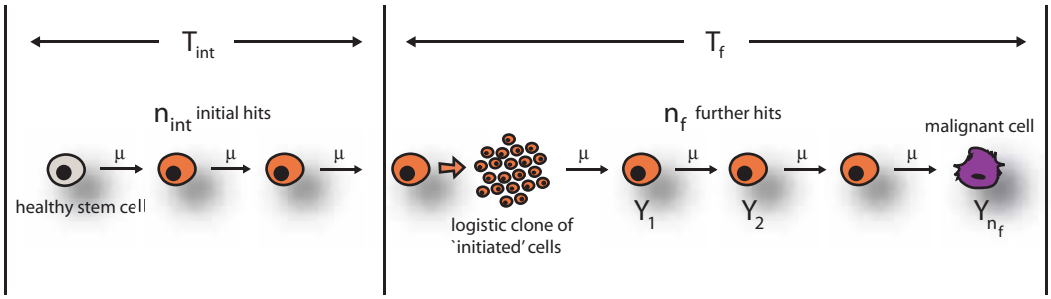

Figure 3.2: In the logistic clonal expansion model, a cell lineage starts to divide symmetrically on receiving n_{int} initial hits. The resulting clone grows logistically. Any cell in the clone can become malignant by receiving n_f further hits.

The time until a single lineage becomes malignant, T_{lin} is equal to $T_{\text{int}} + T_f$. T_{int} is the time taken for the initial n_i hits to occur, initiating the cell lineage. T_f is the time taken for the the initiated lineage to produce a malignant offspring with a further n_f hits. The distribution of $T_{\text{int}} + T_f$ can be expressed as:

$$P[T_{\text{int}} + T_f \leq t] = \int_0^t f_{T_{\text{int}}}(s) P[T_f \leq t - s] ds, \tag{3.3}$$

where $f_{T_{\text{int}}}(s)$ is the density function for T_{int}. Assuming each of the initial hits occurs at rate μ per year, (2.3) can be used for $P[T_{\text{int}} \leq t]$:

$$P[T_{\text{int}} \leq t] = 1 - \sum_{i=0}^{n_{\text{int}}-1} \frac{\mu^i t^i}{i!} \exp[-\mu t].$$

Differentiating with respect to t gives the density:

$$f_{T_{\text{int}}}(t) = \frac{\mu^{n_{\text{int}}} t^{n_{\text{int}}-1}}{(n_{\text{int}}-1)!} \exp[-\mu t]. \tag{3.4}$$

$P[T_f \leq t]$ can be calculated with a method similar to that used to derive the TSCE model. If $\psi[t, y_1, \ldots, y_{n_f}; t]$ is the generating function for the process which starts at $t = 0$ with one initiated cell lineage (i.e. one which has acquired the first n_{int} hits):

$$\begin{aligned}\psi[y_1, \ldots, y_{n_f}; t] &= \sum_{i_1,\ldots,i_{n_f}} P\left[\begin{pmatrix} Y_1(t) \\ \vdots \\ Y_{n_f}(t) \end{pmatrix} = \begin{pmatrix} i_1 \\ \vdots \\ i_{n_f} \end{pmatrix}\right] y_1^{i_1} \ldots y_{n_f}^{i_{n_f}} \\ &= \sum_{i_1,\ldots,i_{n_f}} P\left[\,\underline{Y}(t) = \underline{n}\,\right] y_1^{i_1} \ldots y_{n_f}^{i_{n_f}},\end{aligned}$$

where

$$\underline{n} = \begin{pmatrix} i_1 \\ \vdots \\ i_{n_f} \end{pmatrix}$$

.

Here $Y_i(t)$ is the number of lineages with i further hits at age t.

To represent the growth of the original initiated cell lineage, a continuous function of age, $X(t)$, can be used. Then, by a generalization of equation (2.17) it can be shown that:

$$\psi[y_1, \ldots, y_{n_f}; t] = \exp[\int_0^t \mu X(s)(\phi[y_1, \ldots, y_{n_f}; t, s] - 1)\, ds],$$

where ϕ is the generating function for the process which starts at time s in the state:

$$\begin{pmatrix} Y_1(s,s) \\ \vdots \\ \vdots \\ Y_{n_f}(s,s) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

If the extra hits all occur at rate μ per year then it is possible to give an explicit formula for ϕ:

$$\phi[y_1, \ldots, y_{n_f}; t, s] = \sum_{i_1,..,i_{n_f}} y_1^{i_1} \ldots y_{n_f}^{i_{n_f}} P\left[\underline{Y}(t,s) = \underline{n}\right].$$

$P[\underline{Y}(t,s) = \underline{n}]$ is only non-zero when $\underline{n} = \underline{e}_1, \underline{e}_2, \underline{e}_3$ etc., where $\underline{e}_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 in the ith row zero in all other rows. So

$$\phi[y_1, \ldots, y_{n_f}; t, s] = \sum_{i=1}^{n_f} y_i P\left[\underline{Y}(t,s) = \underline{e}_i\right].$$

From equation (2.1) we know that:

$$P\left[\underline{Y}(t,s) = \underline{e}_i\right] = \frac{\mu^{i-1}(t-s)^{i-1}}{(i-1)!} \exp[-\mu(t-s)], \qquad i = 1, \ldots, n_f - 1$$

and

$$P[\underline{Y}(t,s) = \underline{e}_{n_f}] = 1 - \sum_{i=1}^{n_f-1} P\left[\underline{Y}(t,s) = \underline{e}_i\right].$$

Accordingly,

$$\phi[y_1, \ldots, y_{n_f}; t, s]$$
$$= \exp[-\mu(t-s)] \sum_{i=0}^{n_f-2} \frac{y_{i+1}\mu^i(t-s)^i}{i!} + y_{n_f}\left(1 - \exp[-\mu(t-s)] \sum_{i=0}^{n_f-2} \frac{\mu^i(t-s)^i}{i!}\right),$$

so

$$\psi[y_1, \ldots, y_{n_f}; t] = \exp\left[\int_0^t \mu X(s) \left(\exp[-\mu(t-s)] \sum_{i=0}^{n_f-2} \frac{y_{i+1}\mu^i(t-s)^i}{i!} + y_m n_f \left(1 - \exp[-\mu(t-s)] \sum_{i=0}^{n_f-2} \frac{\mu^i(t-s)^i}{i!}\right) - 1\right) ds\right]$$

and

$$P[T_f \leq t] = 1 - \psi(1, 1, \ldots, 1, 0; t) \tag{3.5}$$

$$= 1 - \exp\left[\int_0^t \mu X(s) \left(\exp[-\mu(t-s)] \sum_{i=0}^{n_f-2} \frac{\mu^i (t-s)^i}{i!} - 1\right) ds\right]. \tag{3.6}$$

Substituting (3.6) and (3.4) into (3.3) gives an expression for $P[T_{lin} \leq t]$:

$$P[T_{\text{lin}} \leq t] = P[T_{\text{int}} + T_f \leq t] = \int_0^t f_{T_{\text{int}}}(s) P[T_f \leq t - s] ds$$

$$= \int_0^t \frac{\mu^{n_{\text{int}}} s^{n_{\text{int}}-1}}{(n_{\text{int}}-1)!} \exp[-\mu s]$$

$$\left(1 - \exp\left[\int_0^{t-s} \mu X(r) \left(\exp[-\mu(t-s-r)] \sum_{i=0}^{n_f-2} \frac{\mu^i (t-s-r)^i}{i!} - 1\right) dr\right]\right) ds.$$

Taking account of the N lineages, the distribution of the time until cancer, T, is:

$$P[T \leq t] = 1 - (1 - P[T_{\text{lin}} \leq t])^N. \tag{3.7}$$

It remains to specify the growth profile $X(t)$. A logistic type growth is given by:

$$X(t) = \frac{K \exp[rt]}{K + \exp[rt] - 1}$$

This is the solution to $\dot{X} = rX\left(1 - \frac{X}{K}\right)$ and $X(0) = 1$.

Fixing K and r, the posterior density at $(\mu, n_{\text{int}}, n_f)$ can be expressed as:

$$P[\mu, n_{\text{int}}, n_f | \underline{D}] = \frac{L[\underline{D} \,|\mu, n_{\text{int}}, n_f] \pi[\mu, n_{\text{int}}, n_f]}{\sum_{i,j} \int_u L[\underline{D} \,|u, i, j] \pi[u, i, j] du}.$$

With a uniform prior, the posterior reduces to:

$$P[\mu, n_{\text{int}}, n_f | \underline{D}] = \frac{L[\underline{D} \,|\mu, n_{\text{int}}, n_f]}{\sum_{i=n_{\text{int}}^{min}}^{n_{\text{int}}^{max}} \sum_{j=n_f^{\min}}^{n_f^{\max}} \int_{u=\mu_{\min}}^{\mu_{\max}} L[\underline{D} \,|u, i, j] du}.$$

Accordingly, the posterior density at a specific value of n , where $n = n_{\text{int}} + n_f$ is given by:

$$P[n|\underline{D}] = \frac{\sum\limits_{\substack{i,j \\ i+j=n}} \int L[\underline{D}\,|u,i,j]\,du}{\sum\limits_{i=n_{\text{int}}^{\min}}^{n_{\text{int}}^{\max}} \sum\limits_{j=n_f^{\min}}^{n_f^{\max}} \int_{u=\mu_{\min}}^{\mu_{\max}} L[\underline{D}\,|u,i,j]du}.$$

For the likelihood function, $L[\underline{D}\,|\mu, n_{\text{int}}, n_f]$, equation (2.23) can be used with $h(t)$ found by differentiating equation (3.7) according to equation (2.4). This is computationally expensive however. To avoid the need to perform calculus on equation (3.7), h_i can be approximated with

$$h_i \simeq \frac{P[T \leq \hat{t}+1] - P[T \leq \hat{t}]}{1 - P[T \leq \hat{t}]}, \tag{3.8}$$

where $\hat{t}$ lies in the ith age group.

With the carrying capacity, K, fixed at 10^6 and $r = 0.1$ years^{-1} the growth profile of the clone is shown in figure 3.3. N was fixed at 10^8 and uniform priors were taken on $1 \leq n_{\text{int}}, n_f \leq 8$ and $10^{-8} \leq \mu \leq 10^{-2}$. Calculations of $P[n|\underline{D}]$ now strongly favour 3 or 4 mutations, rather than 6, with 4 the favourite (figure 3.1). This is a moderate change in conclusion and indicates the instability of the result. If more mechanistic details were included, the uncertainty over mutation numbers would increase further. Unfortunately, there are many more things to consider. At the cellular level, context dependent rates of mutation and selection-driven growth of benign precursor lesions [ISTB99] both play a critical role. At the tissue level, the number and dynamics of healthy target cells are important modulators of incidence. At the population level, genetic heterogeneity among patients can create subpopulations with distinctive risk patterns [PM00]. Similarly, as mentioned previously, temporal trends in incidence can distort population statistics [LM02]. Such trends arise either through real changes in risk, owing, for example, to changes in lifestyle and environment, or sometimes superficially through over-diagnosis or more rapid detection associated with the introduction

of a novel screening program [QB02]. With many confounders, it is extremely difficult to isolate the effects of mutation numbers. This fact is hidden by the standard approach to fitting quasi-mechanistic models of carcinogenesis, an approach which (i) treats a single, predefined clonal growth structure as if it were definitive, and (ii) takes the optimized state of this model (i.e. the parameter values which give the best fit) as a starting point for making inferences. The result is often to exaggerate the specificity of conclusion that can be drawn from the data. Zhang and Simon [ZS05] or Luebeck and Moolgavkar [LM02], in attempts to estimate the number of rate-limiting stages in breast and colorectal cancer respectively, both employ models that depend on a plausible but narrowly constrained description of clonal growth and mutation. These models cannot readily be used for quantifying physiopathological mechanisms of cancer. Translating uncertainty regarding tumorigenesis accurately into statistical uncertainty regarding mutation numbers, or other mechanistic features, is very difficult. However, steps in this direction could be made by, for example, working with a representative collection of possible model structures, rather than a single model. A posterior density for any parameter value of interest could then be calculated by averaging the posterior distributions under each of the models considered, weighted by their posterior model probability [HMRV99].

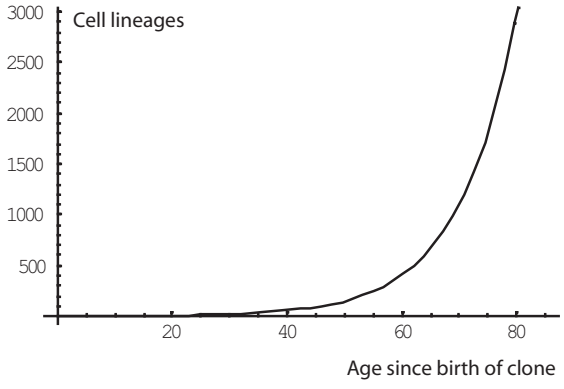

Figure 3.3: The deterministic growth profile of clones in the logistic clonal expansion model.

3.3 Evidence from cancer genome projects

Besides being unstable under changes in model structure, multistage model predictions appear to be inconsistent with evidence generated through the systematic study of cancer genomes. Quantitative analyses of incidence data have conventionally implied ≤ 10 mutational stages in, for example, human breast and colorectal cancer. Evidence produced in a screen of $\backsim 13,000$ genes in cell lines and xenographs derived from these tumour types, however, suggests approximately 14 and 20 genes could be altered via selected mutations in the average colorectal and breast cancer respectively [SJW$^+$06]. Without doubt, these are rough estimates, due to the difficulties inherent in distinguishing genuine selected somatic mutations from passenger mutations or artefacts of sequencing and PCR. Nonetheless, a more recent study of $\backsim 500$ protein kinase genes [GSS$^+$07], across $\backsim 200$ cancer types, using different methods to identify selected mutations, also suggests that a larger number of functionally altered genes than previously anticipated are operative in many human cancers. The apparent discrepancy between low mutation numbers predicted by multistage models and the larger number of alterations found in cancer genomes has a standard explanation; that only certain critical mutations limit the rate at which a cancer is formed. Other mutations, while making essential contributions to the cancer phenotype, occur more quickly than their critical counterparts, for example during the clonal evolution of an established cancer, and are not 'rate-limiting'.

3.4 Rate-Limiting Events

The concept of a rate-limiting step (RLS) originates in the quantitative study of chemical reactions wherein several precise mathematical definitions have been suggested to identify such a step [Tur90]. Common among these definitions is the idea that changes in the speed of a RLS must have a significant impact on the rate of the overall chain of events to which the RLS belongs. As applied to cancer modelling the term is used colloquially and without a strict meaning.

From an incidence modelling perspective, a RLS could be defined as a step whose consequences can be observed by looking at age-distributions. The relevant question in this context is how quick must a mutational / transformational step become before

it ceases to be visible in the age-onset pattern? A rough approach to addressing this question is to (i) build a simple model of tumorigenesis, (ii) fit this model to the incidence distribution of a specific cancer, (iii) add in fast steps of a given rate and observe the effect on quality of fit and (iv) increase the common rate of the extra steps until the effect on quality of fit becomes negligible.

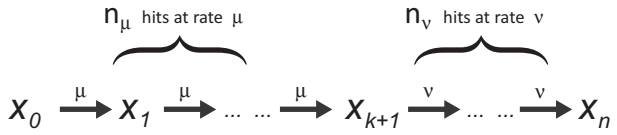

Figure 3.4: A multistage model to test the feasibility of determining 'fast' cellular events from incidence data. A cell lineage becomes malignant following n_μ slow steps (at rate μ), followed by n_ν fast steps at rate ν. When the ν is large enough, the quality of fit becomes insensitive to n_ν.

For parts (i) and (ii) of this method, the Armitage and Doll formula (2.5) can be fit to bowel cancer data, with the number of cell lineages fixed at $N = 10^8$. This is done using the standard maximum likelihood method via the likelihood function (2.23). Adding in more steps to the Armitage Doll model at a faster rate (part (iii)) leads to the Markov chain, shown in figure 3.4. There are n_μ slow hits, followed by n_ν fast hits. The Kolmogorov equations for this system are:

$$\begin{aligned}
\frac{d}{dt}[X_0(t)] &= -\mu X_0(t) \\
\frac{d}{dt}[X_1(t)] &= \mu(X_0(t) - X_1(t)) \\
&\vdots \\
\frac{d}{dt}[X_{k+1}(t)] &= \mu X_k(t) - \nu X_{k+1}(t) \\
&\vdots \\
\frac{d}{dt}[X_{n-1}(t)] &= \nu(X_{n-2}(t) - X_{n-1}(t)) \\
\frac{d}{dt}[X_n(t)] &= \nu X_{n-1}(t).
\end{aligned}$$

where $k = n_\mu - 1$ and $n = n_\mu + n_\nu$. This system is solved by:

$$X_i(t) = \begin{cases} \dfrac{\mu^i t^i}{i!} \exp -\mu t & i = 0, \ldots, k \\ \nu^{i-k-1}\mu^{k+1} \exp[-\mu t] \displaystyle\sum_{j=0}^{k} \dfrac{\binom{i-k-1+j}{j}(-1)^j t^{k-j}}{(\nu-\mu)^{i-k+j}(k-j)!} & \\ + \ \nu^{i-k-1}\mu^{k+1} \exp[-\nu t] \displaystyle\sum_{j=0}^{i-k-1} \dfrac{\binom{k+j}{j}(-1)^{k+1} t^{i-k-1-j}}{(\nu-\mu)^{k+j-1}(i-k-1-j)!} & , \quad i = k+1, \ldots, n-1 \end{cases} \tag{3.9}$$

and

$$X_n(t) = 1 - \sum_{i=0}^{n-1} X_i(t),$$

so that

$$\begin{aligned} P[T \leq t] &= 1 - (1 - X_n(t))^N \\ &= 1 - (\sum_{i=0}^{n-1} X_i(t)). \end{aligned} \tag{3.10}$$

Consider extra stages occurring at rate $\hat{\nu}$. To calculate the quality of fit associated with a particular value of n_ν a uniform prior on $n_\nu = 0, 1, \ldots, 5$ can be taken, leading to a posterior density at $n_\nu = i$ given by:

$$P[n_\nu = i|\underline{D}] = \frac{L[\underline{D}|n_\nu = i, n_\mu = \hat{n}, \mu = \hat{\mu}, \nu = \hat{\nu}]}{\sum_{j=0}^{5} L[\underline{D}|n_\nu = j, n_\mu = \hat{n}, \mu = \hat{\mu}, \nu = \hat{\nu}]}$$

where $n_\nu = \hat{n}$ and $\mu = \hat{\mu}$ optimize $P[\underline{D}|n_\nu = 0, n_\mu, \mu]$.

It is again more computationally efficient to use approximation (3.8) for constructing the likelihood function, $L[\underline{D}|n_\nu = j, n_\mu = \hat{n}, \mu = \hat{\mu}, \nu = \hat{\nu}]$, as this avoids the need to perform calculus on (3.10). Figure 3.5 shows the relative quality of fit to bowel cancer incidence data obtained by adding between 0 and 5 extra stages. These stages have an expected duration of one year or less. One extra stage of three months has a negligible effect, three extra steps of one month go similarly unnoticed.

3.4.1 When is a mutation rate-limiting?

At first glance it does not seem to matter much that it is impossible to sense, at the population level, the effects of an event that takes less than 6 months on average to

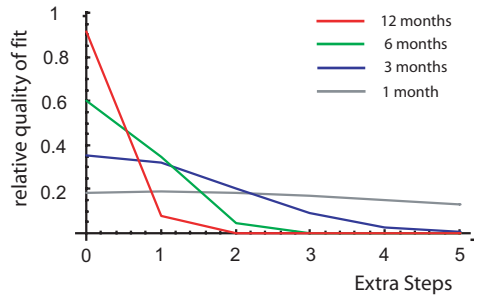

Figure 3.5: Relative quality of fit achieved by adding in one to five extra stages to an optimized multistage model of bowel cancer. The optimized model assumes 10^8 stem cells at risk and that cancer occurs after six mutations each occurring with probability 7×10^{-4} per cell per year. Quality of fit is measured as a posterior density on the number of extra stages. A uniform prior on 0 - 5 extra stages was assumed. Fast steps, expected to occur in less than six months (i.e. with a mutation rate, ν, larger than 2 per cell per year) have a small effect on incidence and so the quality of fit does not decline substantially when these are added. Bowel cancer data taken from [DPW66] (year of diagnosis 1960-1962).

occur. A gene mutation with a rate of 10^{-6} per cell per year for example, is expected to take a million years if it can arise in only one cell lineage. By contrast, the effect of an equally rare event will go unnoticed if it can afflict any of a large population of target cells. For example, among 10^8 healthy target cells, assuming a gene mutation rate of 10^{-8} per cell with each cell division, we would expect the first mutated gene within two divisions. Although the initial step in a genetic pathway is likely to happen very quickly when the number of healthy target cells is large, in many cases the second mutation must arise from a single cell and is expected to take much longer. Towards the end of a genetic pathway, the situation can be reversed again because mutations may arise in a substantial precursor lesion or early stage cancer. How large must such a clone be before one, two, or more sequential mutations cease to be rate-limiting?

The expected time taken for n mutations, occurring at rate μ to arise in any one of N target cell lineages is given by:

$$\int_0^\infty f_T[t]t\,dt,$$

where $f_T[t]$ is the density function for T the time until cancer, found by differentiating (2.3):

$$\begin{aligned} f_T[t] &= \frac{d}{dt}[P[T \leq t]] \\ &= -N\left(\sum_{i=0}^{n}\frac{(\mu t)^i}{i!}\exp^{-\mu t}\right)^{N-1}\frac{d}{dt}\left(\sum_{i=0}^{n}\frac{(\mu t)^i}{i!}\exp^{-\mu t}\right) \\ &= N\left(\sum_{i=0}^{n}\frac{(\mu t)^i}{i!}\exp^{-\mu t}\right)^{N-1}\frac{\mu^n t^{n-1}}{(n-1)!}\exp^{-\mu t}. \end{aligned}$$

Table 3.1 shows the expected time for mutations to occur in cancer stem cell clones of varying sizes. In a very large clone of 1 billion cells (enough to constitute a clinically apparent tumour), two consecutive mutations need not be rate-limiting if they occur with probability 10^{-6} per cell division or higher.

CloneSize	Hits	Mutation Rate 10^{-5}	10^{-6}	10^{-7}
10^7	n=1	10^{-4}	10^{-3}	0.01
	n=2	0.40	3.96	39.64
	n=3	7.55	75.48	754.78
10^8	n=1	10^{-5}	10^{-4}	10^{-3}
	n=2	0.13	1.25	12.53
	n=3	3.50	34.99	349.94
10^9	n=1	10^{-6}	10^{-5}	10^{-4}
	n=2	0.04	0.40	3.96
	n=3	1.62	16.23	162.34

Table 3.1: Expected time lapse in years before one, two or three specific mutations occur in any of a clone of target cells. Clone size is measured in cells. Hits refers to the number of specific gene mutations that are to occur in any one cell of the clone. The mutation rate is quoted per cell per cell generation, assuming 100 generations per year. Since a continuous model of mutation is assumed, mutations can occur at any time and are not limited to fixed points in the cell cycle. This explains why the expected time lapse is less than one cell generation time in some cases

So it is reasonable to assume that many non-rate-limiting mutations occur once the tumour mass has reached a substantial size. This could explain why the age-onset pattern of bowel cancers that have only acquired the potential for local invasion is almost indistinguishable from that of tumours which are aggressively metastasizing and widespread [CMJ+05]. The implication is that clinical stage depends on non-rate-limiting mutations (table 3.1) or other events which occur with high frequency after malignant transformation.

3.5 Discussion

In this chapter it was shown that assumptions about clonal growth can materially effect estimates, made from incidence statistics, of the number of mutations in a given cancer. This is significant because there is real uncertainty over clonal expansion patterns in tumorigenesis. Clonal expansion controls the impact of a given mutation and can determine whether or not subsequent mutations are rate limiting. A novel definition of 'rate-limiting' was given in terms of the observability of a carcinogenic cellular event through registry data. The aim of introducing a concrete metric for measuring the property of being rate limiting is to provide a quantitive framework through which to judge the efficacy and realistic scope of multistage modelling. A marked aspect of this definition is its dependence on the size of the registry in question. It would be interesting to investigate this dependence. For example, could a large patient database notably improve the observability of fast cellular events? Working with bowel cancer data from a UK registry, the 'rate-limiting' definition given was used to show that gene mutations can be undetectably fast, provided they target a sufficient number of cells. Therefore, the concept of a rate-limiting step, provides some explanation for the discrepancy between multistage models predicting low mutation numbers and investigations into cancer genotypes that implicate a greater number of significant DNA modifications. It has been argued, by Moolgavkar et al. [LM02] that if a given gene is inactivated or modified to cause a large clonal expansion, and that if certain critical mutations following this clonal growth happen very quickly (because they can target any cell in the large clone), then it is sufficient to model only the initial mutation causing the clonal growth. The other mutations can be viewed as inevitable consequences of the initial mutation. This argument is a justification for a simple model of tumorigenesis with a small num-

ber of mutations. However, if many mutations conspire together to create an observed incidence profile, it is unlikely that they can be partitioned squarely into those which are rate-limiting and those which are not. The observability of an event is a continuous property and some mutations will occupy the grey area between rate-limiting and not so. For example, one mutation in a clone of a given size may not be rate-limiting but two such mutations may become so. Further, these two mutations together targeting a clone of fixed size could be mistaken for a single mutation in a larger growing clone. To make progress in deciphering the relationship between incidence and aetiology, methods for isolating the effects of a particular mutation or other carcinogenic cellular event are required. In the next chapter such methods are discussed and applied.

In summary, it has been shown in this chapter that quantitative attempts to derive information from age-distributions are sensitive to the assumptions about cancer on which they are made. Therefore, if incidence data are used naively, a false sense of confidence is created over the specificity of conclusion that can be drawn. Care must be taken to ensure that inferences made adequately reflect our current uncertainty over cancer biology as well as our understanding of it.

Chapter 4

Comparative studies of risk in inherited and sporadic tumours

Biologically based models of cancer incidence have not fulfilled their early promise of generating quantitative results on aetiology. How might a further understanding of age-incidence and its dependence on underlying cell and molecular biology be achieved? One approach is to compare the incidence patterns of sporadic and hereditary forms of the same cancer. Observed differences in these incidence patterns can then be ascribed to the gene defect that underlies the hereditary disease. A pioneering example of this type of analysis compared the incidence of familial and sporadic Retinoblastoma. Retinoblastoma is a rare childhood cancer of the nervous system. It is initiated by inactivation of the tumour suppressor gene, *RB1*, in the developing retina. Before the identification of *RB1*, and its role in Retinoblastoma, Knudson [Knu71] had shown that children with familial retinoblastoma develop tumours with an age-onset pattern suggestive of one less causative mutation than in sporadic patients. This finding supported the theory that both alleles of a specific gene must be silenced before a tumour can develop. In familial patients, one of the alleles is already mutated in the germline, causing a shift in age-onset pattern consistent with one less causative mutation.

Retinoblastoma exemplifies a simple relationship between sporadic and hereditary cancer whereby the germline mutation causing the hereditary cancer features as an initiating somatic change in the sporadic cancer. In such a case, patients with a germline mutation are very much like sporadic patients; the increase in risk they experience arises only because their cell lineages begin life with a head-start of one cancerous mutation. In the first part of this chapter, a sporadic and hereditary bowel cancer that

follow this pattern are compared. The hereditary bowel cancer syndrome, familial adenomatous polyposis coli (FAP), is caused by germline mutation in the *APC* gene. *APC* mutations also feature as initiating somatic events in sporadic bowel cancer. The respective incidence patterns of FAP and sporadic bowel cancer are used to estimate the rate of the somatic *APC* mutation which separates them.

4.1 Estimating the rate of *APC* mutation

Cancer arises through successive somatic mutations/epimutations of oncogenes and tumor-suppressor genes. Accurate estimates of the rates at which these (epi)mutations occur are a vital but missing link in our emerging quantitative understanding of tumorigenesis. Their absence has hindered arguments concerning the importance of genetic instability in tumorigenesis and the number of mutations that precede malignant conversion of healthy cell lineages. In this section, a novel method for calculating the in-vivo mutation rate of the *APC* tumor-suppressor gene is presented. The large majority of bowel cancers are thought to be initiated by a partial loss of APC function. Consequently, bowel cancer risk is dramatically altered for the worse in the hereditary syndrome, familial adenomatous polyposis (FAP). This is because FAP patients already harbor germline *APC* mutations so that their cancers require one less genetic aberration at APC. Below, the extra time taken for bowel cancers to develop in sporadic patients is used to estimate the rate of their extra initiating *APC* mutation. A result of approximately 10^{-5} mutations per allele per year, although faster than previous estimates, appears consistent with the high number of different mutations known to target *APC*.

4.1.1 Previous estimates of somatic gene mutation rates

The in vivo rate of somatic tumour suppressor or oncogene mutation cannot be measured directly. Often, it is not possible to isolate the precursor cells of a given cancer or, as a consequence, to probe their DNA for abnormalities. In cases where precursor cells can be isolated, the rarity of any specific mutation makes direct measurements of mutation rate impractical. Even in human cell cultures, determining mutation rate is very difficult [KF88] and has only been possible at a handful of loci. Estimates fall between 10^{-8} and 10^{-6} mutations per gene per cell generation [DH72, SKT^{+}87, AGZ^{+}05].

Similar rates have been observed in yeast [LH89, YK90].

These data provide only a blurred picture of the somatic mutation rate in humans. Aside from the three orders of magnitude over which they are spread, there is a question mark over how well the mutational characteristics of cultured cells mirror those of cells in vivo. Consequently, it is currently very difficult to ascribe a notional rate to tumour suppressor or oncogene mutation. This is especially apparent when considering that the rate in a particular case will depend not only on micro-environmental factors that modify replication fidelity, but also on the specific spectrum of genetic changes that lead to a selected mutant protein product.

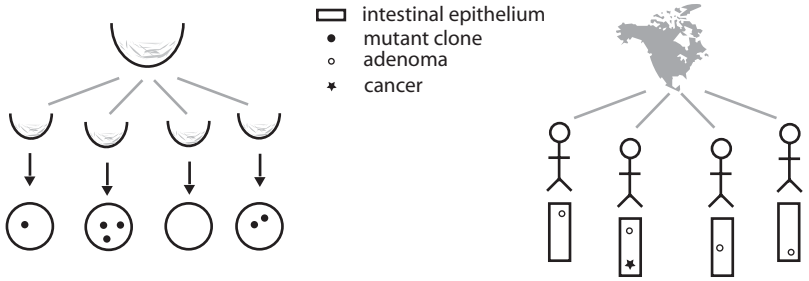

Figure 4.1: In Luria-Delbruck fluctuation analysis (left), the number of mutant colonies arising in plated clones can be used to estimate in vitro mutation rates. By analogy, in vivo, the number of tumours arising in individuals can be used to estimate mutation rates.

Without resorting to cell lines, what alternative methods are available for measuring mutation rates? In vitro, the standard method is Luria-Delbruck fluctuation analysis (see figure 4.1). Many parallel clones are grown in culture from small parent populations containing no mutated genes. After a certain number of cell generations or when the clones have grown to a specific size, the prevalence of mutant cells within each population is recorded. The mean prevalence per clone is then used to estimate the underlying mutation rate. A simple in vivo analogue of this experiment is to observe many patients (rather than clones) and measure the frequency with which neoplasia (rather than mutant colonies) arise:- the neoplasia being markers of mutation. In other words, rather than trying to observe gene mutations at the microscopic level, a practi-

cal alternative is to observe malignancies at the population level (figure 4.1). Luebeck and Moolgavkar [LM02] inferred a rate of 10^{-6} per gene per year for *APC* mutation in patients with colorectal cancer via this approach.

The primary advantage of using population data, in place of cell line data, is that they are indicative of in-vivo rather than in-vitro gene mutations. A second advantage is that neoplasms are markers of precisely the gene mutations whose rates of occurrence we would like to measure. Unfortunately, it is usually difficult to determine the number of precursor cells per patient at risk for a given cancer - a crucial factor in determining the mutation rate. Furthermore, besides the mutation of interest, there are often unknown additional pathogenic events required to produce a cancer. These confound the estimation procedure. For example, Luebeck and Moolgavkar's estimate of the *APC* mutation rate depends on an assumed number of progenitor cells per colorectum. It also depends on more difficult assumptions about the aetiology of bowel cancer. Although it was reasonable for them to suppose that the initiating events are alterations in the two alleles of *APC*, assumptions about the events that follow from the second APC hit are more speculative. These compromise the accuracy of their estimate.

In cancers where there is a well-defined genetic and histological sequence, precursor lesions can be used in place of malignant tumours for estimating rates of gene mutation. The advantage of using neoplasia at an earlier stage of tumourigenesis is that they contain fewer genetic alterations than mature tumours (in addition to the mutation of interest). Iwama treated colonic adenomas as representative of two *APC* hits [Iwa01]. From observed data on the incidence of adenomas in the bowel, he was able to infer a rate between 2×10^{-6} and 3×10^{-6} *APC* mutations per gene per year - similar to the estimate of Moolgavkar et al. While Iwama's approach removes some of the uncertainty associated with aetiology, it is still dependent on assumptions about the number of cells at risk of cancer per patient. The method presented here removes this dependency, while still requiring only a limited knowledge of pathogenesis.

The rate of *APC* mutation can be inferred by comparing incidence of sporadic colon cancer, with that of colon cancers arising in the context of familial adenomatous polyposis coli (FAP). FAP is a hereditary cancer syndrome caused by germline *APC* mutation. It occurs in the general population with a frequency between $\frac{1}{10000}$ and $\frac{1}{7000}$ and accounts for around 2% [dlC04] of CRCs. Patients with FAP develop hundreds

to thousands of adenomas throughout the colon and rectum, beginning in their teens. If untreated at least one of these adenomas will progress to an invasive lesion (the penetrance of FAP is taken to be 100%). It is commonly assumed that colon cancers arising sporadically and in FAP patients proceed along similar pathways of somatic evolution, with FAP cancers requiring one less mutation at the *APC* locus. If this is the case and if partial loss of APC function is initiating for the large majority of cancers of either type, then it is reasonable to use the age-onset pattern of FAP to estimate the time lapse between the first *APC* hit and clinical detection in sporadic bowel cancer (figure 4.2). Using FAP incidence data in this manner removes the need for assumptions about the adenoma - carcinoma sequence after the first *APC* mutation. Additionally, the method is insensitive to the assumed number of progenitor cells, since this information is implicit in the FAP data, provided that the number of progenitors is the same in FAP and sporadic patients.

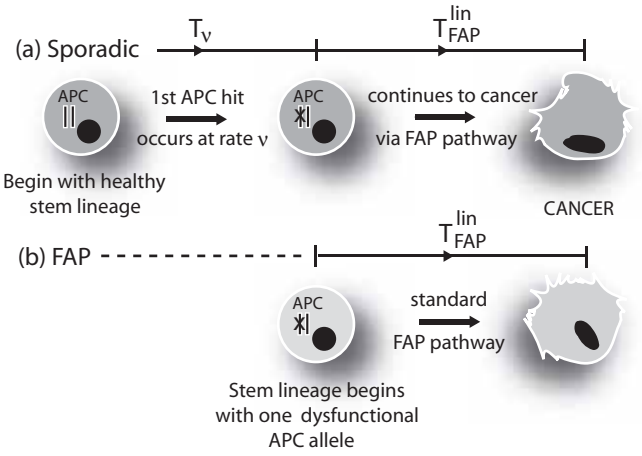

Figure 4.2: The pathways of FAP and sporadic bowel cancer are separated by only a single, truncating *APC* hit. In the case of FAP the first *APC* hit already exists in the germline. In a sporadic patient, a given cell lineage takes T_ν years to acquire a truncating APC mutation in one allele and then a further $T_{\mathrm{FAP}}^{\mathrm{lin}}$ years to become malignant. A FAP lineage only takes $T_{\mathrm{FAP}}^{\mathrm{lin}}$ years to become malignant.

4.1.2 A comparative method for estimating the *APC* mutation rate

FAP incidence was estimated from a retrospective study of British FAP families published in1965 [Vea65]. Sporadic incidence was estimated from UK registry data recorded between 1960 and 1962 by Doll [DPW66]. It was assumed that sporadic bowel cancer emerges via the same genetic pathway as FAP, but with an extra initiating *APC* 'hit' (figure 4.2). A quantitative model, describing the time taken for a sporadic bowel cancer to develop, was derived on the basis of this assumption. The initiating *APC* hit occurs at a rate of ν per allele per year in the model. The time between this hit and clinical detection was estimated from FAP incidence data. Fitting the model separately to sporadic male and female incidence data then allowed ν, the desired mutation rate, to be inferred in each case.

4.1.2.1 Time Until Sporadic Cancer

It was assumed that the colonic epithelium of each patient is sustained by a population of stem cells and that these stem cells form the target population for cancer. Periodically, each divides asymmetrically to give rise to one new stem cell and one non-stem cell daughter. A given stem cell, and its lineal stem cell descendants will be referred to collectively as a 'stem cell lineage' (figure 4.3). We assumed a fixed number of lineages, N, and each was treated as an independent entity. Before the first *APC* mutation has occurred, each lineage retains only a single stem cell (undergoes no symmetric divisions). This restriction allows the time taken for the first *APC* mutation to be modeled with an exponential distribution. After the first *APC* mutation, however, the assumptions of the model do not preclude expansion through symmetric divisions. A cancer is recorded when the first of the lineages has become malignant and grown to a detectable size.

The age at which any single lineage in a sporadic patient becomes malignant is denoted by $T^{\text{lin}}_{\text{spor}}$. It is the sum of two times (figure 4.2): (i) T_ν the time taken for the initial *APC* hit to occur and (ii) $T^{\text{lin}}_{\text{FAP}}$, the time taken for the events which comprise the FAP pathway to follow (again see figure 4.2). So $T^{\text{lin}}_{\text{spor}}$ can be expressed as:

$$T^{\text{lin}}_{\text{spor}} = T_\nu + T^{\text{lin}}_{\text{FAP}}. \tag{4.1}$$

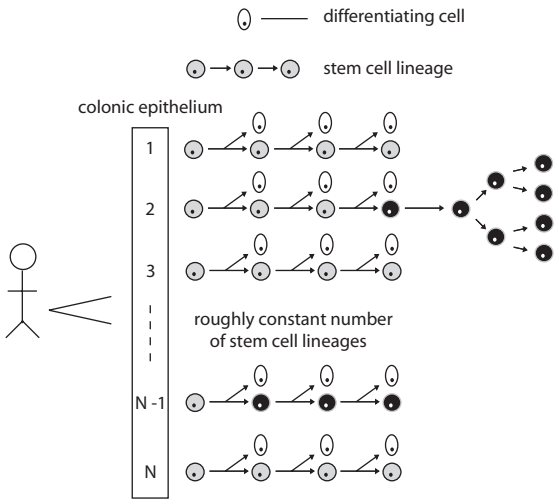

Figure 4.3: The epithelial sheet is replenished, initially, by a fixed number of independent stem cell lineages. These are shown in grey and their differentiating progeny are shown in white. Black cells represent stem cell descendants that have at least one mutant APC allele. The assumptions of the model do not preclude expansion through symmetric division, in the black cell lineages

T_ν is assumed to follow to an exponential distribution, with rate 2ν (since there are two alleles and ν is the mutation rate per allele per year). Two contrasting methods were then used to derive a distribution for $T_{\text{FAP}}^{\text{lin}}$ from the observed FAP data.

First of all, to avoid forcing a generic shape onto the distribution of $T_{\text{FAP}}^{\text{lin}}$, it was constructed as a survival curve directly from the FAP data. The construction proceeded in two stages. Initially, the observed FAP data, as summarized by Ashley [Ash69], were used to estimate the distribution of T_{FAP} , the time at which a patient (rather than a lineage) first develops a malignancy. So, $P[T_{\text{FAP}} \leq t]$ was inferred for $t_i = 10, 15, 20, 25, \ldots$ via the 'actuarial method' (see for example Parmar and Machin [PM95]).

The FAP data from [Vea65] are reproduced in table 4.1. If T_{FAP} represents the time at which a FAP patient presents with a tumour, then the data in table 4.1 suggests an estimate for $P[T_{\text{FAP}} \leq 10]$ is 0, since there were no cases of cancer occurring before age 10. The chance of a patient presenting with cancer before age 15 given that the patient was healthy at age 10 is taken as $P[T_{\text{FAP}} \leq 15 |\ T_{\text{FAP}} > 10] = \dfrac{1}{(151 - \frac{7}{2})}$. This

Age	Cases treated, S	Cases of cancer, D	Patients for Analysis
0-10	5	0	156
10-15	7	1	151
15-20	11	0	144
20-25	10	6	133
25-30	14	14	123
30-35	15	17	109
35-40	5	8	94
40-45	9	14	89
45-50	3	6	80
50-55	1	3	77
55-60	0	2	76
60-65	1	1	76
65-70	0	1	75
70-75	0	2	75

Table 4.1: FAP data (males only) from [Ash69], the patients treated during each age interval are removed from the study, and no longer form part of the analysis. This is reflected in the 'Patients for Analysis' column, which contains the total number of patients who began the study and have not yet received treatment.

is the number of cancers occurring in the age range, divided by the average number of patients at risk during the period who did not already have cancer. So, although there were 151 healthy patients at age 10, 7 were treated with prophylactic surgery and removed from the analysis before age 15. Assuming the times at which they were removed are uniformly distributed (the actuarial assumption - [PM95]), then there were on average $151 - \frac{7}{2}$ patients. Similarly, $P[T_{\text{FAP}} \leq 25 |\ T_{\text{FAP}} > 20] = \dfrac{6}{(133 - 1 - \frac{10}{2})}$. Again, this is the number of cancers, 6, divided by the average number of healthy and untreated patients $133 - 1 - \frac{10}{2}$. 133 patients had not yet been treated, subtract 1, as one case of cancer has been recorded, and then subtract $\frac{10}{2}$, as 10 patients were treated during the interval. The probability of cancer occurring before age 30, for example, is then constructed as a product:

$$
\begin{aligned}
P[T_{\text{FAP}} \leq 30] &= 1 - P[T_{\text{FAP}} > 30] \\
&= 1 - P[T_{\text{FAP}} > 10]P[T_{\text{FAP}} > 15|T_{\text{FAP}} > 10]...P[T_{\text{FAP}} > 30|T_{\text{FAP}} > 25] \\
&= 1 - (1 - P[T_{\text{FAP}} \leq 10]) \\
&\times (1 - P[T_{\text{FAP}} \leq 15|\ T_{\text{FAP}} > 10]) \\
&\times \ldots \\
&\times (1 - P[T_{\text{FAP}} \leq 30|T_{\text{FAP}} > 25]).
\end{aligned}
$$

This completes the actuarial method for calculating $P[T_{\text{FAP}} \leq t_i]$ when t_i is $10, 15, 20, 25, ...$ Subsequently, it was observed that if a patient has N independent lineages at risk of becoming cancerous, then there is a simple relationship between T_{FAP} and $T_{\text{FAP}}^{\text{lin}}$. By equation (2.2):

$$P[T_{\text{FAP}} \leq t_i] = 1 - (1 - P[T_{\text{FAP}}^{\text{lin}} \leq t_i])^N. \tag{4.2}$$

Rearranging gives

$$P[T_{\text{FAP}}^{\text{lin}} \leq t_i] = 1 - (1 - P[T_{\text{FAP}} \leq t_i])^{1/N}.$$

Hence a smooth approximation to $P[T_{\text{FAP}}^{\text{lin}} \leq t]$ can be derived by interpolating the co-ordinates:

$$(t_i, 1 - (1 - P[T_{\text{FAP}} \leq t_i])^{1/N}), \quad t_i = 10, 15, 20, \ldots \qquad .$$

Using this approximation to $P[T_{\text{FAP}}^{\text{lin}} \leq t]$ and equation (4.1) it follows that:

$$
\begin{aligned}
P[T_{\text{spor}}^{\text{lin}} \leq t] &= P[T_\nu + T_{\text{FAP}}^{\text{lin}} \leq t] \\
&= \int_0^t \frac{d}{ds} \left(P[T_\nu \leq s]\right)\ P[T_{FAP}^{lin} \leq t - s]\, ds.
\end{aligned}
$$

This gives the age at which cancer arises in a single lineage, $T_{\text{spor}}^{\text{lin}}$. For the patient, any of the N lineages has the potential to become a cancer. Consequently, the time

taken for a patient to present with cancer, T_{spor}, is much quicker than that for a lineage. The two waiting times are related by (equation (2.2)):

$$P[T_{\text{spor}} \leq t] = 1 - (1 - P[T_{\text{spor}}^{\text{lin}} \leq t])^N. \tag{4.3}$$

4.1.2.2 Fit to sporadic data 1

With N fixed, and $T_{\text{FAP}}^{\text{lin}}$ defined as above, $P[T_{\text{spor}} \leq t]$ (equation (4.3)) can be used to construct a likelihood function, $L(\underline{D}_{\text{spor}}|\nu)$, for the sporadic data, $\underline{D}_{\text{spor}}$. This is done as described in the previous chapter (equation (3.8)). In turn, $L[\underline{D}_{sport}|\nu]$ is used to calculate a posterior distribution $P[\nu|\underline{D}_{\text{spor}}]$. Assuming a uniform prior for ν on $10^{-7} - 10^{-3}$, gives

$$P[\nu|\underline{D}_{\text{spor}}] = \frac{L[\underline{D}_{\text{spor}}|\nu]}{\int_{\nu_{\text{min}}}^{\nu_{\text{max}}} L[\underline{D}_{\text{spor}}|v]\, dv}.$$

The variance of this posterior is likely to be too narrow because uncertainty over the distribution of $T_{\text{FAP}}^{\text{lin}}$ was ignored in its derivation. To address this issue, in the second method for characterizing $T_{\text{FAP}}^{\text{lin}}$, the assumption was made that its distribution follows a predefined functional form whose parameters are to be inferred. Specifically, Armitage and Doll's formula was used (equation (2.1)). Recall that this gives the probability that a cancer requiring n successive mutations, each occurring at a rate μ, has developed from an immortal lineage by time t:

$$P[T_{\text{FAP}}^{\text{lin}} \leq t] = 1 - \sum_{i=0}^{n-1} \frac{\mu^i t^i}{i!} e^{-ut}, \tag{4.4}$$

so that

$$P[T_{\text{FAP}} \leq t] = 1 - (1 - P[T_{\text{FAP}}^{\text{lin}} \leq t])^N. \tag{4.5}$$

4.1.2.3 Fit to sporadic data 2

Using equations (4.4) and (4.5), a likelihood function, $L[(\underline{D}_{\text{FAP}}, \underline{D}_{\text{spor}})|\mu, n, \nu]$ was constructed for the FAP and sporadic data together, given the approximated FAP curve

(parametrized by μ and n) and the model for the sporadic disease. This likelihood was the product of the likelihood for the sporadic data ($L[\underline{D}_{\text{spor}}|\nu]$ - equation (3.8)) and a new likelihood, $L[\underline{D}_{\text{FAP}}|\mu, n]$ for the FAP data. The FAP data is drawn from a small study involving M patients where M is ~ 100. The patients are observed at 5-year age intervals. During the ith age interval, a total of S_i patients are treated with surgery before being removed from the study and D_i new cancers occur among the remaining patients. To account for loss of patients through treatment, the effective number of patients, Pop_i, at risk during the ith interval, Pop_i, is defined by:

$$Pop_i = M - \sum_{k=0}^{i-1}(D_k + S_k) - \frac{1}{2}S_i. \tag{4.6}$$

The likelihood $L[\underline{D}_{\text{FAP}}|\mu, n]$ at μ given the FAP data can be expressed as a product of binomial probabilities. If the probability of a cancer arising in a given patient during the ith interval is p_i, then the probability that D_i cancers arise among the effective population Pop_i over the period is

$$\binom{Pop_i}{D_i} p_i^{D_i}(1-p_i)^{Pop_i - D_i}.$$

$L[\underline{D}_{\text{FAP}}|\mu, n]$ is then equal to the probability of the given sequence of case numbers $\{D_1, D_2, \ldots\}$,

$$L[\underline{D}_{\text{FAP}}|\mu, n] = \prod_i \binom{Pop_i}{D_i} p_i^{D_i}(1-p_i)^{Pop_i - D_i}.$$

If s_i and e_i are the start and end points of the ith age interval then p_i can be evaluated, using equation (4.5), as:

$$p_i = P[T_{FAP} \leq e_i | T_{FAP} > s_i] = \frac{P[T_{FAP} \leq e_i] - P[T_{FAP} \leq s_i]}{1 - P[T_{FAP} \leq s_i]}. \tag{4.7}$$

To fit both the sporadic and FAP data simultaneously, the combined likelihood function, $L[(\underline{D}_{\text{FAP}}, \underline{D}_{\text{spor}})|\mu, n, \nu]$ was used:

$$L[(\underline{D}_{\text{FAP}}, \underline{D}_{\text{spor}})|\mu, n, \nu] = L[\underline{D}_{\text{FAP}}|\mu, n] \cdot L[\underline{D}_{\text{spor}}|\nu].$$

The posterior density at ν, was calculated assuming uniform priors for ν,μ and n respectively on $10^{-7} - 10^{-3}$, $0 - 10^{-2}$ and $2 - 7$:

$$P[\nu|(\underline{D}_{\text{FAP}}, \underline{D}_{\text{spor}})] = \frac{\sum_{n=n_{\text{min}}}^{n_{\text{max}}} \int_{\mu_{\text{min}}}^{\mu_{\text{max}}} L[(\underline{D}_{\text{FAP}}, \underline{D}_{\text{spor}})|\mu, n, \nu] d\mu}{\sum_{n=n_{\text{min}}}^{n_{\text{max}}} \int_{\nu_{\text{min}}}^{\nu_{\text{max}}} \int_{\mu_{\text{min}}}^{\mu_{\text{max}}} L[(\underline{D}_{\text{FAP}}, \underline{D}_{\text{spor}})|\mu, n, v] d\mu \, dv}.$$

4.1.3 Results

It was assumed that sporadic bowel cancer emerges via the same genetic pathway as FAP, but with an extra initiating *APC* 'hit' (figure 4.2). A stochastic model, describing the time taken for a sporadic bowel cancer to develop, was derived on the basis of this assumption. The intestinal cell lineages of a sporadic patient acquire initiating APC hits at a rate of ν per allele per year in the model. The remaining time taken for any such lineage to become malignant and present clinically was estimated from FAP incidence data. Fitting the resulting model back to the sporadic incidence curve allowed a posterior distribution for ν, the desired mutation rate, to be inferred.

The posterior distribution, $P[\nu|\underline{D}]$, was derived via two separate methods. In the first, the time taken for an APC+/- cell lineage to become malignant and present clinically was assumed to be distributed according to a survival curve constructed from the FAP data. The shape of $P[\nu|\underline{D}]$ in this case (see figure 4.4), implies that ν - the annual rate of APC mutation per allele - falls between 6×10^{-5} and 9×10^{-5} for both males and females. These estimates are insensitive to two-order of magnitude changes in the assumed number of cell lineages per patient. However, substituting the optimum ν for males or females into the model gives only a rough approximation to the sporadic incidence profiles (figure 4.5). The discrepancy is systematic, in that for both males and females we overestimate the risk prior to 65 years of age and underestimate it there-

after. It is likely that ascertainment bias in the FAP data is the underlying cause of this error.

Ascertainment bias in the FAP data arises because only those patients who present clinically before death from other causes are included in the study. Consequently, relative risk of cancer at young ages, when mortality is low, is exaggerated while risk in older age as mortality increases is downplayed. Another problem is that the FAP data consist of only a small number of observations.

By assuming the 'real' distribution of FAP cases followed exactly a survival curve constructed from a small sample, the variances of the posterior distributions on ν were underestimated. To quantify the extent of this problem, in addition to using the exact survival curve suggested by observed data points, a parameterized class of possible approximations to the FAP curves was used. The posterior densities on ν in this case are only slightly wider (figure 4.6). The optimized models naturally give a better fit to the sporadic data (figure 4.7) and seem, crudely, to correct for the ascertainment bias described.

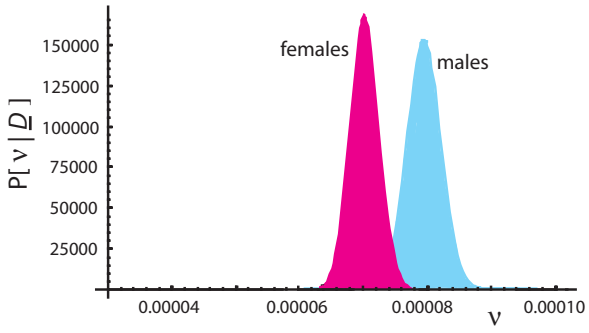

Figure 4.4: Posterior distribution of the *APC* mutation rate, ν, measured in mutations per allele per year, calculated using a single parameter likelihood function. Sporadic data colon cancer incidence taken from Doll [DPW66], cases diagnosed between 1960 and 1962.

4.1.4 Discussion

The above calculations suggest that alleles of the *APC* tumour suppressor mutate about 6×10^{-5} times a year. This is 30 times faster than previous estimates [Iwa01, LM02]. A

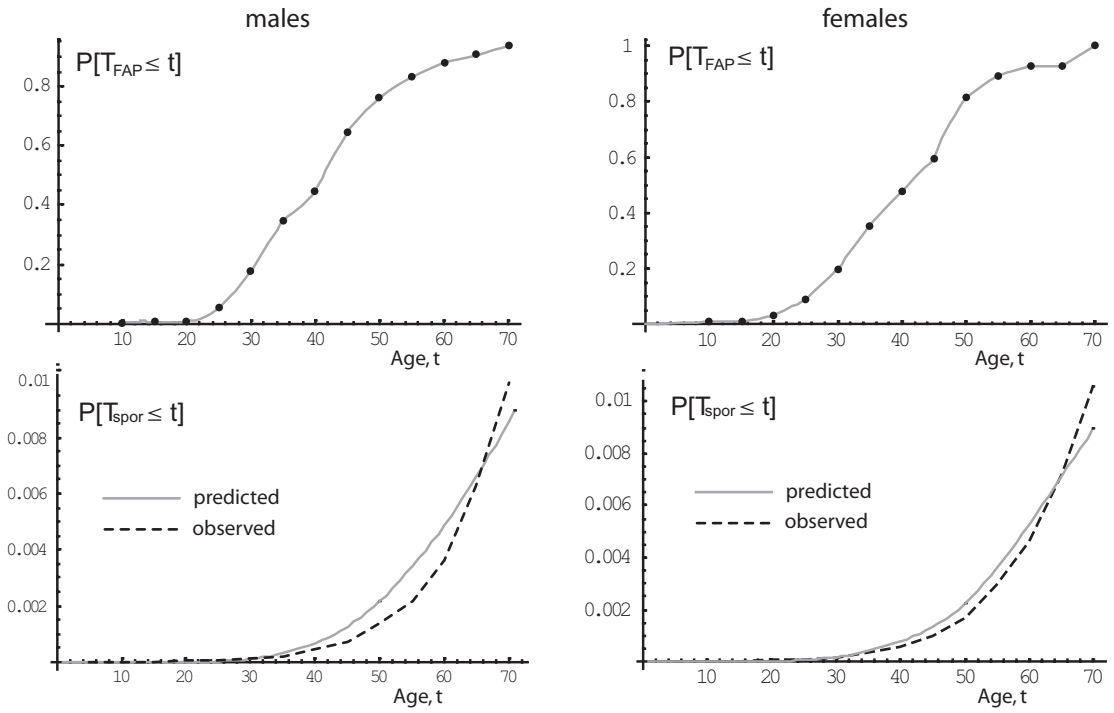

Figure 4.5: $P[T_{\text{FAP}} \leq t]$, the cumulative risk of FAP at age t, was calculated separately for males (left column) and females (right column) by interpolating observed FAP data. $P[T_{\text{spor}} \leq t]$, the cumulative risk of sporadic bowel cancer at age t, was constructed from $P[T_{\text{FAP}} \leq t]$, according to assumptions about the relationship between FAP and sporadic bowel cancer described in the text. Using the optimum ν, provides an adequate approximation to observed sporadic data.

rough consistency check can be done by comparing the estimate against the mutational spectrum of the gene. More than 95% of the APC mutations found in bowel cancers are nonsense or frameshift mutations. They take the form of small deletions / insertions or point mutations that result in truncation of the protein [FWB02]. Assuming an error rate (insertions / deletions or mismatches) of 10^{-10} per base per cell generation [KB00] and assuming order 10^2 stem cell divisions per annum, the number, B, of base pairs through which *APC* can be suppressed (either by insertion / deletion or point mutation) is:

$$B = \nu \times 10^8$$

Assuming $\nu \simeq 10^{-5}$ gives $B \simeq 1000$. This seems a sensible value for B as more

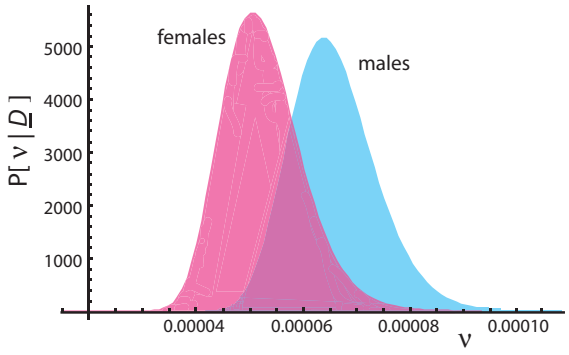

Figure 4.6: Posterior distribution of the *APC* mutation rate ν, measured in mutations per allele per year, calculated using a three parameter likelihood function. Data on sporadic colon cancer incidence taken from Doll [DPW66], cases diagnosed between 1960 and 1962.

than 700 distinct, somatic APC mutations, have been reported in bowel cancers to date [LPBS98].

Confidence in the estimates for v, should be based primarily upon the accuracy of the assumptions / approximations used in their calculation. The most crucial assumptions are listed below:

1. FAP and sporadic bowel cancer are both initiated by genetic alterations in the *APC* gene,

2. A lineage in a sporadic patient, with one *APC* hit, becomes malignant via the same mechanism as a FAP lineage with the germline genotype,

3. The target lineages for bowel cancer act independently,

4. The time of loss of the first *APC* allele in a sporadic patient follows an exponential distribution.

To evaluate the accuracy of these assumptions, a definition of 'sporadic' colon cancer is necessary. We take sporadic to mean any case occurring in the general population, excluding those arising in the context of a known, highly-penetrant germline mutation. Although initiating *APC* mutations appear to feature in the large majority of such sporadic cases [RLI$^+$00], alternative genetic pathways are available for progenitor cells in the colonic epithelium [ST06] and cases with no *APC* mutation are a source of

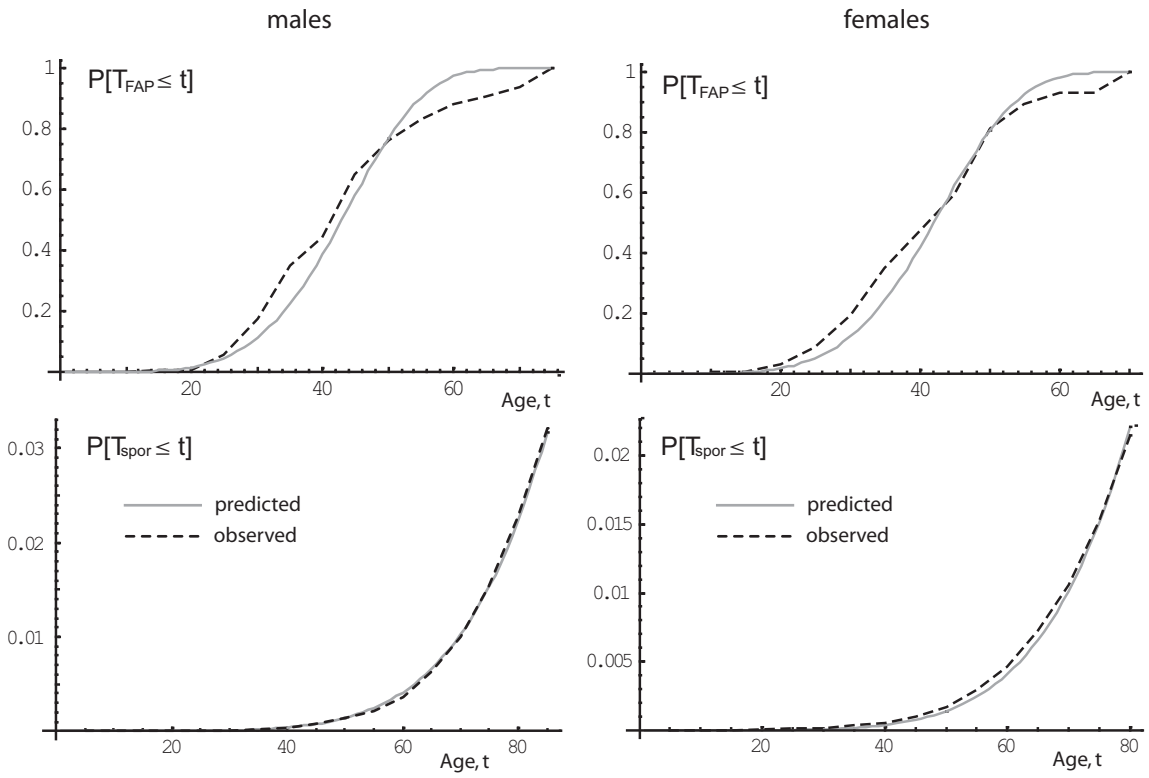

Figure 4.7: $P[T_{\text{FAP}} \leq t]$, the cumulative risk of FAP at age t, is represented by a smooth function, parametrized by μ and n. $P[T_{\text{spor}} \leq t]$, the cumulative risk of sporadic bowel cancer at age t, was again constructed separately for males and females from $P[T_{\text{FAP}} \leq t]$, as described in the text. Using the optimum parameter vector $(\hat{\mu}, \hat{n}, \hat{\nu})$, which maximizes the likelihood function, a good fit to the sporadic data can be made.

noise in the experiment. Since these cases account for less than 30% and since their age-relatedness is unlikely to be remarkably different from those initiated by *APC*, the impact should be minimal.

A similar but potentially more serious source of noise arises, because the registry data we use to represent sporadic cases are likely to be contaminated with FAP cases and those of other highly penetrant hereditary bowel cancers. However, these account for less than 5% of all colorectal cancers [dlC04], so the effect is negligible.

Concerning approximation 2, *APC* hits can broadly be categorized into two types: truncating mutations and loss of heterozygosity (LOH). It is well established that the type, as well as the position of the two hits at *APC* are not independent in either FAP or sporadic bowel cancer. The first hit in sporadic cases and the germline mutations

in FAP are both predominantly truncating mutations but their spatial distribution is not identical. Consequently, due to non independence of the two hits, or otherwise, LOH may occur at a different frequency as a second hit in sporadic cancers. Fortunately, the difference in the spatial distributions is primarily due to certain germline mutations occurring at high frequencies relative to somatic mutations in sporadic cases. Specifically, germline mutations affecting codon 1061 and codon 1309 are relatively frequent in FAP. As these tend to lead to different second hits (truncating mutation and LOH respectively) the overall frequency of LOH as a second hit in FAP and sporadic colon cancer is similar. For example, Rowan et al. [RHG$^+$05] and Prall et al. [PWO07] detected LOH in 23 of 99 (23%) and 20 of 99 (20%) sporadic cancers respectively, while Lamlum et al. [LIR$^+$99] found APC LOH in 42 of 210 (20%) FAP tumours.

Other factors that usually confound estimates taken from incidence data include the assumed number of cell lineages at risk and also calendar year effects whereby incidence changes over calendar time. Happily, the comparative nature of the method presented here eliminates any sensitivity to the assumed number of lineages. It remains to be shown that the estimate is also robust under age-related changes in cell number. Calendar year effects are mild for bowel cancer before the 1980s [LM02] and are unlikely to have a significant effect. Finally, a difficult problem, when modelling incidence data, is to decide how long it takes for a cancer to be detected once it has come into existence [MIN06]. In this study, the FAP data are used to estimate the time taken for an APC+/- cell lineage to mutate into a cancer and present clinically. Assumptions about how this time is split between tumour formation and tumour progression are not required.

Given a small set of fairly conservative assumptions regarding sporadic colon cancer and its relationship to FAP, incidence data on the two diseases imply that the rate of truncating and disease-causing *APC* mutation is of order 10^{-5} mutations per allele per year. This estimate, although higher than previous estimates, seems to be consistent with the mutational spectrum of *APC*. Further, the estimate neither requires accurate determination of the number of target cells in the colon, nor depends on assumptions about time spans between the appearance of a cancer and its detection clinically. The quality of fit provided by the model supports the theory that FAP and sporadic bowel cancer follow the same genetic pathway and are separated by only one mutation.

4.2 HNPCC and sporadic colorectal cancer

The penetrance of *APC* germline mutations can be explained by a simple modification of the sporadic genetic pathway for bowel cancer. In general, however, many factors may influence the penetrance of a given germline variant. It may target a non-initiating and non-rate limiting step from a sporadic cancer and hence have low or negligible penetrance. Alternatively, a germline mutation may cause selective pressures that give rise to a syndromic cancer with an aetiology distinct from that of sporadic cancer in the same tissue. For example, hereditary non-polyposis colorectal cancer (HNPCC) patients have germline mutations targeting certain DNA repair pathways. They are prone to particular gene mutations which they accumulate more quickly than sporadic bowel cancer patients. In the second part of this chapter quantitative methods are used to probe for consequent differences in the aetiology of HNPCC and sporadic bowel cancer. First of all, it is argued that only a subset of HNPCC patients actually has a raised risk of cancer. This explains the plateauing of HNPCC penetrance with age in the population. Subsequently, a scale free measure of the rate of change in incidence with age, referred to as 'log-log acceleration' or 'LLA' is used to argue against a simple relationship between HNPCC and sporadic bowel cancer. The change in 'LLA' produced by a germline mutation under various hypothetical scenarios is compared with that observed for the DNA miss-match repair (MMR) mutations found in HNPCC patients. A model in which these MMR defects act only to increase the rate of the transitions found in the sporadic cancer is inconsistent with the incidence shift observed in HNPCC. A more consistent hypothesis is that HNPCC tumorigenesis begins with slower transitions than sporadic bowel cancer but then finishes with a series of faster transitions as a result of reduced DNA repair capacity,

4.2.1 HNPCC

Hereditary non-polyposis colorectal cancer (HNPCC) (also known as Lynch syndrome) is a familial cancer syndrome affecting patients with a germline mutation in an allele of one of the mismatch repair (MMR) genes. Somatic loss of the wild-type copy causes a failure in post replicative MMR. A portion of the cancer susceptibility in HNPCC, therefore, is thought to originate from an elevated rate of copying errors during DNA replication. 90% of all known HNPCC - associated germline mutations are found in

just two of the nine human genes shown to possess MMR function [PV04]. *MLH1* and *MSH2* mutations account for approximately 50% and 40% respectively. The fact that germline mutations in HNPCC target specific components of the MMR machinery suggests the possibility that an additional fitness advantage, distinct from dysfunctional MMR, is also being selected. Indeed, the ability to signal an apoptotic response following certain types of DNA damage has been proposed for both MLH1 and MSH2 [Fis01]. Regardless of the precise mechanisms through which variants of *MLH1* and *MSH2* confer risk of colorectal cancer, it is clear from the high penetrance of HNPCC and the finding of biallelic inactivation of *MLH1* in a significant majority of sporadic bowel cancers, that silencing of these genes represents a pivotal rate-limiting step in the somatic evolution of the disease. Therefore, it will be of interest to quantify age-specific risk in the context of a germline MMR mutation. A comparison with age-specific risk of sporadic bowel cancer may then provide clues as to the biological consequences of inherited MMR dysfunction.

The incidence curve for HNPCC is of general interest, not least for the purposes of genetic counselling, and so, much effort has been expended in trying to estimate the probability of developing cancers associated with the syndrome. This is, however, more difficult than in the case of FAP. HNPCC is a more heterogeneous disease with lower penetrance, and so not as easily diagnosed or as cleanly defined. It is hard to obtain a representative sample of patients within a given population that fit a clear definition of HNPCC.

4.2.2 Defining HNPCC

When the term 'HNPCC' was introduced by Henry Lynch in 1985, it was intended to describe early-onset and predominantly right-sided CRCs arising in an autosomal dominant pattern, sometimes in combination with certain extra colonic cancers and always in the absence of the multiple premonitory polyps associated with FAP. Ambiguity crept in, however, after the genetic basis of the disease was discovered in the early 1990's and clinical definitions used to diagnose HNPCC families were incrementally updated to improve their sensitivity. In what follows, the terms HNPCC and Lynch Syndrome will be used synonymously and specifically in reference to germline MMR mutation carriers. For example, an individual with a de novo germline MMR mutation shall be

considered an HNPCC patient despite an inconspicuous family history.

4.2.3 Calculating age-specific risk of colorectal cancer in HNPCC patients

An idealized method of measuring age-specific risk in Lynch syndrome would involve screening the whole population of interest for germline MMR defects and then recording every primary cancer occurring among mutation carriers over a fixed period, for example 12 months. Such a method is clearly impractical. Instead, a sample of HNPCC families must typically be used. Collecting a representative sample is very difficult. As pointed out by Mitchell et al. [MFDC02] "the identification of families with mismatch repair gene mutations using any phenotypic selection criteria introduces ascertainment bias, and such kindreds may not be representative of all mutation-carrying families in the general population".

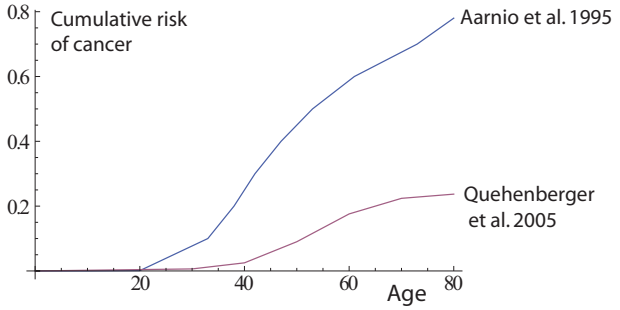

Figure 4.8: Penetrance estimates for bowel cancer in HNPCC have been lowered in light of concerns over ascertainment bias. The original studies, for example Aarnio et al. [AMA$^+$95], that used family history to identify HNPCC kindreds, were enriched for multiple case families and so are thought to have overestimated risk. More recently, Bayesian statistical methods have been used by Quehenberger et al. [QVvH05] to correct for ascertainment bias by conditioning on sample phenotype. This has resulted in a markedly reduced penetrance estimate (figure redrawn from Aarnio et al. and Quehenberger et al. [AMA$^+$95, QVvH05].

The first penetrance calculations for *MLH1* and *MSH2* mutation typically yielded lifetime risk figures of 70% - 90% (figure 4.8), but used family history as part of the selection criteria for index cases [AMA$^+$95, ASP$^+$99, VWM$^+$96]. Such estimates are likely to be inflated, since multiple case families are over-represented in samples

collected on the basis of family history. To reduce the "ascertainment bias" associated with sampling according to family history, Dunlop et al. [DFC+97] instead used early-onset as a sampling criterion, gathering a small collection of 67 relatives of early-onset cases. The result was a lower penetrance estimate, with cumulative risks to age 70 for CRC of 70% in males and only 34% in females (a lower risk in females is consistent with other studies that have separated the sexes in Finland, the United States, Holland and Australia [MFDC02, JBD+06]). The possibility remains that early-onset probands may either tend to carry mutations that correlate with a severe phenotype or tend to occur in families whose members share other genetic or environmental risks.

As an alternative to synthesizing a small sample of population based probands, statistical methods can be used to correct for ascertainment bias in larger clinic-based studies that may still utilize databases of HNPCC kindreds. The obvious method of correcting for ascertainment, is to estimate penetrance with a likelihood function that models the ascertainment process. Unfortunately, this is frequently intractably difficult for HNPCC, since the clinical criteria are complex. To circumvent this problem, at the expense of likelihood efficiency, "retrospective likelihood" methods [KT00, CBP04] can be used. The idea is to condition on the observed phenotypic information and maximize the probability of observing the genotypic information. Such a method, was taken up by Quehenberger et al. [QVvH05] and used to estimate a cumulative risk to age 70 for colorectal cancer in men at 26.7% and in women at 22.4% (figure 4.8) using over 2000 patients from a national HNPCC database in the Netherlands. Another study, by Jenkins et al. [JBD+06] using early-onset cases unselected for family history as probands, calculated comparable, but higher risks of CRC by age 70, 45% for men and 38% for women.

Figure 4.9 contains a collation of estimated penetrance functions from the three studies mentioned above that each use a methodology designed to minimize ascertainment bias. Although not shown, to avoid a cluttered picture, wide 95% confidence intervals were provided for the penetrance estimates of Quehenberger et al. and Jenkins et al. In the case of Quehenberger et al. risk to age 80 of CRC was quoted at 28.5% (13.7 - 53.4%) in males and at 23.7% (11.3 - 45.5%) in females. In the case of Jenkins et al. risk to age 70 in males was 55.9% (36.7 - 75.0%) and in females 48.1% (26.2-65.3%). Hence it is difficult to say to what extent the disparity in these risk estimates could be

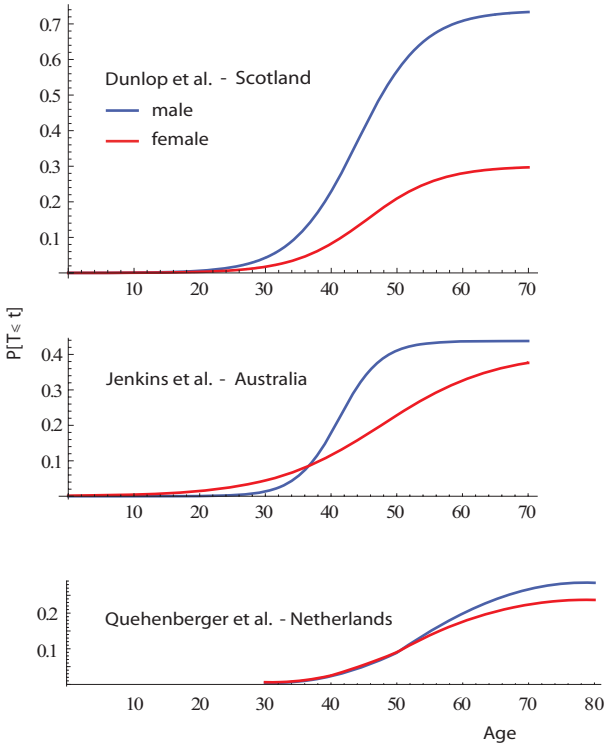

Figure 4.9: Penetrance functions estimated in three studies that have attempted to mitigate, as far as possible, ascertainment bias. An unexpected but common feature, is the shallow gradient after age 50.

caused by experimental bias and low statistical power. Their most striking common feature is the plateauing of penetrance after age 50. The majority of theoretical penetrance functions do not plateau until they reach full penetrance, i.e. a cumulative risk of 100%. When the multistage evolution of cancer is modelled as a simple series of transitions through states of a Markov chain, cumulative risk (the risk of transitioning into the final malignant state), will increase with age until transition to the malignant state becomes a practical certainty. Hence the incidence (rate of transition into the final malignant state) is a monotonic increasing function of age. The plateauing penetrance functions of figure 4.9 translate into peaked rather than monotonic incidence patterns (figure 4.10).

Figure 4.10 shows the incidence profiles for CRC in MMR mutation carriers as predicted by Dunlop et al., Quehenberger et al. and Jenkins et al. They appear as a series of humps centred around the 40's and 50's.

What could cause risk per unit time to be less in older individuals, who under

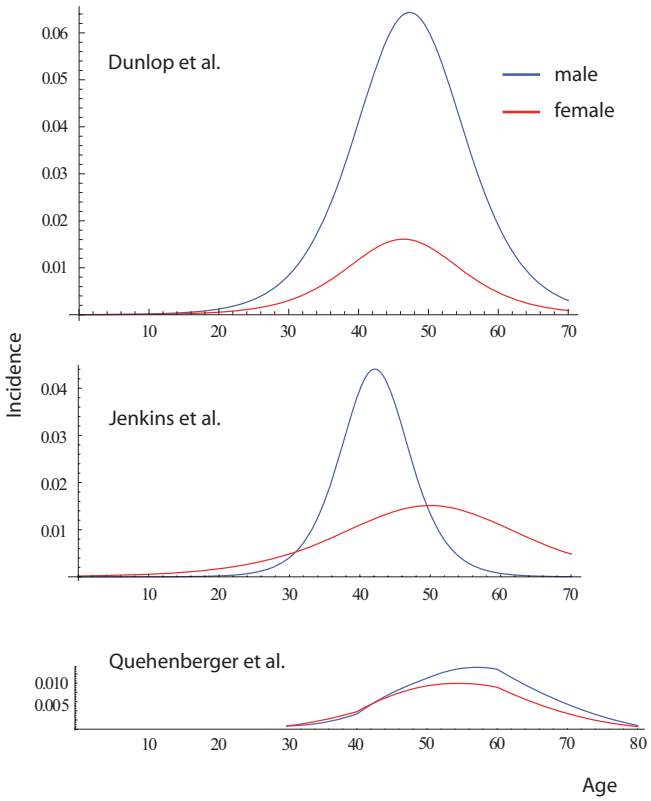

Figure 4.10: Incidence of CRC in MMR mutation carriers derived from the penetrance functions given in figure 4.9. Declining incidence, after age 50, could be indicative of a combination of factors. For example a severe decline in susceptible target cells, heterogeneity in liability, age-related cell behaviour and study design issues, e.g. ascertainment bias.

normal circumstances would be expected to carry more mutated cell lineages and pre-cancerous lesions? Some suggestions for declining incidence have already been put forward. As per a previous discussion on cancer in old age (section 2.5.2), there is the idea that susceptible individuals may have, in the main, developed cancer by a certain age. Beyond that age, the population will predominantly be made up of low susceptibility individuals and so the number of cases per 100,000 population will drop. There is also the idea of age dependent cell kinetics. For example, the number of target cell divisions per unit time in a tissue may be unusually high during a particular period of growth (c.f. peak in incidence of oesteosarcoma and its coincidence with adolescence when the long bones are growing rapidly [Pri58]). Alternatively, the rate of target cell

divisions may be unusually low at some point in old age, for example during involution of the breast [MDS80]. Another possibility is that some essential event in the natural history of a particular cancer must occur during a certain window of opportunity, as in a model of childhood leukaemia which hypothesizes a necessarily-in-utero mutation [SCS97]. Before considering such hypotheses, artefacts of the estimation methods should be ruled out. It is possible that the logistic parameterizations used to infer the three penetrance functions described, naturally tend to plateau even when the real penetrance does not. To test this, sample data can be simulated using a log-log linear hazard function (i.e. a hazard that is monotonically increasing rather than humped). These simulated sample data can then be fit using the logistic-type penetrance functions employed in the studies under scrutiny to see how they perform.

4.2.3.1 Performance of logistic penetrance estimators under log-log linear simulated patient data.

Perhaps the simplest method of simulating the life histories of a sample of patients is to independently draw a time at death, d_i, from a mortality distribution for each patient, and then a time at cancer, c_i, from a log-log linear hazard (see equation (2.3)). The likelihood of the data can be computed as:

$$L[\underline{C}, \underline{D}|\underline{X}] = \prod_i \dot{P}_{\underline{X}}[T \leq c_i]^{(c_i \leq d_i)} (1 - P_{\underline{X}}[T \leq d_i])^{(c_i > d_i)}, \tag{4.8}$$

where $\underline{C}$ holds the times at cancer, and $\underline{D}$ holds the times at death. $\underline{X}$ is the vector of parameters controlling the logistic penetrance function used to fit the data, $P_{\underline{X}}[T \leq t]$. $(c_i \leq d_i)$ is an indicator function evaluating to one when $c_i \leq d_i$ and zero otherwise. $(c_i > d_i)$ also represents an indicator function defined similarly.

Numerically maximizing the log of the likelihood given in equation (4.8) yields a maximum likelihood estimate, $\hat{\underline{X}}$. The maximization was performed in 'Mathematica' using the 'NMaximize' routine By repeating 5000 times, the mean optimum parameter vector can be used to plot a typical penetrance estimate against the true penetrance. Figure 4.11 shows two such plots. One with

$$P_{\underline{X}}(T \leq t) = \frac{x_1}{1 + e^{x_2(t-x_3)}}, \qquad \underline{X} = (x_1, x_2, x_3) \tag{4.9}$$

as in Jenkins et al. and the other with

$$P_{\underline{X}}[T \leq t] = 1 - (1 - P_s(t)) \left(1 - \frac{x_1}{1 + e^{x_2(t-x_3)}}\right), \tag{4.10}$$

as in Dunlop et al., where $P_s(t)$ is sporadic penetrance fit from sporadic incidence data. In both cases the mortality distributions were constructed from life-tables relating to the appropriate populations.

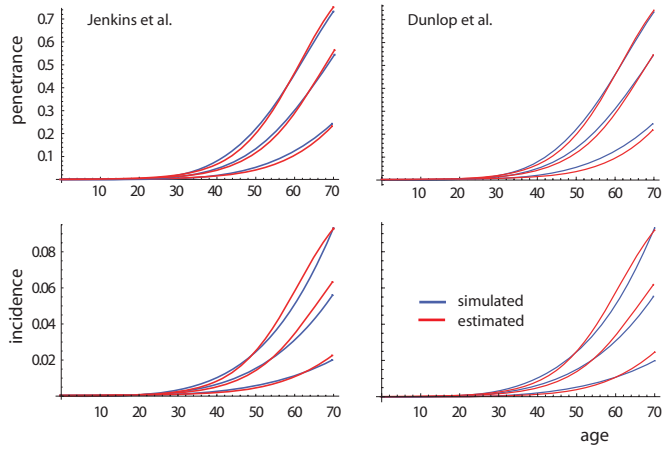

Figure 4.11: Penetrance / incidence estimates from log-log linear simulated patient data, using the inference methods of Jenkins et al. and Dunlop et al. Three different functions were used to simulate the patient data with cumulative risks to age 70 of 0.73, 0.54 and 0.24. In each case the number of patients in the simulated samples were matched to the real sample sizes used by Jenkins et al. and Dunlop et al.

The estimation methods used by Jenkins et al. and Dunlop et al. show little bias (figure 4.11). It is unlikely that an artefact of these methods could be responsible for the humped incidence patterns shown in figure 4.10. In Quehenberger et al. the log-ratio of the sporadic and syndrome-related hazards is taken to be a polynomial function of age. This model has many more degrees of freedom than the logistic functions tested above and so should be able to match a log-log linear hazard at least as well.

4.2.4 Heterogeneity and acceleration matching

Supposing then, the incidence of CRC in HNPCC is humped in reality. What explanation can multistage theory provide? Several theories for peaked incidence patterns already exist in the literature as noted above. Of particular interest is a study which encountered a similar pattern for breast cancer risk among Askenazi Jewish women carrying germline *BRCA1* and *BRCA2* mutations [SHW+97]. Because of the high prevalence of these mutations in the Askenazi Jewish population (2%), Struewing et al. were able to survey 3742 female volunteers and yield 89 population based (2.4%) mutation carriers. The penetrance estimate from the study is shown in figure 4.12 along with the associated humped incidence.

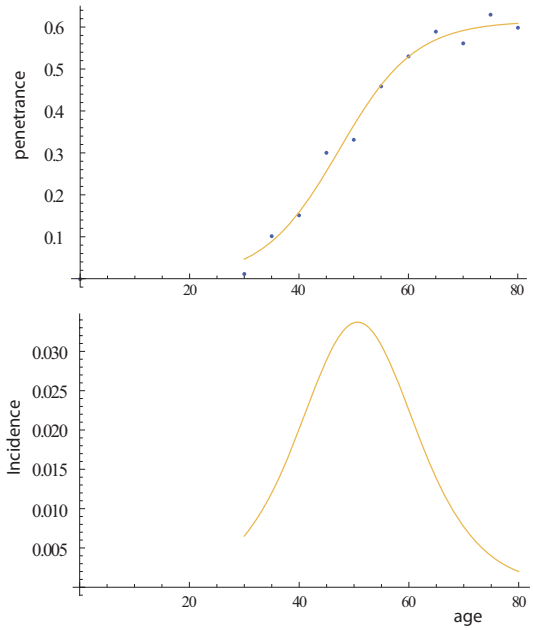

Figure 4.12: Dots - penetrance of breast cancer in *BRCA1* and *BRCA2* female mutation carriers as calculated by Struewing et al. [SHW+97]. Line - fit of Struewing et al.s estimates. The incidence shown was calculated from the logistic function (equation (4.9)).

Observing the data from Struewing et al. (figure 4.12) led Frank to suggest that perhaps only a fraction of the mutation carriers participating in the survey had fully elevated risk [Fra07]. This is the same as the heterogeneity in liability hypothesis put forward to explain declining incidence in the elderly [HJTMF+00, SMT+06]. It allows incidence in the high risk subjects to be restored to a monotonically increasing function, in line with a theoretical Markov - multistage model. Frank had an additional but related

motivation for his hypothesis, concerned with acceleration matching between sporadic patients and those with a predisposing germline mutation.

The concept of age-specific acceleration was mentioned briefly in section 2.5.4, and will now be used further. Recall that acceleration is a measure of the change in incidence with age, $\frac{d(I(t))}{dt}$ where $I(t)$ is the incidence at age t. So, if the incidence is increasing with age, then the acceleration is positive. It is useful to calculate acceleration on a log-log scale to produce a scale free measure of fractional change in incidence with fractional change in age [Fra04a]. Such a measure will be referred to as log-log acceleration or 'LLA'.

In section 2.5.4, it was mentioned that LLA can be thought of as the gradient of a log-log plot of incidence against age, in other words as the gradient of the parametric curve:

$$(\ \log(t),\ \log(I(t))\ ).$$

Therefore at age t

$$\begin{aligned} \mathrm{LLA}(t) &= \frac{\mathrm{d}\log(I(t))}{\mathrm{d}\,t} \times \left(\frac{\mathrm{d}\log(t)}{\mathrm{d}\,t}\right)^{-1} \\ &= t\frac{\dot{I}(t)}{I(t)}. \end{aligned} \tag{4.11}$$

As observed by Frank [Fra07], under the simple model of progression described by equation (2.5) with $>> 1$ lineages per patient (i.e. $N >> 1$), the number of steps by which a mutant germline genotype advances progression is approximated by the difference between the LLA for cancer arising in healthy patients and in mutation carrying patients. This is because equation (2.5), roughly equates the LLA (gradient of log-log age-incidence plot) with the number of stages in progression. Therefore, by plotting $\Delta\mathrm{LLA}(t)$, the difference in LLA between the sporadic and inherited cancer, a roughly constant function equal to the number of rate-limiting steps abrogated by the inherited mutation is expected.

As well as restoring a monotonic increasing incidence in high risk mutation carriers, Frank's hypothesis that not all BRCA1/2 germline mutation carriers are at relatively high risk also creates a ΔLLA between sporadic patients and high risk mutation carriers which seems to remain approximately constant at one, with age (figure 4.13).

However, to achieve the constant ΔLLA shown in figure 4.13, Frank has varied the smoothing parameter used in his fit of observed incidence.

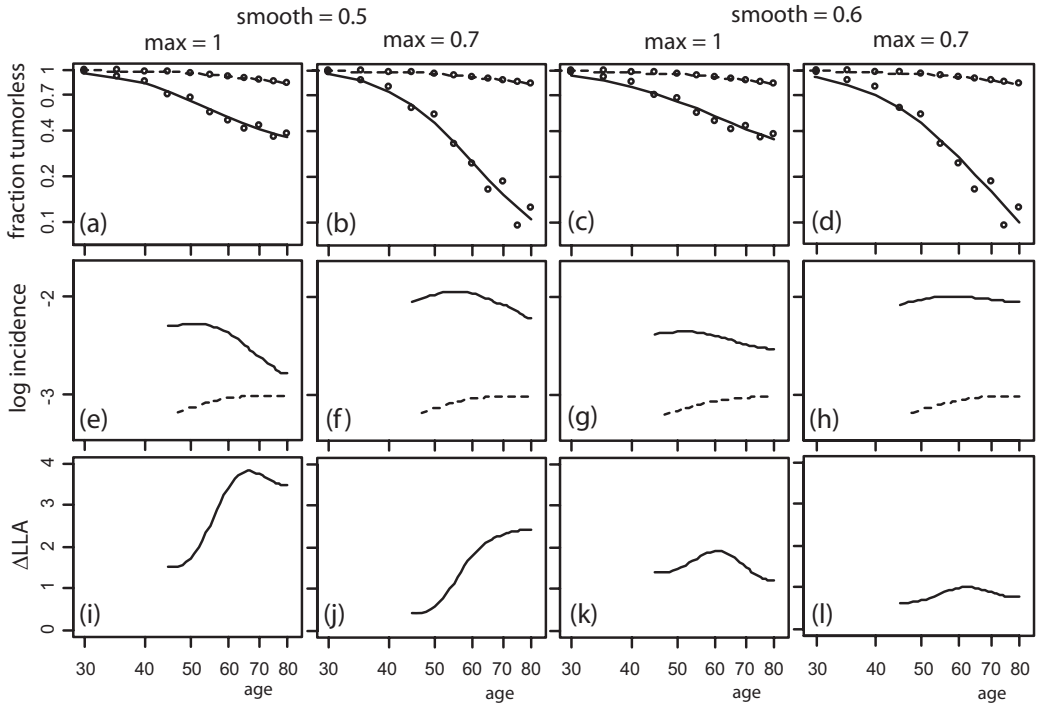

Figure 4.13: Breast cancer rates for females who carry a mutation in BRCA1 or BRCA2, shown as solid lines, versus those females who do not have a mutation shown as dashed lines. The circles in (a) and (c) mark the estimated fraction of females in each class that have not yet developed tumors, taken from figure 1B of Struewing et al. [SHW+97]. In (b) and (d), the observed fraction tumorless, S_{obs}, is transformed to the 'real' fraction tumorless, S_r, via $S_r = \dfrac{max - (1 - S_{obs})}{max}$, where *max* is the fraction of carriers who have fully elevated risk. Panels (a) and (b) used the smooth.spline function of the R computing language (R Development Core Team 2004) to fit a smooth curve to the logarithms of the observed points, with smoothing parameter set to 0.5; (c) and (d) force a stiffer, less curved fit with a smoothing parameter of 0.6. The second row shows incidence on a $\log_{10}$ scale, obtained from -dln(S)/dt, where S, is the fraction tumorless in the curves of the top row. The bottom row shows ΔLLA, the difference in the log-log slopes of incidence in the second row of plots (Redrawn from Frank [Fra07]).

This smoothing parameter dictates the weight of penalty applied to the integral of the squared derivative of the smoothing function during the fitting procedure. The

effect on ΔLLA, as can be seen from the two rightmost columns of figure 4.13, is significant. While it is not unreasonable to use a smoothing algorithm when trying to obtain information from the derivative of an incidence curve, since it is unclear what the most parsimonious choice of the smoothing parameter is, care must be taken in interpreting the smoothed fit. To illustrate this point figure 4.14 shows an analagous plot, comparing the FAP and sporadic colorectal cancer incidence data discussed in section 4.1. Although it is not necessary to use the parameter max to obtain $\Delta\text{LLA} \equiv 1$, as predicted in the previous chapter, in this case a lower smoothing parameter is required to ensure a roughly constant ΔLLA.

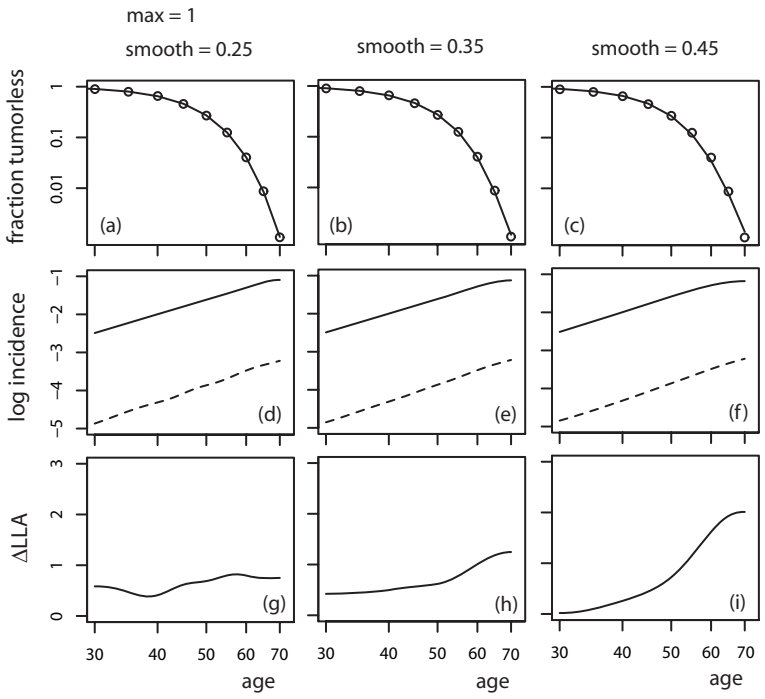

Figure 4.14: Panels (a) to (c): survival rates for male FAP patients who carry a mutation in *APC*. The circles mark the estimated probability of being tumorless at various ages, taken from figure 4.7. Panels (d) through (f) show incidence for carriers and non-carriers (dashed line) on a $\log_{10}$ scale. Non-carrier incidence relates to British males diagnosed in 1961 and is taken from [DPW66]. Panels (g) through (i) show ΔLLA, the difference in the log-log slopes of incidence in the second row of plots.

4.2.5 Heterogeneity and acceleration matching in the case of HNPCC

Can the same technique of correcting for heterogeneity in liability also work to restore the incidence observed in MMR gene mutation carriers to a monotonic increasing function and perhaps a roughly constant acceleration? Figure 4.15 shows ΔLLA calculated from Quehenberger et al. [QVvH05]. ΔLLA was also calculated from Jenkins et. al and Dunlop et al. In each case, the lowest value of max used was the smallest value theoretically possible, i.e. the smallest value larger than the lifetime penetrance. Even using such extreme values for max, the resulting ΔLLA is always an increasing (rather than constant) function of age, starting low and rising to between 3 and 7.

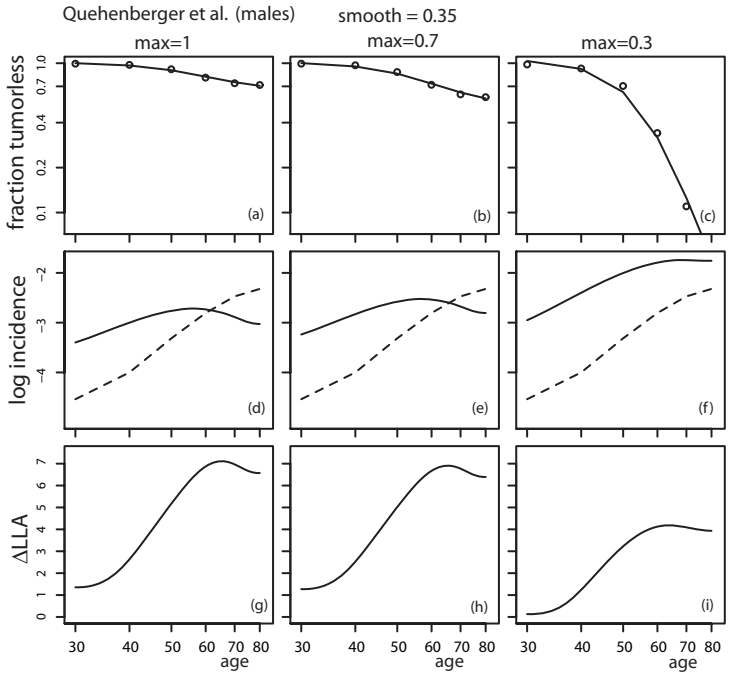

Figure 4.15: Panels (a) to (c): survival rates for male colorectal cancer patients who carry a mutation in *MLH1* or *MSH2*. The circles mark the estimated probability of being tumorless at various ages, taken from table 4 of Quehenberger et al. [QVvH05]. Panels (d) through (f) show incidence for carriers and non-carriers (dashed line) on a $\log_{10}$ scale. Non-carrier incidence is taken from table 3 of Quehenberger et al. Panels (g) through (i) show ΔLLA, the difference in the log-log slopes of incidence in the second row of plots.

4.2.6 Theoretical ΔLLA patterns

The rising ΔLLA observed for CRC in MMR mutation carriers relative to sporadic patients contrasts with the flatter ΔLLA seen in the case of BRCA mutation. This could be explained by the contrasting nature of the *relative* dysfunction caused by MMR and BRCA mutation respectively. Frank has shown that in the simple case of theoretical Armitage and Doll incidence (equation (2.5)), a rising ΔLLA with age is expected if the syndrome-associated genotype causes an increase in the rate of transitions, relative to the healthy genotype. In general, acceleration decreases with age, proportional to the rate at which the average transitional stage, occupied by the separate cell lineages within a tissue, rises with age [Fra07].

4.2.6.1 ΔLLA under Armitage and Doll hazard

As noted above, under the Armitage and Doll hazard, inherited mutations which effectively remove one stage of progression are predicted to cause a roughly constant drop in LLA of around one with age. This is because the Armitage and Doll hazard is approximately log-log linear, with log-log slope equal to one less than the number of stages in progression. This can be seen from the incidence / hazard function, $h(t)$, given by equation (2.5) where N is the number of target lineages in a tissue, n - the number of stages and μ the mutation rate between stages:

$$\begin{aligned} h(t) &= \frac{N\mu^n t^{n-1}}{(n-1)!\sum_{i=0}^{n-1}\frac{(\mu t)^i}{i!}} \\ &\simeq \frac{N\mu^n}{(n-1)!}t^{n-1}, \text{ when } \mu t \ll 1. \end{aligned}$$

Figure 4.16 shows ΔLLA for instances of the Armitage and Doll model, where the healthy patients require their cell lineages to progress through n stages to become malignant, and syndrome-associated-mutation-carrying patients cell lineages need only pass though $n-1$ stages. As expected, in each case, ΔLLA is roughly constant at one. This is the case regardless of which stage in progression is inactivated by the inherited mutation.

Suppose an inherited mutation is assumed to cause an increase in the rates of transitions rather than abrogation of a rate-limiting step. The hazard in mutation carriers can still be described by equation (2.5), but with a new transition rate ν, larger than μ -

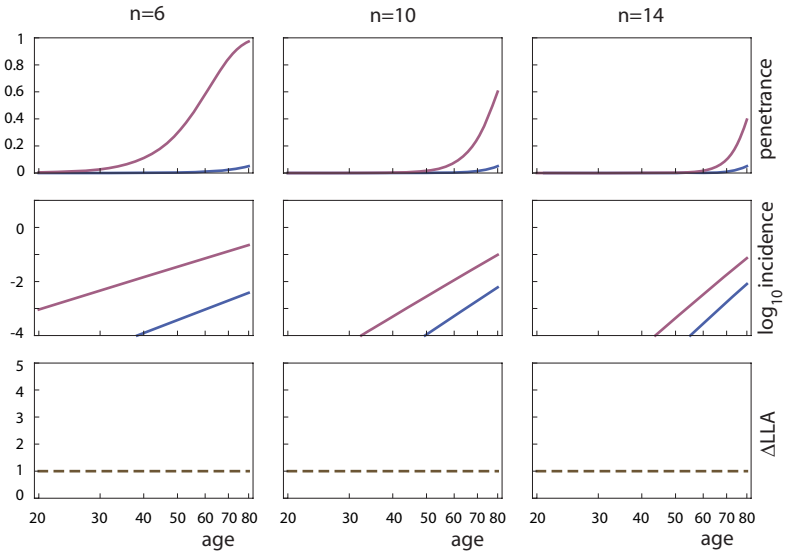

Figure 4.16: ΔLLA arising from a germline mutation which abrogates one of n stages in progression under the Armitage and Doll hazard (equation (2.5)) with number of cell lineages $N = 10^8$, $n = 6, 10, 14$ and μ chosen in each case so that the penetrance at age 80 in the healthy genotype (blue line) is equal to 5%. While the effect on penetrance of the mutant gene (red line) is diminished as the rate of transition between stages increases (with increasing n), ΔLLA remains roughly equal to one.

the transition rate in non carriers. Then the incidence in syndrome associated patients loses acceleration more strongly with age in proportion to ν. However, in order for the fall in acceleration to be as strong as that observed in HNPCC, the transition rate ratio has to be high, $\frac{\nu}{\mu} >> 1$, which results in a much higher incidence than observed in HNPCC (figure 4.17).

Suppose an inherited mutation only increases the rate of some but not all transitions. This situation can be simulated by using the standard Armitage and Doll hazard (equation (2.5)) to model the incidence in non carriers and equation (3.9) for the incidence in carriers. Recall that equation (3.9) describes a lineage transitioning through stages at two different rates. The first $k + 1$ transitions are at rate μ and the remaining steps are at rate ν. Thus, the number of steps with an increased mutation rate ν can then be controlled through the parameter k in equation (3.9), also see figure 3.4. Naturally the penetrance is lower in mutation carriers if fewer transitions are quickened but, again, the model will not reproduce the observed deceleration in MMR mutation

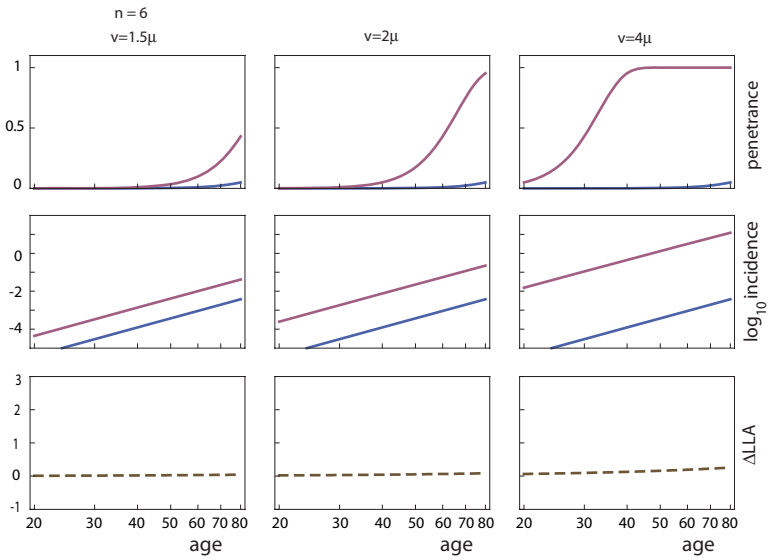

Figure 4.17: ΔLLA arising from a germline mutation which increases the rates of transitions under the Armitage and Doll hazard (equation (2.5)) with $N = 10^8$, $n = 6, 10, 14$ and μ chosen in each case so that the penetrance at age 80 in the healthy genotype (blue line) is equal to 5%. All of the age axes are logarithmic to base 10. Incidence also is plotted on a $\log_{10}$ scale. The increased transition rate, ν, even at only 4 times the original transition rate, causes a strong increase in incidence but only a small deceleration in mutation carriers (red line). Incidence of CRC observed in HNPCC is never higher than 10^{-1}

carriers without too large an increase in incidence (figure 4.18).

A more appropriate hypothesis for the effect of MMR mutations may be that some transitions are quicker in HNPCC patients while some are slower. This hypothesis can produce falling acceleration in the HNPCC incidence curve without raising the overall HNPCC incidence unrealistically. It is consistent with HNPCC following a pathway that is initially distinct from sporadic CRC, requiring extra or slower mutations to precipitate loss of MMR. Figure 4.19 shows ΔLLA for a situation in which the sporadic case has ten stages with μ chosen to force a lifetime penetrance of 5%. The inherited case also has 10 stages but three of these are 100 times slower than in the sporadic case, while the other 7 are 30 times faster.

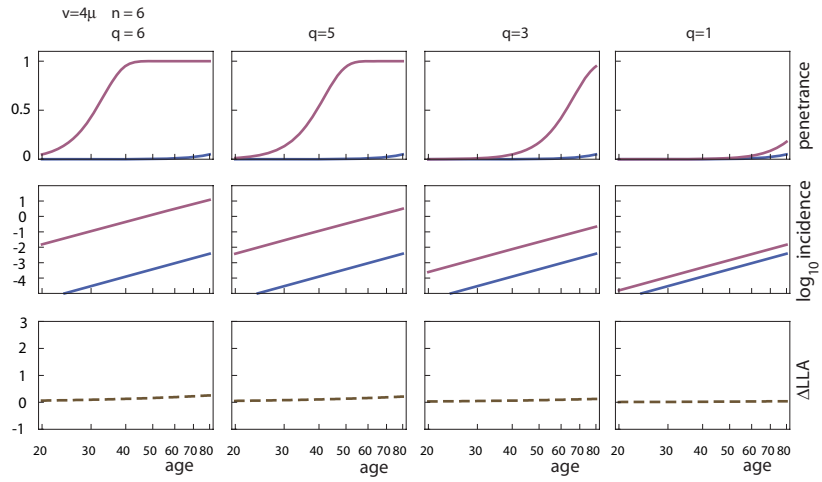

Figure 4.18: ΔLLA arising from a germline mutation which increases the rates of q out of n transitions by a factor of 4 under the Armitage and Doll hazard (equation (2.5)) with $N = 10^8$, $n = 6$, $q = 1, 3, 5, 6$ and μ chosen so that the penetrance at age 80 in the healthy genotype (blue line) is equal to 5%.

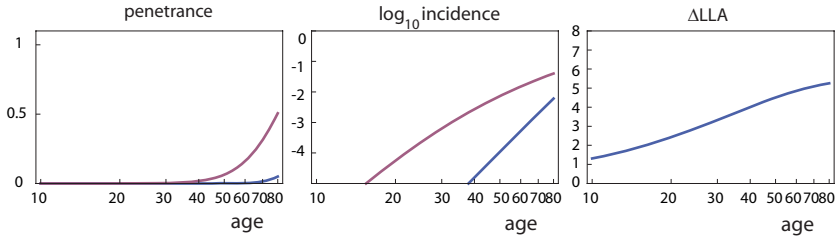

Figure 4.19: ΔLLA arising from a germline mutation which slows the rates of 3 out of n transitions while quickening all other transitions. Incidence in the healthy genotype (blue line) is modelled by equation (2.5) with $N = 10^8$, $n = 10$ and μ chosen so that the penetrance at age 80 is equal to 5%. The heterogeneity in the syndrome associated transition rates produces a plateauing incidence and rising ΔLLA (red line).

4.2.6.2 ΔLLA under clonal expansion

To investigate ΔLLA when clonal expansion features in the multistage sequence, Frank's multistage model [Fra04b], mentioned in section 2.5.4, can be used and will now be described. Let the probability that a cell lineage is in stage i at age t be denoted by $x_i(t)$. Having entered a given stage, i, at age s, the time until transition to stage $i+1$ is governed by an inhomogeneous Poisson process with intensity $v_i y_i(\alpha)$. $\alpha = t - s$

is the time since entry into stage i, v_i is the transition rate per lineage and $y_i(\alpha)$ is a continuous approximation to the number of copies of the lineage existing in stage i, α years after the lineage first entered that stage. The copies are produced through clonal expansion so $y_i(0) = 1$, i.e. there is only one copy of the lineage when first it enters a particular stage. The lineage then multiplies according to:

$$y_i(\alpha) = \frac{K_i e^{r_i \alpha}}{K_i + e^{r_i \alpha} - 1}.$$

This is the same logistic expression used in a different clonal expansion model presented in section 3.2. K_i is the carrying capacity and r_i is the initial growth rate of the clone.

Hence, the probability that a lineage which entered stage i at time s is still there at time t (denoted by $D(t, s)$ where $t > s$) is:

$$\begin{aligned} D(t,s) &= e^{-\int_s^t v_i y_i(z-s)dz} \\ &= \left(\frac{K_i}{K_i + e^{r_i(t-s)} - 1}\right)^{\frac{v_i K_i}{r_i}}. \end{aligned} \tag{4.12}$$

The time until transition from state i to state $i + 1$ given the lineage is in state i at age t, but where the time of entry into state i is unspecified, can also be modelled by the waiting time of a non-homogeneous Poisson process. The intensity in this case, denoted $u_i(t)$, is harder to define because $u_i(t)$ depends on the size of the clone in stage i at time t, which in turn depends on the unknown time of entry into stage i, which in turn depends on all the previous rates of transition $u_j(s)$ where $s < t$ and $j < i$. In effect, $u_i(t)$ must be defined iteratively. It is equal to the following limit:

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} P[X(t + \Delta t) \geq i + 1 | X(t) = i].$$

Here, $X(t)$ is the discrete state random process representing the stage occupied by a cell lineage which starts in the healthy compartment, $i = 0$, at age $t = 0$. Suppose

that the entry time into stage i is known to be s, with $s \leq t$. In this case, by time t the clone has grown to be of size $y_i(t-s)$, so:

$$P[X(t+\Delta t) \geq i+1 | X(t) = i] = 1 - e^{-v_i \int_0^{\Delta t} y_i(t-s+z)dz},$$

i.e. the lineage, having entered state i at time s, and still remaining in state i at time t, will leave state i with density $y_i(t-s+z)$ at age $t+z$. However, since the entry time, s, is not known, an integral over all possible entry times is required. Let S_i denote a random variable representing the entry time of the lineage into state i. The cumulative density for the entry time, conditional on the event $X(t) = i$ is:

$$\begin{aligned} P[S_i \leq s | X(t) = i] &= \frac{P[(S_i \leq s) \cap (X(t) = i)]}{P[X(t) = i]} \\ &= \frac{\int_0^s u_{i-1}(z) x_{i-1}(z) D_i(t,z) dz}{x_i(t)}. \end{aligned}$$

The numerator in the last integral reflects that the probability of entering state i before age s and remaining there until age t is given by integrating the unconditional density for the entry time, $u_{i-1}(z)x_{i-1}(z)$, against the probability of remaining in state i given an entry time of z, $D_i(t,z)$.

Therefore

$$\frac{d}{ds}[S_i \leq s | X(t) = i] = \frac{u_{i-1}(s) x_{i-1}(s) D_i(t,s)}{x_i(t)},$$

and hence

$$\begin{aligned} &P[X(t+\Delta t) \geq i+1 | X(t) = i] \\ &= \int_0^t \frac{u_{i-1}(s) x_{i-1}(s) D_i(t,s)}{x_i(t)} \left(1 - e^{-v_i \int_0^{\Delta t} y_i(t-s+z)dz} \right) ds. \end{aligned}$$

So

$$\begin{aligned}\lim_{\Delta t\to 0}\frac{1}{\Delta t}P[X(t+\Delta t)\geq i+1|X(t)=i] &= u_i(t)\\ &= \int_0^t \frac{u_{i-1}(s)x_{i-1}(s)D_i(t,s)}{x_i(t)}v_i y_i(t-s)ds\\ &= \frac{v_i\int_0^t u_{i-1}(s)x_{i-1}(s)D_i(t,s)y_i(t-s)ds}{x_i(t)}\\ &= v_i\overline{y}_i, \qquad (4.13)\end{aligned}$$

where $\overline{y}_i$ is the expected clone size in the ith department. Hence, the transition rate from state i at age t is v_i, the rate per lineage, multiplied by this expected clone size.

The probabilities of being in stages i through n are given by:

$$\begin{aligned}x_0(t) &= D_0(t,0)\\ x_i(t) &= \int_0^t u_{i-1}(s)x_{i-1}(s)D_i(t,s)ds \quad i=1,\ldots,n-1\\ x_n(t) &= \int_0^t u_{n-1}(s)x_{n-1}(s)ds,\end{aligned}$$

where $u_i(s)$ is given by equation (4.13) and $D_i(t,s)$, (4.12).

This model can be used to test the effect of clonal expansion on ΔLLA. First, consider the basic case where precursor lineages of the syndrome-associated cancer must traverse one less stage than those of the sporadic cancer to become malignant. Assume that a single compartment in the inherited and sporadic cancer undergoes clonal expansion. A constant ΔLLA of one, as predicted under the Armitage and Doll model with no clonal expansion, is no longer expected. The expanding clone initially causes acceleration as the number of target cells at risk increases. However, once the expected clone size of a lineage in that compartment becomes large, the effective transition rate out of the compartment becomes extremely rapid. Acceleration lowers again as transition out of the compartment ceases to be rate-limiting. The net result is a peaked acceleration pattern and associated concave incidence curve on the log-log scale (figure 4.20). The effect of a peaked acceleration pattern on ΔLLA is limited, since acceleration peaks equally in both the sporadic case and the mutant case. However, the peak occurs at an earlier age in the mutant case because the expanding compartment is preceded by

fewer stages. Hence there is a falling ΔLLA at young ages, which rises again once the sporadic acceleration catches up. Even for a large clone, the effect is small though (see figure 4.20) and the stability of ΔLLA under the complex acceleration patterns suggested by figure 4.20 is an endorsement of its potential utility.

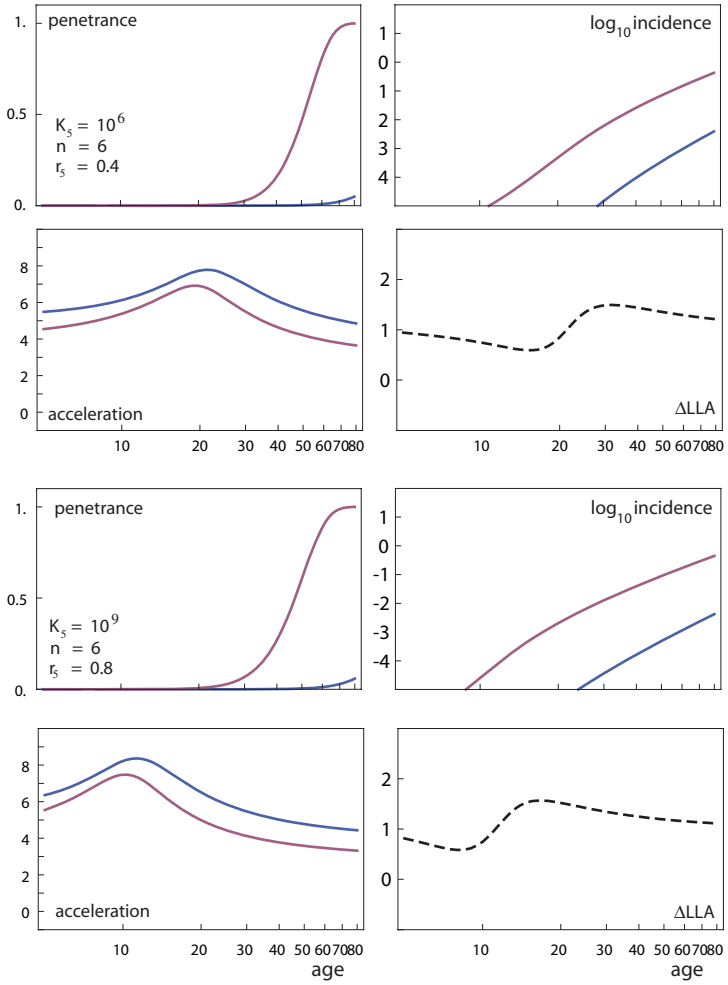

Figure 4.20: Penetrance, log-incidence and ΔLLA assuming a six-step pathway with one clonal expansion in healthy patients (blue line) and the same pathway with one step deleted in mutation carrying patients (red line). The top four panels are calculated assuming a clonal expansion in the final stage with capacity of $K_5 = 10^6$ cell lineages and initial growth rate of 0.4. The bottom four panels assume a faster growing and larger clone. In either case the mutation rate per lineage is chosen so that cumulative risk to age 80 is 5% in healthy patients. The effect of the more aggressive clone is only to shift the kink in ΔLLA to earlier ages. The departure from a constant ΔLLA of one remains small in either case.

Given the strong modulating effect clonal expansion has on acceleration, it is tempting to suggest that differences in clonal expansion could contribute to the rising ΔLLA between HNPCC and sporadic colorectal cancer. However, if a difference in clonal expansion is responsible for elevating risk in the context of syndrome-associated patients, then necessarily the clonal expansion must be stronger in the syndrome-associated patients in order that the penetrance of the disease be increased. Since clonal expansion increases acceleration, higher clonal expansion in mutation carriers will cause a negative ΔLLA (see figure 4.21). So under this particular model of progression it is unlikely that faster growing clones are a primary mechanism of risk modulation in mutation carriers.

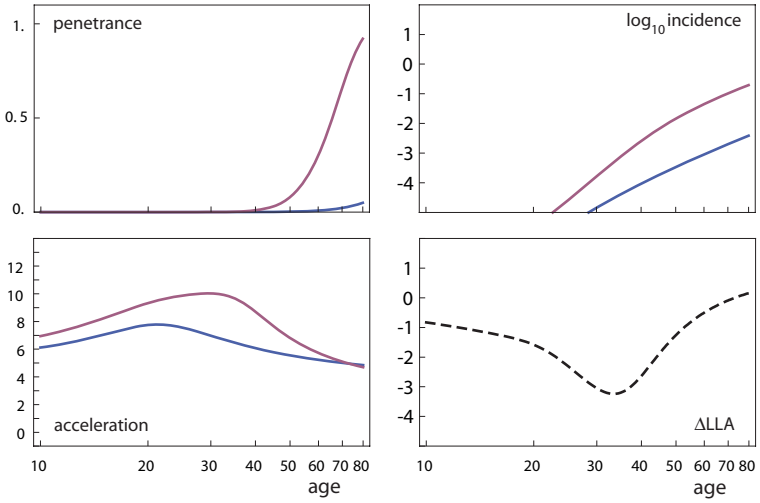

Figure 4.21: Penetrance, log-incidence and ΔLLA assuming a six step pathway with one clonal expansion in healthy patients (blue line) and the same pathway but with two clonal expansions in mutation carriers (red line). All clones have capacity $K = 10^6$ and initial growth rate $r = 0.4$. The mutation rate per lineage is chosen so that cumulative risk to age 80 is 5% in healthy patients. The mutation carriers out-accelerate the healthy patients throughout mid-life causing a negative ΔLLA.

4.2.7 Sporadic MSI+ colorectal cancer

A limitation of the data used to generate the observed ΔLLA patterns discussed above is that HNPCC colorectal cancers are known to progress along a pathway which diverges, to an extent, from that of sporadic CRC [ISTB99]. One cause of this divergence is thought to be the contrasting type of genome destabilization found in each of

the cancers. Most sporadic cancers show evidence of chromosomal instability (CIN) (chromosome losses/duplications, mitotic recombinations, and large deletions) while HNPCC cancers show evidence of microsatellite instability (MSI). MSI is manifest as variable length microsatellites (repetitive DNA sequences with a short repeating unit) and an increased point mutation rate indicative of lost miss-match repair (MMR) function. In sporadic CRCs which are microsatellite stable (MSS or MSI-) certain genes are mutated that are not selected in HNPCC. Alternative genes, lying, for example, in the same signalling pathways, but containing coding micro-satellites, are targeted in preference, by the MSI phenotype [SKP$^+$99, YAN$^+$98]. Other distinctions between HNPCC and MSS sporadic CRC, may be caused by non-hypermutation-related changes in selection pressure arising in the context of MMR mutation [JVH$^+$05]. Irrespective of the driving force behind the divergence of HNPCC and MSS sporadic CRC, their phenotypes are clearly distinguishable, in terms of their position in the colon, prognosis and histological features [dlC03]. A closer relation of HNPCC CRC, therefore, ought to be the minority of sporadic colorectal cancers that are MSI+. This seems to be the case. While observable phenotypic differences exist [YSB$^+$01] and there is debate over the extent of their similarity, HNPCC and MSI+ sporadic CRC are certainly closer in terms of their aetiology than HNPCC and MSS sporadic CRC. For example, MSI+ sporadic CRC shows biallelic inactivation of *MLH1* by promoter hypomethylation [BDR$^+$07, YSB$^+$01]. Other similarities distinguishing MSI+ sporadic and HNPCC CRC from MSS CRC include a lower frequency of *APC* mutation compared with MSS CRC [JBF$^+$03, KKYT$^+$96, SRV$^+$06, SKP$^+$99] and raised frequencies of *TGFβR2* mutation [YIM$^+$06, FSW$^+$98, YAN$^+$98, YSB$^+$01, JSD$^+$06, TSO$^+$01, SKP$^+$99, FPNO$^+$05] and *BAX* mutation [YSW$^+$98, YAN$^+$98, RYI$^+$97, FPNO$^+$05] with associated reduction in *P53* mutation [LdLJ$^+$97, YAN$^+$98, KKYT$^+$96, SKP$^+$99, KKYT$^+$96]. Notable differences between HNPCC and MSI+ sporadic CRC include a significant frequency of *β-catenin* mutations in HNPCC which are never found in sporadic MSI+ CRC [MIK$^+$99, JLC$^+$05, JVH$^+$05, SKP$^+$99]. Conversely, *BRAF* mutations are very common in sporadic MSI+ while very rare in or absent from HNPCC where *K-Ras* mutation is more likely found [KSW$^+$04, MWW$^+$04, DBC$^+$04, MVSC$^+$07]. *BRAF* mutation is not only associated with *MLH1* promoter methylation but also thought to correlate with a general increase in the frequency of promoter methylation at other genes [STK$^+$07].

So, another potential distinction between MSI+ sporadic and HNPCC is a higher level of promoter methylation in the former that could influence tumorigenesis beyond the inactivation of *MLH1* [WSC+06, eaY02]. On balance, it seems reasonable to assume that HNPCC and MSI+ sporadic CRC differ at least in terms of the mechanism through which MMR is silenced and possibly beyond that. Still, by comparing penetrance in HNPCC specifically with the risk of MSI+ sporadic CRC, the effect of an inherited mutant MMR gene ought to be isolated more cleanly.

4.2.7.1 Estimating the penetrance of MSI+ CRC

An estimate of age-related sporadic MSI+ CRC penetrance can be derived from the data of Salovarra et al. [SLK+00] and Aaltonen et al. [ASK+98]. They attempted to determine the MSI status of every CRC diagnosed at nine regional hospitals in southeast Finland over a four-year period running from May 1994 until June 1998. They managed to achieve this for just over one thousand cases (approximately 60% of all cases diagnosed in the catchment area of the hospitals in question [SLK+00]). Figure 4.22 shows the number of cases typed as MSS and MSI+ respectively, binned into five-year age groups. MSI+ cancers are less prevalent and account for 12% of all cases, as expected from other studies [BTH+98]. Figure 4.22 suggests that incidence of sporadic MSI+ CRC is not strictly proportional to that of sporadic MSS CRC. Hence the two cancer types must have different acceleration patterns.

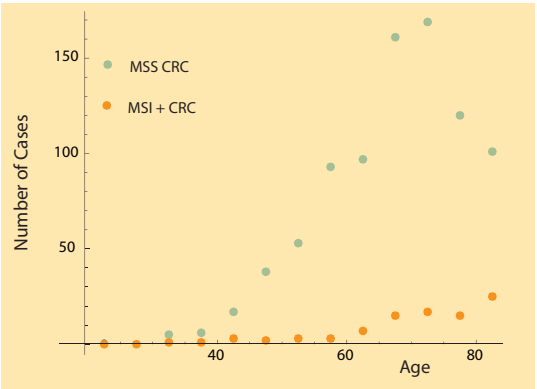

Figure 4.22: Number of cases of MSS and MSI+ CRC occurring at 9 regional hospitals in southeast Finland over a four-year period.

Estimating the penetrance or incidence of MSI+ CRC from these data is difficult because the population in which the cases arise is not clearly defined. The

nine hospitals included Helsinki University Hospital and hospitals serving defined healthcare districts in southeast Finland: Kymenlaakso, Etelä-Karjala, Etelä-Savo, Itä-Savo, Pohjois-Karjala, Pohjois-Savo and Keski-Suomi. The structure of their catchment population can be estimated. 'Statistics Finland' (a government agency - `http://www.stat.fi/index_ en.html`) holds historical population figures for ages 0 - 75 by geographical region rather than healthcare district. However, the geographical regions Uusimaa, Itä-Uusimaa, Kymenlaakso, South Karelia, Etelä-Savo, Pohjois-Savo, North Karelia and Central Finland together overlap with the healthcare districts in question. Figure 4.23 shows an estimate of the catchment population based on these geographical regions. For comparison an estimate made from recent data from the Finnish Cancer Registry (`http://www.cancerregistry.fi`), that run to age 85 and are tabulated by healthcare district, is also given. Most of the regions have a very similar age structure, with the exception of Uusimaa, which includes Helsinki and has a younger population.

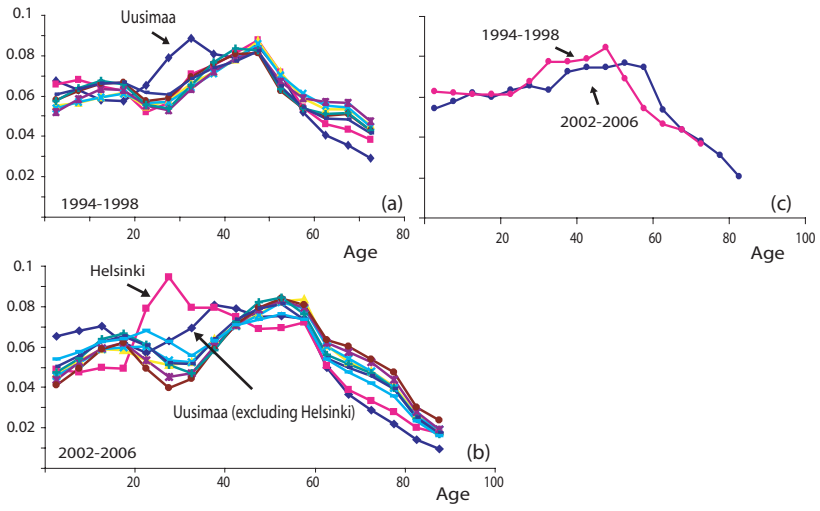

Figure 4.23: Estimated age-structure of the population served by nine regional hospitals in southeast Finland. (a) Frequency by 5-year age group, averaged over the years 1994-1998, taken from 'Statistics Finland' for 8 geographical regions of Southeast-Finland. (b) Frequency of population by age in 8 healthcare regions, taken from the Finnish cancer registry. (c) Population age-structure of the combined regions in (a) and in (b).

By combining the (1994-1998) population structure for ages 0-75 with the more recent (2002-2006) population structure estimate for older ages, a simple estimate of

the hazard function for MSI+ CRC can be made. First of all, the combined age structure is used to estimate the population at risk. A total population size is chosen so that the number of cases of CRC per unit population matches roughly the incidence in Finland as a whole over the period (see figure 4.24, left panel). The incidence of MSI+ CRC can then be calculated using the notional population (defined by this total size and age structure), and the MSI+ case count data (figure 4.24, right panel).

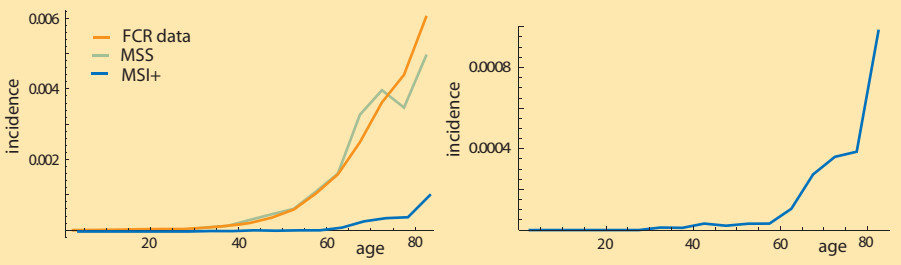

Figure 4.24: (Green line): average annual cases of CRC per unit population during 1994-1998, estimated from nine hospitals in southeast Finland, assuming a 300,000 catchment population. (Orange line): nationwide cancer incidence over the period 1993-1997, taken from the Finnish Cancer Registry [PWF+97]. (Blue lines): MSI+ incidence estimated from the nine hospitals data.

The incidence of MSI+ CRC in the catchment population of the nine hospitals seems to rise more sharply in old age than the incidence of CRC in general. As a consequence, MSI+ CRC has a strongly rising acceleration (figure 4.25). The strongly rising acceleration pattern in turn confirms the rising ΔLLA relative to HNPCC (figure 4.26).

4.3 Discussion

In the first half of this chapter, a comparative analysis of the incidence of sporadic CRC and the hereditary bowel cancer syndrome, FAP, was presented. Molecular analyses indicate that the two cancers differ only by virtue of an inherited germline mutation in the *APC* gene. Assuming this simple relationship enabled an estimate of the rate of *APC* mutation that did not require knowledge of clonal expansion patterns or other aetiological details. This estimate is a useful addition to the few estimates of in-vivo gene mutation already in the literature and shows the potential of studies which focus

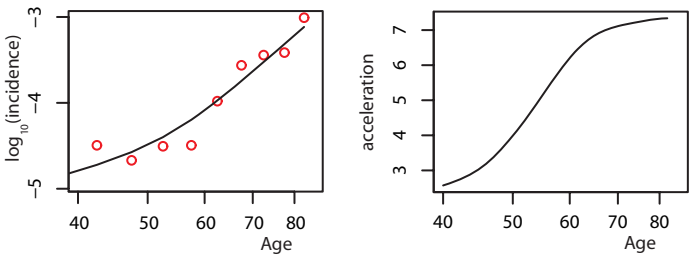

Figure 4.25: (Left): The red circles are log of incidence of MSI+ CRC, estimated from the data of Salovarra et al. [SLK$^+$00] and Aaltonen et al. [ASK$^+$98]. The circles are fit with the smooth.spline function of the R computing language, with smoothing parameter set to 0.5. (Right): LLA calculated from the smoothing spline opposite.

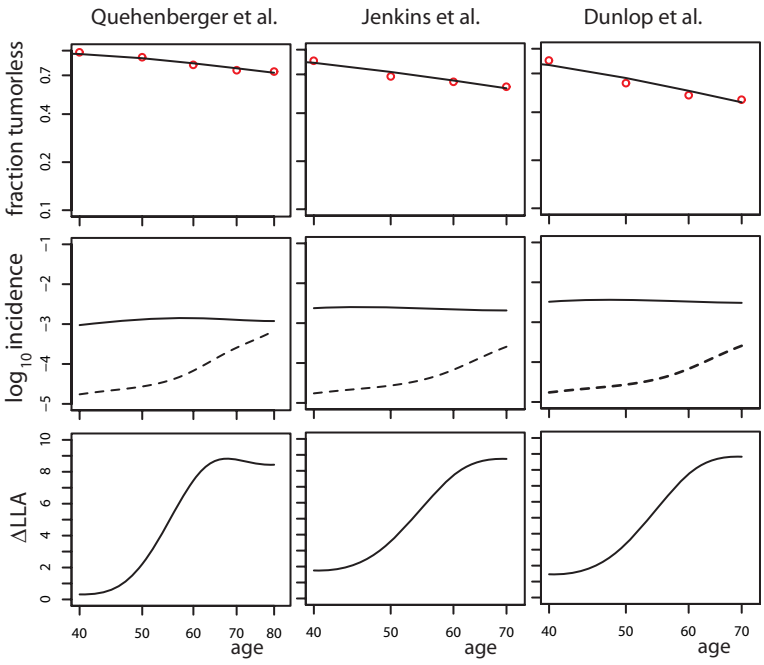

Figure 4.26: (Top row): the red circles show the fraction of MMR mutation carriers who are tumourless on a log scale, estimated in three studies of HNPCC penetrance. The data are fit with smoothing splines as in figure 4.13. (Middle row): the solid lines show incidence derived from the smoothing splines above. The dashed lines show incidence of MSI+ sporadic CRC as estimated above (see figure 4.24). (Bottom row): ΔLLA calculated as the difference in gradient between the solid and dashed lines from the middle row.

on changes in incidence patterns arising in the context of identified germline mutations [HPT08].

Attention was then turned to another hereditary bowel cancer syndrome, HNPCC. CRC penetrance in HNPCC, in contrast to FAP, suggests a more complex relationship between HNPCC and sporadic CRC. Using the computational machinary outlined by Frank [Fra07] it was difficult to isolate a single effect that could cause the observed plateauing of HNPCC incidence relative to sporadic CRC or, in particular, relative to sporadic MSI+ CRC. The most promising hypothesis for this phenomenon, suggested by Franks measure of ΔLLA in the exploratory analysis above, is that MMR mutation causes a slowing of some transitions in HNPCC patients coupled with a quickening of other transitions. This combination allows a significant drop in acceleration to be generated in HNPCC patients without penetrance becoming too severe. Nevertheless, the data can be fit in a variety of different ways that suggest other, more complicated hypotheses, drawing on combinations of changes in clonal expansion parameters, numbers of transitional stages and mutation rates. The simple idea of increased transitions in some stages and slower transitions in others is appealing because it fits with the known function of MMR genes and also the observation that adenomas in HNPCC develop no faster than those in sporadic patients, but progress to malignancy more quickly. However, the incidence data alone support this hypothesis only very weakly. In fact, the observed plateau in penetrance of CRC associated with an MMR mutation is at odds with multistage theory. To resolve this issue it was necessary to hypothesize that as few as 30% of MMR carriers are at increased risk of CRC, but that this small sub-population has complete penetrance. Heterogeneity of this kind has the potential to distort population incidence, so that it ceases to reflect the risk profiles of individuals. In the next chapter direct methods for quantifying heterogeneity are developed in order to better understand the relationship between individual risk and population incidence.

Chapter 5

Population variance in cancer liability

5.1 Introduction

The models presented so far in this thesis have made the assumption that all patients within the population of interest have the same cancer risk. Exception has only been made for distinguishable groups with rare cancer syndromes such as HNPCC and FAP, or defined and observable environmental risk factors such as smoking. This dichotomous view of cancer liability is probably inadequate. It is possible that many interacting loci influence cancer risk and that exposure to environmental risk factors varies widely within populations. The extent to which genetic and environmental factors cause population variance in liability informs the validity of inferences made on the assumption of a homogeneous population. It also tells us how concentrated the cancer burden may be in high risk subsets of a given population. One of the strongest pieces of evidence for variance in cancer liability is the existence of high risk families whose members presumably share genetic and / or environmental risk factors. Known genetic syndromes are the most obvious candidates to explain such clustering. To start this chapter a model of susceptibility owing to a single dominant locus is presented and used to show that, in fact, such rare Mendelian cancer syndromes have insufficient impact to explain the familial clustering of cancer. A polygenic model, positing many common low penetrance risk alleles, is then developed as an alternative theory for the high relative risks observed in first degree relatives of affected patients. Evidence for environmental influences on cancer susceptibility are subsequently discussed and finally an estimate of population variance owing to the combined effects of genes and environment is presented. This estimate is based on the observation that patients who have had cancer

once are more likely to be afflicted again than healthy age-matched controls.

5.2 Genetic liability to cancer

Consider an intuitive definition of liability, quantified by the parameter (l). Let l for a given patient be equated with that individual's lifetime risk of cancer (probability of cancer before age 80), so that $0 \leq l \leq 1$. Ignoring environmental factors and ascribing an increase in risk only to those patients with a particular Mendelian cancer syndrome is tantamount to assuming a discrete liability distribution. Figure 5.1 shows such a discrete liability distribution based on a simplistic notion of colorectal cancer susceptibility, i.e. that the large majority of patients have a lifetime risk of 5% while a small fraction with known CRC cancer syndromes have higher liability.

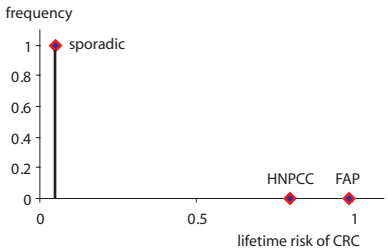

Figure 5.1: Simplistic liability distribution for CRC. Population frequency of FAP taken as 1:10000 [BFB$^+$94] and population frequency of germline MMR mutation taken as 1:3000 [DFN$^+$00]. For a review of further hereditary CRC syndromes that are rarer still see Lynch and de la Chapelle [LdlC03]

A discrete distribution of the type shown in figure 5.1 is unsuitable for many cancers. It is at odds with the extent to which cancer typically clusters in families. Risk of cancer in first degree relatives of affected individuals is around twice the risk in the general population [HC02]. In particular, the CRC risk to siblings of patients affected with CRC is more than twice the population risk [BHP06].

5.2.1 Sibling risk owing to a rare dominant single gene syndrome

It is perhaps intuitively obvious that rare highly penetrant germline variants such as mutant *APC* or *MLH1*/*MSH2* cannot be responsible for a doubling of risk in siblings or first degree relatives in general. Nevertheless, it is instructive to see what type of familial aggregation such rare predisposing mutations can produce. Consider the simple

nuclear family shown in figure 5.2.

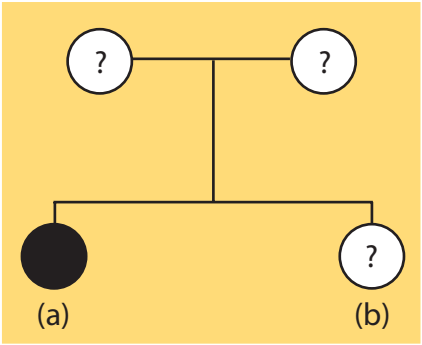

Figure 5.2: Given that one offspring is a confirmed case of cancer (sibling (a) shown in black), the probability that the other will develop cancer within a lifetime depends on, among other things, the existence of predisposing allelic variants in the population.

Supposing there is a single, autosomal susceptibility locus. If the population of interest contains two variants at this locus, D and d, then a simple dominant model of liability is that individuals with at least one D allele, (i.e. heterozygotes or DD-homozygotes) have lifetime risk $R \times s$, with $R > 1$ while dd individuals have a baseline risk s. If the allelic frequencies of D and d are p and q respectively ($q = 1 - p$), then, assuming Hardy-Weinberg equilibrium, the lifetime risk, K, of cancer in an individual of unknown genotype is given by:

$$K = (p^2 + 2pq)Rs + q^2 s. \tag{5.1}$$

By comparison, lifetime risk, K_s, for the sibling of a confirmed case (sibling (b) in figure 5.2), is:

$$\begin{aligned} K_s &= P[\text{(b) has cancer}|\text{(a) has cancer}] \\ &= \frac{P[\text{cancer in both siblings}]}{P[\text{(a) has cancer}]}. \end{aligned} \tag{5.2}$$

$P[\text{(a) has cancer}] = K$. $P[\text{cancer in both siblings}]$ depends on the genotypes of the parents. For example if the parents are $DD \times DD$ then both siblings must be DD and have risk Rs. Hence the chance they both get cancer would be $(Rs)^2$. Under

random mating, parents with $DD \times DD$ occur with frequency $p^2 \times p^2$. Considering all possible parental genotypes in this manner leads to:

$$\begin{aligned} P[\text{cancer in both siblings}] = (p^2)^2(Rs)^2 &+ 2(p^2)(2pq)(Rs)^2 \\ &+ 2(p^2)(q^2)(Rs)^2 \\ &+ (2pq)(2pq)((3/4)Rs + (1/4)s)^2 \\ &+ 2(q^2)(2pq)((1/2)(s + Rs))^2 \\ &+ (q^2)^2(s^2). \end{aligned} \tag{5.3}$$

Equations (5.1), (5.3) and (5.2), can be used to calculate sibling relative risk, λ_s, through:

$$\lambda_s = \frac{K_s}{K^2}, \tag{5.4}$$

which is a function of p and R but independent of s.

Table 5.1 and figure 5.3 show λ for various allele frequencies, p, and relative risk values R.

p \ R	4	8	16
0.00001	1.00	1.00	1.00
0.0001	1.00	1.00	1.02
0.001	1.01	1.05	1.21
0.01	1.08	1.37	2.30
0.1	1.27	1.68	2.14

Table 5.1: Sibling relative risk, λ_s, as a function of disease allele frequency, p, and genotype relative risk, R.

Under the dominant model, the deleterious allele frequency, p, is roughly half the frequency of affected carriers when p is small. So, taking the example of HNPCC, where the frequency of affected carriers is roughly 1/3000 [DFN$^+$00], the deleterious allele frequency is 1/6000. The penetrance of colorectal cancer in HNPCC is at most 0.80 (vs 0.05 in the general population), which translates into a maximum relative risk of 16. Hence, by equation (5.4), the resulting sibling risk owing to HNPCC is only

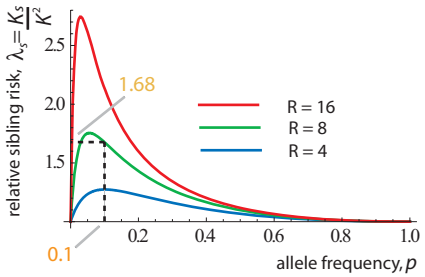

Figure 5.3: Relative sibling risk as a function of deleterious allele frequency for various genotype relative risk values. Calculated from equation (5.4) which assumes random mating. The biphasic nature of this graph can be explained as follows: for very low allele frequencies an affected sibling is only in rare cases likely to carry the deleterious allele, and hence the sibling risk approaches the population risk as the allele frequency tends to zero. For very high allele frequencies, an affected sibling will likely carry the allele but then so will most of the population so the sibling risk approaches the population risk also as the allele frequency tends to one. The data point highlighted in red in table 5.1 is shown.

1.04. Other heritable effects or environmental sharing must be present to cause an observed sibling risk of 2. One possible explanation for a lack of observable genetic syndromes to explain the familial clustering of cancer is the existence of many common low penetrance susceptibility alleles, which are difficult to identify individually by linkage analysis, but which can nevertheless act in combination to produce sizable effects [TWCC+07].

5.2.2 Offspring risk under multiplicative polygenic susceptibility

The following model can be used to quantify risk in offspring inherited through such multiple low penetrance alleles. Consider the offspring of an affected parent, with otherwise unknown pedigree information (e.g. child (a) in figure 5.4).

Suppose there are n susceptibility loci and that the population in question contains two variants at each locus, one dominant risk-conferring allele, D_i, at frequency p, and one wildtype allele, d_i, at frequency $1-p$ for $i = 1, ..., n$. Suppose these loci are unlinked and that in each case the dominant allele confers a relative risk of R. Under this scenario the liability, l, of an individual with genotype $(g_1, g_2, ..., g_n)$ where $g_i \in \{d_i d_i, D_i d_i, D_i D_i\}$ is:

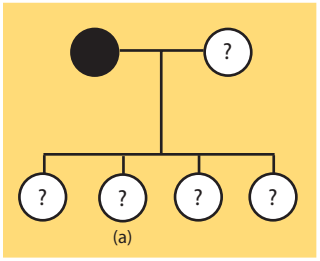

Figure 5.4: A nuclear family with undetermined number of offspring. Given that one parent is a confirmed case of cancer, the probability that a given offspring will develop cancer within a lifetime again depends on the existence of predisposing allelic variants in the population.

$$l[(g_1,...,g_n)] = s \prod_{i=1,...,n} \left(\frac{l[(d_1d_1,\ldots,d_{i-1}d_{i-1},g_i,d_{i+1}d_{i+1},\ldots,d_nd_n)]}{s} \right),$$

Here, $s = l[(d_1d_1,...,d_nd_n)]$ is the wildtype liability and

$$l[(d_1d_1,\ldots,d_{i-1}d_{i-1},g_i,d_{i+1}d_{i+1},\ldots,d_nd_n)] = \begin{cases} s & g_i = d_id_i \\ R \times s & g_i \in \{D_id_i, D_iD_i\} \end{cases}.$$

The genotype frequency of $(g_1, g_2, ..., g_n)$ under Hardy-Weinberg equilibrium is:

$$f[(g_1,\ldots,g_n)] = \prod_{i=1,...,n} f[g_i], \tag{5.5}$$

where $f[D_iD_i] = p^2, f[D_id_i] = 2p(1-p)$ & $f[d_id_i] = (1-p)^2$.

Consequently, the general population risk, K is:

$$K = \sum_{g_1 \in \mathbb{G}_1} \sum_{g_2 \in \mathbb{G}_2} \cdots \sum_{g_n \in \mathbb{G}_n} f[(g_1,\ldots,g_n)] l[(g_1,\ldots,g_n)], \tag{5.6}$$

where $\mathbb{G}_i = \{d_id_i, D_id_i, D_iD_i\}$ for $i = 1,\ldots,n$.

The risk to the offspring of an affected parent can be calculated similarly as in the case of the sibling risk derived above. The aim is to calculate the chance of cancer in a particular offspring (denote this event C_o) given the event C_p - cancer in the parent. Again,

$$P[C_o|C_p] = \frac{P[C_o \cap C_p]}{P[C_p]} = \frac{P[C_o \cap C_p]}{K},$$

since $P[C_p] = K$ (equation (5.6)). The relative offspring risk is then:

$$\frac{P[C_o|C_p]}{K} = \frac{P[C_o \cap C_p]}{K^2}, \tag{5.7}$$

$P[C_o \cap C_p]$ can be calculated by considering the different possible combinations of genotypes in the parents and offspring. If the genotype of the affected parent is $\underline{G}_p = (g_1^{(p)}, \ldots, g_n^{(p)})$ and the offspring genotype is $\underline{G}_o = (g_1^{(o)}, \ldots, g_n^{(o)})$ then $P[C_o \cap C_p|\ \underline{G}_p, \underline{G}_o] = l[\underline{G}_p]l[\underline{G}_o]$. So,

$$P[C_o \cap C_p] = \sum_{\underline{G}_p} \sum_{\underline{G}_o} P[\underline{G}_p]P[\underline{G}_o|\underline{G}_p]l[\underline{G}_p]l[\underline{G}_o],$$

where $P[\underline{G}_p] = f[\underline{G}_p]$ (equation (5.5)) and $P[\underline{G}_o|\underline{G}_p]$ is given by:

$$P[\underline{G}_o|\underline{G}_p] = \prod_{i=1,\ldots,n} P[g_i^{(o)}|g_i^{(p)}].$$

$P[g_i^{(o)}|g_i^{(p)}]$ is calculated by considering the three possible genotypes, $g_i^{(m)}$, at locus i in the affected parents mate. This is done below for some examples of $g^{(o)}$ and $g^{(p)}$ (the i subscript is dropped for convenience) assuming allele D has frequency p:

$$\begin{aligned}
P[g^{(o)} = dd|g^{(p)} = dd] &= P[g_i^{(m)} = dd] + \frac{1}{2}P[g_i^{(m)} = Dd] + 0P[g_i^{(m)} = DD] \\
&= (1-p)^3 \\
P[g^{(o)} = Dd|g^{(p)} = dd] &= 0P[g_i^{(m)} = dd] + \frac{1}{2}P[g_i^{(m)} = Dd] + P[g_i^{(m)} = DD] \\
&= p(1-p)^2 \\
P[g^{(o)} = DD|g^{(p)} = dd] &= 0.
\end{aligned}$$

Figure 5.5 shows relative offspring risk (equation (5.7)) assuming different numbers of risk loci, n, and allele frequencies p. At each locus the risk conferring variant

is assumed to give a doubling of risk (i.e. R=2). For bowel cancer the relative familial risk, when HNPCC, FAP and other known syndromes are excluded is estimated at 1.5 [JH01]. At least eight common, low risk (R=2) alleles with minor allele frequency $0.1 < p < 0.3$ are required to produce such a familial clustering.

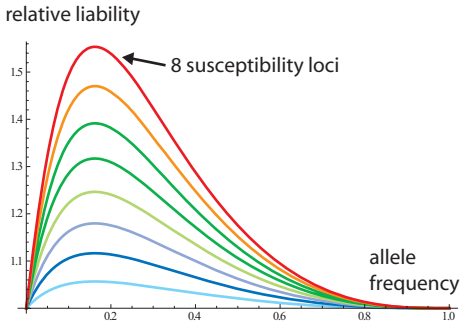

Figure 5.5: Relative offspring risk based on 1,2, ... ,8 susceptibility loci plotted against allele frequency.

If many alleles underlie the familial aggregation of cancer, then a large proportion of cases will involve a hereditary component and the discrete liability distribution shown in figure 5.1 will be violated, giving way to a lognormal distribution of liability (figure 5.6).

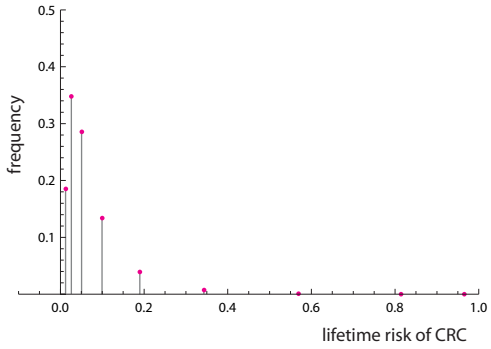

Figure 5.6: Notional liability distribution for "sporadic" CRC based on a multiplicative polygenic model with 8 risk loci, p=0.1 and R=2. The baseline risk, s, is fixed so that mean lifetime risk is 0.05. The distribution arising from the product of many independent positive valued random variables tends to be lognormal.

5.3 Environmental liability to cancer

Abundant evidence exists for a strong environmental component to cancer incidence. Migrants frequently adopt the cancer rates of their new country. Consequently, the greater than ten-fold differences in incidence between populations worldwide have largely been attributed to environment and lifestyle factors [HBF06]. It seems reasonable to hypothesize that shared environment among family members also accounts for a degree of the familial aggregation of cancer described in the previous section. Lichtenstein et al. [LHV$^+$00] used twin data and the multi-factorial threshold (MFT) model to estimate the relative contribution of genes and environment to many cancers. In the MFT model, liability in a population of twin pairs, l, is assumed to vary normally about its mean, μ, in response to independent and normally distributed genetic and environmental factors:

$$l - \mu = g + c + e.$$

g represents inherited genetic effects, c represents shared environmental effects between twins raised in the same environment and e represents non-shared environmental effects. Cancer is treated as a binary trait arising in all those individuals whose liability exceeds a threshold value. The liability of a twin pair, (l_1, l_2), is assumed to follow a bivariate normal distribution. The covariance of l_1 and l_2 depends on whether the twins are monozygotic or dizygotic. In the monozygotic case:

$$\text{Cov}(l_1, l_2) = \sigma_g^2 + \sigma_c^2,$$

whereas, in the dizygotic case, the covariance arising from the genetic effect is halved to $\frac{1}{2}\sigma_g^2$ on account of the reduced relatedness between dizygotic twins [Yan00]. So, if the genetic effect is large, the MFT model predicts that cancer status will concord more often in monozygotic twins than in dizygotes.

Fitting the MFT model to twin data seems to imply that shared environment has a limited role in familial aggregation for most cancers. For example, in the case of colorectal cancer, Lichtenstein et al. estimate that while 35% of the variance in liability is inherited / genetic, only 5% by contrast can be apportioned to shared environmental

effects. Results of this kind are supported by data showing low spouse concordance at many sites [HDV01], excepting stomach and lung cancer. Although the MFT model implies that environment has a limited role in familial aggregation, the environment is nevertheless strongly implicated in overal disparities in risk between individuals. In CRC for example, liability variance due to non-shared environmental effects was estimated at 60% of the total liability variance [LHV$^+$00].

5.4 Temporal environmental variance in cancer liability

Genes and environment both appear to make significant contributions to cancer susceptibility. Twin studies suggest that inheritance is largely responsible for familial clustering but that environmental factors dominate overall liability variance. A precise apportioning of blame to genes and environment through such studies, however, is wrought with difficulties [HBF06]. The classical assumption that dizygotic and monozygotic twins share environmental risks to the same extent remains controversial as does the assumed additive interaction of genetic influences and the environment [BTH05]. Another unsatisfactory aspect of the MFT model is that the implied genetic contribution to susceptibility depends on population prevalence. For example, heritability, a relative measure of the genetic effect, is defined by:

$$h = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_c^2 + \sigma_e^2}. \tag{5.8}$$

Figure 5.7 shows that for a fixed ratio in the risk to a monozygotic vs dizygotic twin of an affected patient, the implied heritability increases with increasing prevalence, K.

If determining the relative contributions of distinct factors to liability is difficult, perhaps a more feasible aim is simply to quantify the total variance in cancer susceptibility within a population. In section 5.5 this is attempted by comparing risks of second primary malignancies in cancer patients (i.e. new primaries unrelated to the first cancer), with risks in unaffected individuals of the same age and birth cohort. Before addressing the question of intra-cohort variance in liability, also of interest are changes in susceptibility patterns that occur with calendar time. Figures 5.8, 5.9 and 5.10 show

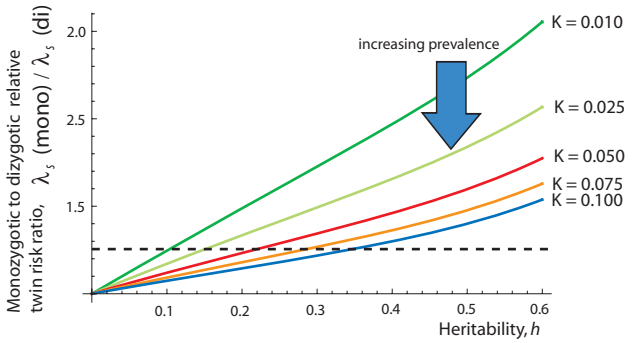

Figure 5.7: Under MFT, the heritability h (equation (5.8)) implied by a fixed ratio in twin relative risk, between monozygotic or dizygotic twins, increases with increasing population prevalence, K. $\sigma_g^2 + \sigma_c^2 + \sigma_e^2$ was normalized to one in the calculations of twin relative risk and σ_c^2 was fixed at 5%.

temporal trends in cancer incidence for breast, colorectal and prostate cancer. For each age between 0 and 85, the incidence rates recorded in Connecticut each year between 1973 and 2005 are plotted. This gives an impression of the variance in disease counts per unit population.

In the absence of calendar year effects, Poisson variance, i.e. variance equal to the mean, is expected. If liabilities among individuals of the same birth cohort are independent then intra-cohort variance in susceptibility does not translate into extra-Poisson variance in the disease counts. Any over-dispersion (variance greater than predicted by the Poisson distribution) can be attributed to calendar year effects. The following model is designed to quantify such over dispersion. Let d_{ij} denote the cancer counts in individuals of age i at calendar year j. Let n_{ij} denote the population at that age and in that year. d_{ij} is modelled as a Poisson variable with mean $\nu_{ij} \cdot n_{ij}$.

ν_{ij} is the hazard for a given individual from the population at age i and calendar year j:

$$\nu_{ij} = \varepsilon_j \cdot h(i),$$

where ε_j is a multiplicative term for calendar year j that acts on the baseline hazard, $h(i)$.

If ε_j is drawn from a lognormal distribution with mean fixed at 1, variance σ_E^2 and

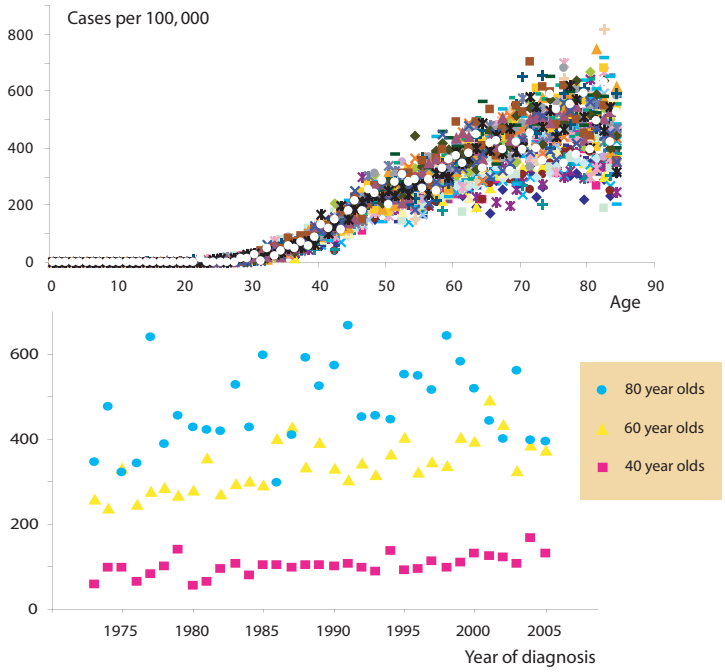

Figure 5.8: Top: age-specific female breast cancer incidence in Connecticut. Plotted for the years 1973 - 2003, giving a rough picture of dispersion in the disease counts at each age. Bottom: a rising temporal trend for 40, 60 and 80 year olds can be seen following the initiation of mammography screening in the early 1980s [AJD06]. This trend contributes to the count dispersion.

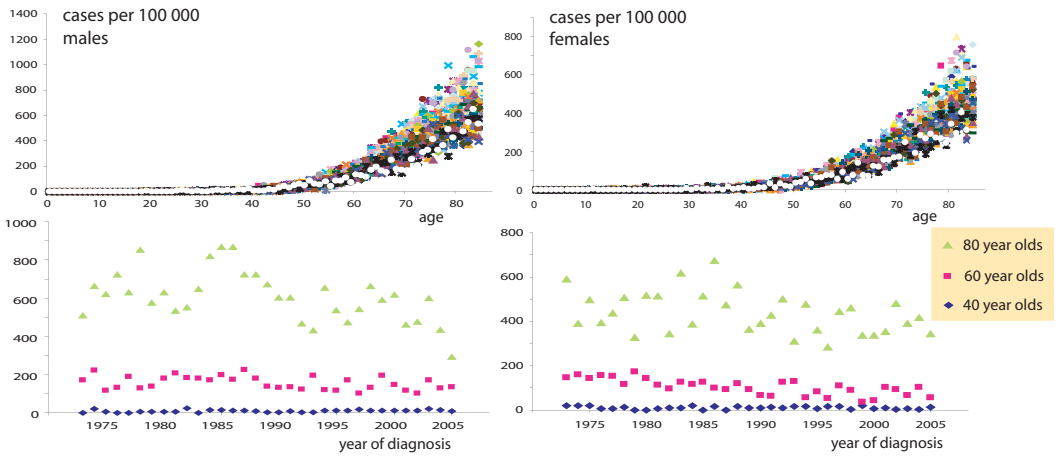

Figure 5.9: Top: colorectal cancer incidence in Connecticut for the years 1973 - 2003. Bottom: downwards temporal trends in incidence for 40, 60 and 80 year olds are significant from the 80s onwards. Increased use of sigmoidoscopy and fecal occult blood tests (triggering colonoscopy) beginning in the 70s seems to have precipitated the early detection and removal of precancerous legions (e.g. adenomas) eventually impacting on incidence in the following decade [CTC$^+$94]. Lifestyle changes may also play a role in the continuing steady reduction in colorectal cancer incidence.

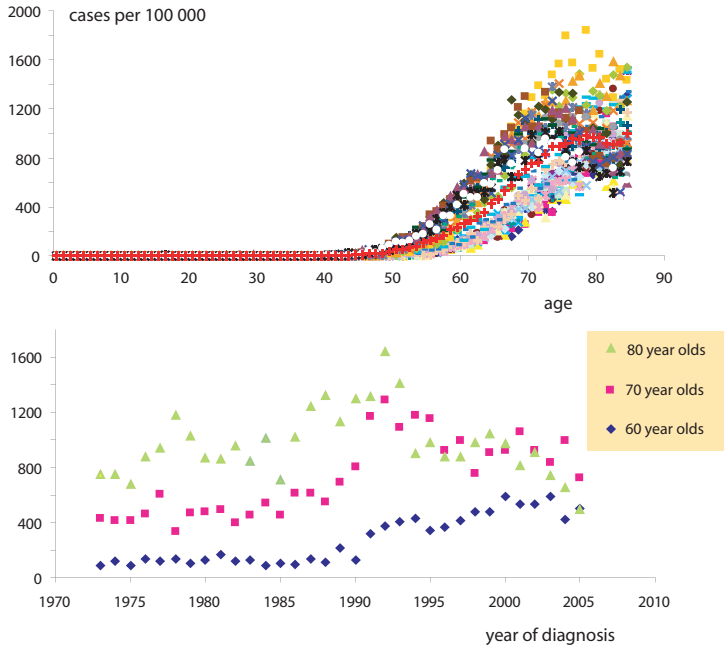

Figure 5.10: Top: age-specific prostate cancer incidence in Connecticut (1973 - 2003). Bottom: strong temporal incidence trends following the introduction of prostate specific antigen (PSA) screening in the late 1980s [KFFM00]. PSA testing is highly sensitive. Its use has meant cancers are registered at earlier stage / age and has also led to the detection of some cases that would never have become clinically apparent over the lifetime of the patient in the absence of PSA testing.

density function $f_E(\varepsilon)$, then the likelihood of the count d_{ij} is:

$$P[d_{ij}|\sigma_E^2, h(i)] = \int_0^\infty f_E(\varepsilon)\frac{(\varepsilon \cdot h(i) \cdot n_{ij})^{d_{ij}}}{d_{ij}!} \cdot \exp^{-\varepsilon \cdot h(i) \cdot n_{ij}}\, d\varepsilon. \tag{5.9}$$

Under the simplifying assumption that the calendar year coefficients are independent, the likelihood, $L(\underline{d}|\sigma_E^2, h(i))$, of a sequence of counts, $\underline{d}$, among birth cohorts of the same age, with $\underline{d} = (d_{i1}, d_{i2}, \ldots, d_{in})$ is:

$$L(\underline{d}|\sigma_E^2, h(i)) = \prod_{j=1}^{n} P[d_{ij}|\sigma_E^2, h(i)],$$

with $P[d_{ij}|\sigma_E^2, h(i)]$ given by equation (5.9).

The posterior density for σ_E^2 is:

$$P[\sigma_E^2|\underline{d}] = \frac{\int L(\underline{d}|\sigma_E^2, h) \cdot \pi_{\Sigma_E^2}(\sigma_E^2) \cdot \pi_H(h)dh}{\int\int L(\underline{d}|s, h) \cdot \pi_{\Sigma_E^2}(s) \cdot \pi_H(h)dsdh},$$

where π_H and $\pi_{\Sigma_E^2}$ are uniform prior densities for the hazard and variance terms respectively.

Figure 5.11 shows the marginal posterior, $P[\sigma_E^2|\underline{d}]$, for breast, colorectal and prostate cancer data recorded in Connecticut. These data show a small but not insignificant extra-Poisson variance caused by calendar year effects in breast and colorectal cancer, with modal variances less than 0.03. Prostate cancer shows stronger temporal incidence trends. Modal variances at different age groups of up to 0.6 are suggested for prostate cancer. Figure 5.12 shows lognormal distributions with mean one and variances ranging from 0.01 to 0.64. These distributions give an idea of the volatility in mean hazard with time described by various lognormal variances. However, only a fraction of this volatility is due to bona fide changes in susceptibility (i.e. changes in genetic and environmental influences on cancer risk). Much of the fluctuation in incidence is controlled by changing patterns of medical interventions and screening. The estimates of variance in mean hazard can be viewed as conservative upper bounds for the effects of changing susceptibility patterns with time.

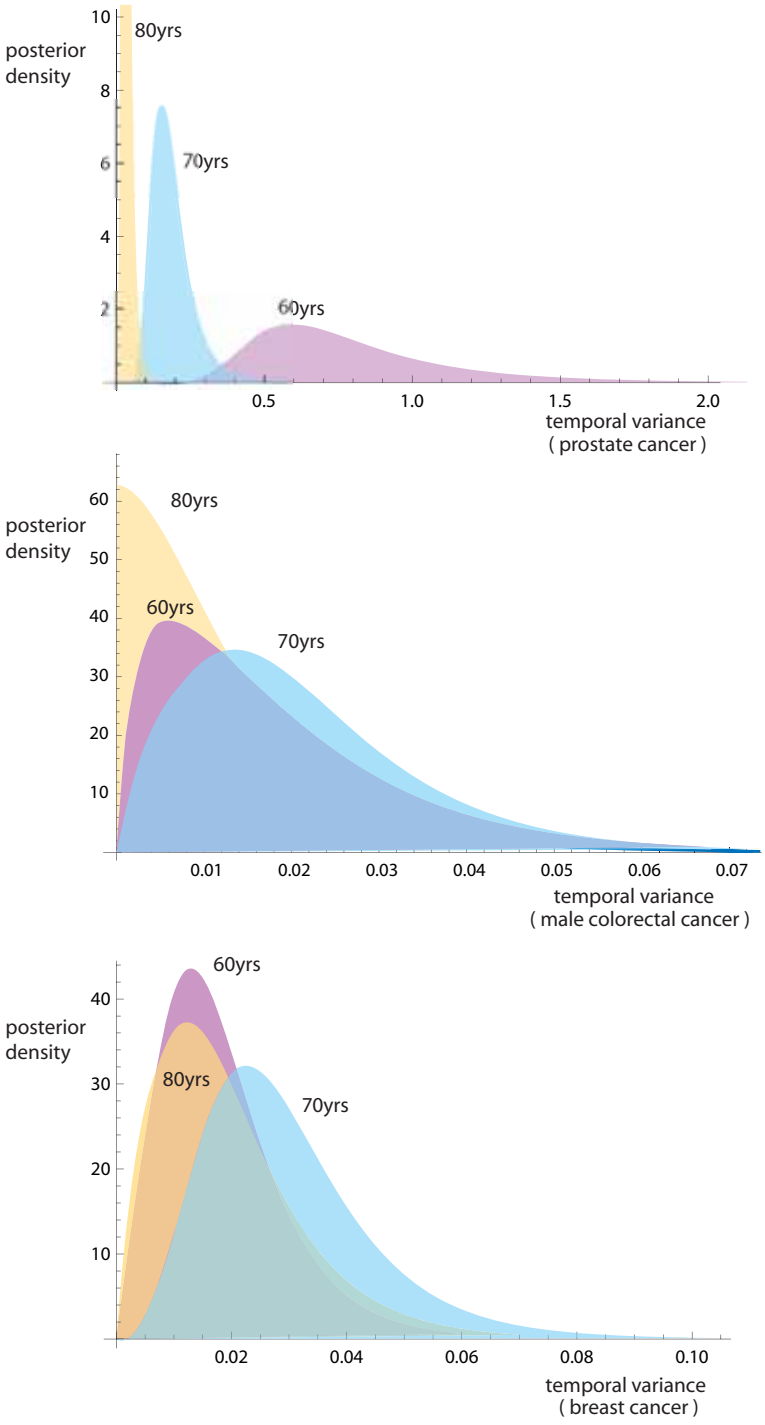

Figure 5.11: Posterior densities for Σ_E^2 in the case of prostate, breast and colorectal cancer.

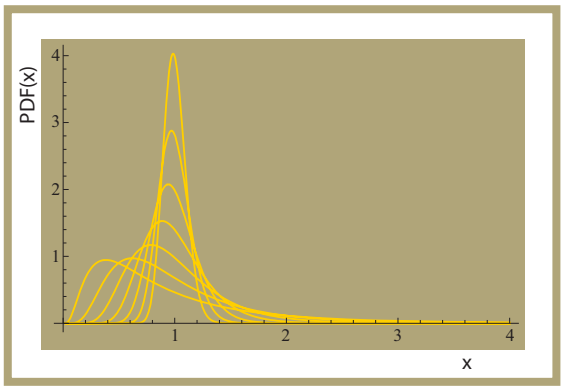

Figure 5.12: Lognormal distribution with mean set to one, and variance given by $\frac{2^n}{100}$ for $n = 0, 1, \ldots, 6$.

5.5 Estimating total population variance

In the MFT model used to determine environmental and genetic contributions to cancer susceptibility via twin data (discussed in section 5.3), liability is treated as a latent variable and as such its total variance cannot be estimated [Els81]. Locatelli et al. in an alternative analysis of breast cancer among twins, treated liability explicitly within a proportional hazards framework [LRLY07] and modelled age at cancer onset in a cohort of patients, rather than treating cancer as a binary trait. This model enabled an estimate of total liability variance as well as the absolute sizes of the genetic and environmental contributions to this variance. The hazard function for an individual was defined as:

$$h(t, l) = l \cdot h(t),$$

with liability, l, lognormally distributed and the baseline-hazard, $h(t)$, parameterised as a Gompertz curve [YVI95]. Log-normality for quantitative traits has been argued for elsewhere [LSA01] and arises naturally, for example, in the polygenic model of genetic risk described above [PAB$^+$02]. Locatelli et al. took, (l_1, l_2), the liability of a twin pair, to have a bivariate log-normal distribution (see appendix B) with mean equal to $(1, 1)$ and covariance matrix given by:

$$\begin{pmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{pmatrix}.$$

For monozygotic twins, $\rho = \rho_M$ and for dizogotic twins, $\rho = \rho_D$. ρ_M and ρ_D were estimated along with σ^2 and the Gompertz parameters. The estimated total population variance was $\sigma^2 \approx 45$.

The population lifetime-risk distribution implied by this variance can be derived. If $f_L(l)$ represents the PDF of the log-normal hazard ratio (with mean 1 and variance $\sigma^2 = 45$) then the density, $f_R(r)$, at a given lifetime risk, r is:

$$f_R(r) = f_L[w(r)] \cdot \frac{dw}{dr},$$

where w is the inverse of $g(l,t)$ with t fixed. g maps from hazard ratio for an individual, l, to lifetime risk, r (r is calculated to age $t = T_{\text{life}}$) . g is given by:

$$g(l,t) = 1 - \exp[-l \cdot \int_0^t h(s)ds]. \tag{5.10}$$

Assuming a Gompertz hazard, $h(a,b,t) = a \cdot \exp[b \cdot t]$, $g(l, T_{\text{life}})$ becomes:

$$g(l, T_{\text{life}}) = 1 - \exp[-l \cdot \frac{a}{b}(\exp[b \cdot T_{\text{life}}] - 1)],$$

and

$$w(r) = \frac{-b \cdot \ln(1-r)}{a(\exp[b \cdot T_{\text{life}}] - 1)}$$

and further

$$\frac{dw}{dr} = \frac{b}{a(\exp[bT_{\text{life}}] - 1)(1-r)}.$$

$f_R(r)$ is plotted in figure 5.13(a) using the Gompertz parameters inferred by Locatelli et al. Because the mean hazard ratio is constrained to 1, a large variance translates into a highly skew distribution with mode very close to zero. Another consequence

of a large population variance in hazard, is that while the expected hazard among newborns is equal to the baseline hazard, the theoretical population incidence deviates from the baseline hazard with age. This is because higher susceptibility individuals have higher mortality and so comprise a smaller proportion of the population at older ages. To quantify this effect, consider the expected population hazard, h_{pop}:

$$h_{\text{pop}}(t) = \int_0^\infty f_L(l,t) l \cdot h(t) dl \tag{5.11}$$

where $h(t)$ is the baseline hazard and $f_L(l,t)$ is the density function for the hazard ratio at age t and $f_L(l,0) = f_L(l)$. To calculate $f_L(l,t)$ requires an assumption concerning the mortality impact associated with a particular hazard ratio, l. Suppose the age related mortality rate from the cancer in question is $m_c(t)$ and that total mortality, $m(t)$, is $m_0(t) + m_c(t)$. Here $m_0(t)$ represents mortality from all causes other than the cancer in question. Assuming for simplicity that $m_c(t) = \alpha \cdot h_c(t)$ where $h_c(t)$ is the cancer incidence function, it follows that for an individual with hazard ratio l, the mortality is given by $m(l,t) = m_0(t) + l \cdot \alpha \cdot h(t)$. Hence, the survival to age t of such an individual, $S(l,t)$, is:

$$S(l,t) = \exp\left(- \int_0^t m_0(s) + l \cdot \alpha \cdot h(s) ds \right).$$

Now,

$$\begin{aligned} f_L(l,t) &= f_L(l) \cdot \frac{S(l,t)}{\int_0^\infty f_L(l) S(l,t) dl} \\ &= f_L(l) \cdot \frac{\exp\left(-l \cdot \alpha \cdot \int_0^t h(s) ds \right)}{\int_0^\infty f_L(l) \exp\left(-l \cdot \alpha \cdot \int_0^t h(s) ds \right) dl}. \end{aligned} \tag{5.12}$$

Equations (5.12) and (5.11) together give the population hazard, $h_{\text{pop}}(t)$. $h_{\text{pop}}(t)$ is plotted in figure 5.13 (b) with $\alpha = 1$ and assuming two different values of σ^2 in the log-normal distribution of hazard ratio. The baseline Gompertz hazard, $h(a,b,t)$, is shown alongside. There is a trend for increasing divergence from the baseline hazard with

increasing population variance in hazard ratio. A variance of 45 implies that inferences made from the observed population hazard are likely to be inaccurate reflections of the hazard function in individuals.

Another implication of such a large variance, is that approximately 80% of cases would be expected to occur among the top 20% of the population stratified by hazard ratio. To see this, note that for a given hazard ratio, l, the proportion of the population with hazard ratio greater than l is $1-\int_0^l f_L(s)ds$. The proportion of affected individuals with hazard ratio above this amount is given by:

$$1 - \frac{\int_0^l g(s,t) f_L(s) ds}{\int_0^\infty g(s,t) f_L(s) ds}.$$

Figure 5.13(c) plots $\overline{r}(l)$ for $l = 0$ to $l >> 1$, where

$$\overline{r}(l) = \left(1 - \int_0^l f_L(s)ds, 1 - \frac{\int_0^l g(s,t) f_L(s) ds}{\int_0^\infty g(s,t) f_L(s) ds} \right).$$

The trend with increasing σ^2 is shown.

But is $\sigma^2 = 45$ realistic on closer scrutiny? Can such a large subset of cases really be concentrated in such a small subset of the population? One potential problem with Locatelli et al.'s study, is their choice of parameterisation for the hazard function. A Gompertz hazard, $h(a, b, t)$, has a linear LLA with age, given simply by $b \cdot t$. Breast cancer incidence however, as discussed in section 2.5.1, has an approximate step function LLA. So, some of the variance in liability estimated by Locatelli et al. is actually likely to arise from poor representation of the hazard shape. Figure 5.14 shows the implied population incidence from Locatelli et al. against Swedish population incidence from Doll [PWF+02]. The discrepancy in the hazards is apparent.

5.6 Using incidence of second primary cancers to estimate liability variance

Pharoah et al. presented a more robust method of estimating σ^2 from a lognormal model of genetic liability variance [PAB+02] based only on observed twin relative risk

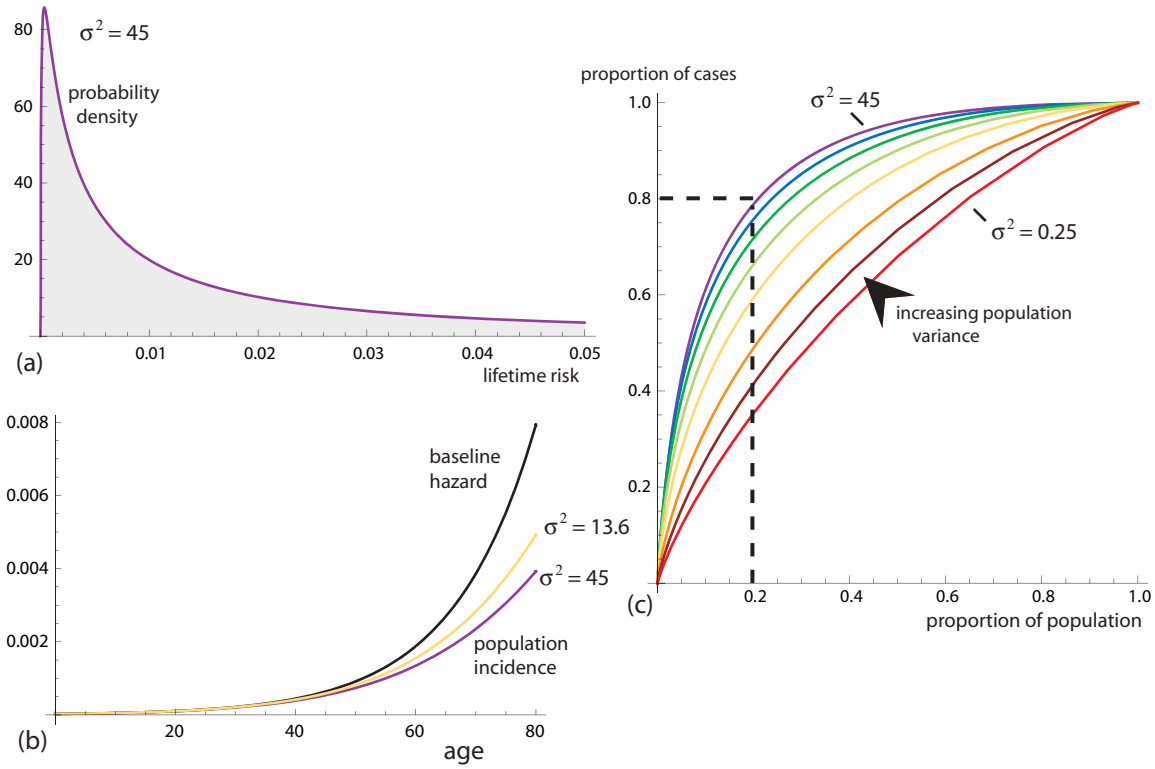

Figure 5.13: (a) Distribution of lifetime risk implied by the study of Locatelli et al. (b) Increasing liability variance causes a decoupling of expected population incidence and the baseline hazard rate. (c) A larger proportion of cases arise in a smaller minority of the population as the lognormal variance is increased. Under a variance of 45, 80% of cases occur in the 20% of the population at highest risk.

or sibling relative risk and not individual patient age-at-onset data. The technique rests on the idea that relative risk in siblings of affected patients is a function of the population variance in liability owing to genetic factors. To calculate the relative probability, λ_d, that the sibling of an affected patient is also affected, genetic liability in siblings is assumed to be correlated according to a theoretical value. Pharoah et al. took correlation on the log-scale to be $\frac{1}{2}$, so the distribution of (l_1, l_2), where l_1 and l_2 are the liabilities of each member of the sibling pair, is bivariate lognormal with covariance matrix given by:

$$\begin{pmatrix} \sigma^2 & (1+\sigma^2)^{1/2} - 1 \\ (1+\sigma^2)^{1/2} - 1 & \sigma^2 \end{pmatrix}.$$

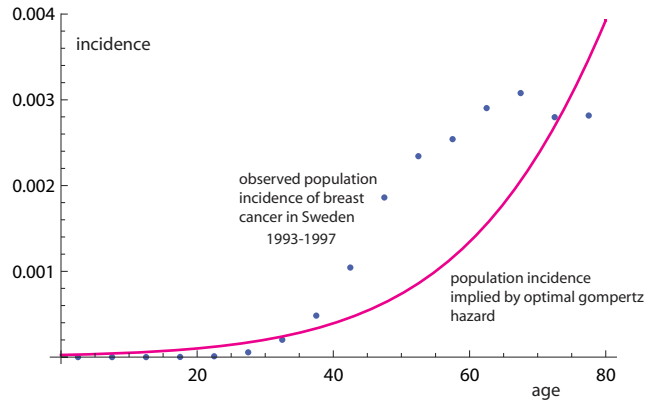

Figure 5.14: Locatelli et al. [LRLY07] used a Gompertz hazard to model age of breast cancer onset in Swedish twins born between 1886 and 1967. A Gompertz hazard is a questionable model for Breast cancer incidence.

Assuming simply that risk of cancer is proportional to liability (through the constant α say), Pharoah et al. calculated λ_d:

$$\begin{aligned}\lambda_d &= \frac{\int_0^\infty \int_0^\infty f_L(l_1,l_2)(\alpha\cdot l_1)(\alpha\cdot l_2)dl_1dl_2}{\left(\int_0^\infty f_L(l)(\alpha\cdot l)dl\right)^2}\\ &= \int_0^\infty \int_0^\infty f_L(l_1,l_2)(l_1\cdot l_2)dl_1dl_2, \quad \text{since} \quad \mathbf{E}(L)=1.\\ &= \frac{1+\sigma^2}{2}. \end{aligned} \tag{5.13}$$

Taking $\lambda_d = 2$ [PDD$^+$97], equation (5.13) is solved with $\sigma^2 = 3$. However, if rather than assuming risk of cancer is proportional to liability, risk of cancer is modelled in the relative hazards context, then λ_d can alternatively be expressed as:

$$\lambda_d = \frac{\int_0^\infty \int_0^\infty f_L(l_1,l_2)P_{l_1}[T<t_{\text{life}}]P_{l_2}[T<t_{\text{life}}]dl_1dl_2}{\left(\int_0^\infty f_L(l)P_l[T<t_{\text{life}}]dl\right)^2}. \tag{5.14}$$

where, $P_l[T\le t] = 1-e^{-\int_0^t l\cdot h(s)ds}$. and $h(t)$ is given a suitable parametrization.

Solving equation (5.14) with $h(t) = ae^{b\cdot t}$ (where (a,b) is chosen so that the hazard roughly fits breast cancer incidence data) and λ_d=2 gives $\sigma^2 = 9.6$. A much higher

estimate than the $\sigma^2 = 3$ predicted by Pharoah et al. For calculating lifetime risks, the relative hazards framework is more precise than the risk proportional to liability assumption, but the greater difficulty is in determining an appropriate theoretical relationship between the liabilities of siblings or twins. If the correlation of the actual liabilities of dizygotes, rather than their logarithms is taken to be one half so that the covariance matrix for the bivariate lognormal distribution of (l_1, l_2) becomes:

$$\begin{pmatrix} \sigma^2 & \sigma^2/2 \\ \sigma^2/2 & \sigma^2 \end{pmatrix}.$$

then equation (5.14) solves to give $\sigma^2 = 3.73$.

Even in the case of monozygotic twins, although the genetic component of liability can confidently be set equal between a twin pair, the extent of their correlation in environmental liability is undetermined.

To avoid modelling relatedness in the liabilities of relatives, total liability variance can be estimated by looking at second primary risk in individual patients rather than recurrence risk in their relatives. Assuming variable liability within the population of interest the risk of a second cancer in patients with an initial primary malignancy is higher than the risk of first cancer in patients of the same age. This is simply because the expected liability of a cancer patient is higher than for a healthy individual. The recurrence risk for a cancer patient depends on the age at diagnosis of the first primary. If, for example, the first primary is diagnosed before age s, then the smaller s is, the higher the patients expected hazard after that age. Further the larger the population variance in liability the greater the relative risk in patients compared with unaffected individuals. To see this, consider a random process, $x(t)$, that counts the number of primaries a patient has accumulated by age t. As before let $f_L(l)$ represent the PDF of the log-normal density for the hazard ratio. The probability of second primary before age t given one primary before age s, with $t > s$, is given by:

$$P[x(t) \geqslant 2 | x(s) = 1] = \int_0^\infty f_{L|x(s)=1}(l) \left(1 - e^{-\int_s^t l \cdot h(x) dx} \right) dl, \qquad (5.15)$$

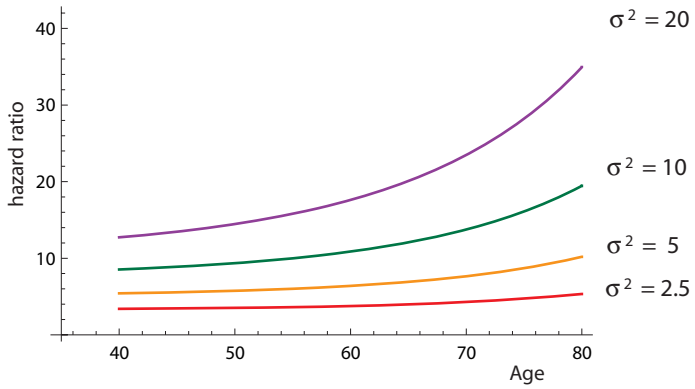

Figure 5.15: Relative hazard for second primary cancer given one primary diagnosis accumulated by age 40.

where

$$f_{L|x(s)=1}(l) = \frac{f_L(l)\left(\int_0^s l \cdot h(x)dx\right) e^{-\int_0^s l \cdot h(x)dx}}{\int_0^\infty f_L(y)\left(\int_0^s y \cdot h(x)dx\right) e^{-\int_0^s y \cdot h(x)dx} dy}. \tag{5.16}$$

Figure 5.15 shows the relative hazard derived from equation (5.15) evaluated for several values of σ^2. The upward trend in relative hazard with increasing variance can be exploited to estimate liability variance from observed risks of second primary cancers.

5.6.1 Incidence of second primary colon cancers

Hoar et al. published a large study of second cancers following initial cancer of the digestive system in Connecticut [HWB+85]. The results for 26 804 initial cases of colon cancer diagnosed between 1935 and 1982 are shown in table 5.2. Average follow up per patient was 4.5 years. Expected numbers of cases were calculated by applying appropriate age specific and calendar year specific incidence rates to the person years at risk accumulated in the follow up period after each initial colon cancer diagnosis. Relative risk was then calculated as the ratio of observed to expected cases.

A model of the results of Hoar et al. can be used to estimate population variance in liability. Suppose the initial 26804 cases of colon cancer were distributed by age according to the population structure and colon cancer risk profile in Connecticut (taken

colon, male and female	years after first primary cancer diagnosis				totals
	<1 year	1-4years	5-9 years	10+years	
number starting interval	26804	18813	8071	3984	26804
person years in interval	18167	47696	28621	26825	121309
observed cases	48	184	120	154	506
expected cases	32.31	89.46	59.33	64.48	245.42
incidence	0.26%	0.39%	0.42%	0.57%	0.42%
RR	1.49	2.06	2.02	2.39	2.06

Table 5.2: Expected verses observed incidence of second primary colon cancer in Connecticut. Data from Hoar et al. Relative risk is calculated as the ratio of observed to expected cases. Incidence is the ratio of observed cases to person years in interval.

from the SEER database for a calendar year within the study period). i.e. that the probability a given case falls into the ith age interval (where the age intervals are the typical 5 year bins 0-4, 4-9, etc..) is given simply by the total number of cases in this interval for a given year within the study period, divided by the total number of cases in that calendar year. Suppose that incidence is modeled by a Weibull hazard, with relative hazard lognormally distributed in the population, so that:

$$h(t) = l \cdot at^b, \tag{5.17}$$

where l is a lognormal variable with variance σ^2 and mean 1. For a given value of σ^2, equation (5.17) can be fit to incidence data from the Connecticut registry using the standard likelihood function (equation (2.23)). Then the ratio of observed to expected cases, assuming n years of follow up, predicted by the model can be compared with that published by Hoar et al. The relative risk predicted by equation (5.17) is:

$$\frac{Ob(\hat{a}_{\sigma^2}, \hat{b}_{\sigma^2}, \sigma^2)}{Ex(\hat{a}_{\sigma^2}, \hat{b}_{\sigma^2}, \sigma^2)} \tag{5.18}$$

where $\hat{a}_{\sigma^2}$ and $\hat{b}_{\sigma^2}$ are the maximum likelihood estimates of the Weibull parameters a and b (equation (5.17)) with liability variance fixed at σ^2. The number of cases, $Ex(\hat{a}_{\sigma^2}, \hat{b}_{\sigma^2}, \sigma^2)$, expected per patient when applying the hazard described by equation (5.17) to n years of follow up after each initial diagnosis is:

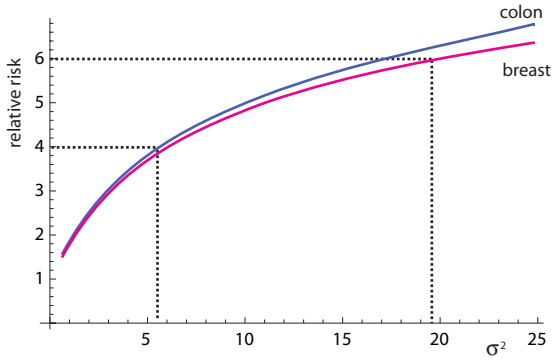

Figure 5.16: Theoretical relationship between relative risk and liability variance based on the studies of Hoar et al. and Harvey and Brinton.

$$Ex(\hat{a}_{\sigma^2}, \hat{b}_{\sigma^2}, \sigma^2) = \sum_i P[T_1 \in [5(i-1), 5i)] \cdot \int_0^\infty \left(1 - \exp[-l \cdot \int_{5i-2.5}^{5i-2.5+n} \hat{a}_{\sigma^2} t^{\hat{b}_{\sigma^2}} dt]\right) \cdot f_L(l, \sigma^2) dl,$$

where $f_L(l, \sigma^2)$ is the lognormal density with mean 1 and variance σ^2. T_1 is the time at first cancer, and $P[T_1 \in [5(i-1), 5i)]$ is estimated by the ratio of the number of cases over the period falling in the age bracket $[5(i-1), 5i)$ in Connecticut, divided by the total number of initial colon cancer cases recorded over the period.

The observed number of cases expected per patient under equation (5.17), is:

$$Ob(\hat{a}_{\sigma^2}, \hat{b}_{\sigma^2}, \sigma^2) = \sum_i P[T_1 \in [5(i-1), 5i)] \cdot \int_0^\infty \left(1 - \exp[-l \cdot \int_{5i-2.5}^{5i-2.5+n} \hat{a}_{\sigma^2} t^{\hat{b}_{\sigma^2}} dt]\right) \cdot f_{L|T_1 \in [5i-3, 5i-2)}(l, \sigma^2) dl,$$

where $f_{L|T_1 \in [5i-3, 5i-2)}(l, \sigma^2)$ is defined by analogy to equation (5.16).

Hoar et al. calculate a relative risk of 2.06, however they excluded cases occurring less than 2 months following the initial diagnosis or multiple cases diagnosed simultaneously. Including these cases leads to a relative risk estimate of 4.2. Figure 5.16 shows (5.18) evaluated for various values of σ^2 with the follow up period n set to 4.5 years. A relative risk of 4 implies a variance of $\backsim 6$.

5.6.2 Incidence of second primary breast cancers

Harvey and Brinton published a large study of second cancers following initial cancer of the female breast in Connecticut [HB85]. The results for 41109 initial cases of breast cancer diagnosed between 1935 and 1982 are shown in table 5.3. Average follow up per patient in this study was 6.6 years.

breast, female	years since first primary breast cancer				total
	less than 1 year	1-4years	5-9 years	10+years	
number starting interval	41109	36068	18609	9306	41109
person years in interval	32043	103274	66162	70046	271524
observed cases	241	754	467	465	1927
expected cases	66.81	225.87	155.66	188.51	636.46
incidence	0.75%	0.73%	0.71%	0.66%	0.71%
relative risk	3.61	3.34	3.00	2.47	3.03

Table 5.3: Expected verses observed incidence of second primary breast cancer in Connecticut. Data from Harvey and Brinton. Relative risk is calculated as the ratio of observed to expected cases. Incidence is the ratio of observed cases to person years in interval.

As above, in the case of colon cancer, the age distribution of the initial 41109 breast cases recorded in Harvey and Brinton [HB85] can be inferred from the Connecticut registry statistics. A simple model of the age related breast cancer hazard can then be used to estimate population variance in liability. Suppose that incidence is modelled by a piecewise linear hazard with relative hazard lognormally distributed in the population, so that:

$$h(t) = \begin{cases} 0 & t \le a \\ l \cdot b(t-a) & t > a \end{cases}, \tag{5.19}$$

with l a mean-one lognormal variable with variance σ^2. Then equation (5.18) gives the relative risk predicted by equation (5.19).

Harvey and Brinton calculate a relative risk of 3.03. However, the second primaries they observe are all in the contralateral breast, but they have used population rates (i.e. twice the hazard per breast) to calculate expected numbers of cases. Therefore their estimate of relative risk should be six rather than three. A relative risk of six implies a large population liability variance of roughly 20 (figure 5.16).

5.7 Discussion

In this chapter evidence for genetic and environmental influences on cancer susceptibility was discussed. A common argument was presented showing that known Mendelian cancer syndromes are insufficient to explain the doubling of risk in first degree relatives of cancer patients. A popular theory accounting for this discrepancy states that many common but undiscovered low penetrance alleles must act to confer the susceptibility patterns observed in relatives. The notion of many factors, environmental and genetic, interacting multiplicatively to influence cancer risk, leads naturally to the assumption of a lognormal distribution of susceptibility in populations [PAB$^+$02]. A lognormal distribution has been used previously to quantify variance in cancer risk by Pharoah et al. Their method was based on assumptions about twin pairs or siblings and the extent to which they share risk factors. It is difficult to argue convincingly for a definitive theoretical relationship between the genetic or environmental components of liability in twins or siblings. The problem can be avoided, however, via the novel method outlined above which focuses on recurrence risk in individual cancer patients rather than in the relatives of these patients. Liability variances calculated from recurrence data in this manner suggest large variation in cancer susceptibility within human populations. A lognormal variance value of six for colon cancer implies that 80% of colon cancers occur in the 30% of the population which is at highest risk (figure 5.17) . Similarly, for breast cancer, with a lognormal variance of 20, the relative hazard model predicts that 80% of cases occur in the 25% of the population at highest risk. Intra-cohort variance on this scale can be contrast with temporal variation in susceptibility which is modest by comparison. Modifications to the risk of being diagnosed with cancer show a variance of less than 0.03 for colon and breast cancer over the last 30 years. Further, this variance seems to be dominated by changes in screening and medical interventions rather than bona fide changes in susceptibility. A pattern of susceptibility, relatively stable with calendar time, but showing strong differences between individuals underlines the importance of continuing efforts to pinpoint genetic and environmental risk factors.

A more theoretical implication of wide intra-cohort variance in liability is that it forces age-specific risk in an individual to be different from that observed in the population (figure 5.13) (b). The exact relationship between individual hazards and pop-

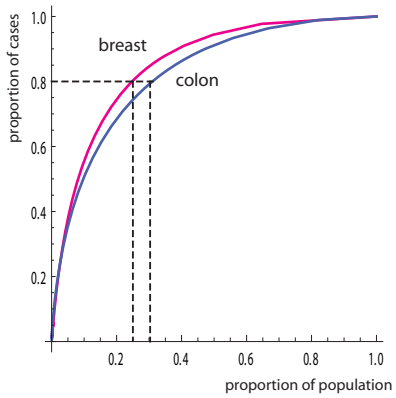

Figure 5.17: Relationship between fraction of the population at highest risk and the fraction of cases occurring in that high risk subset for breast and colon cancer - estimated from the data of Harvey and Brinton and Hoar et al. respectively.

ulation incidence is dictated by the mortality impact of the cancer in question and by the manner in which risk varies between people. It is likely that a simple proportional hazards model, with hazards changing by only a constant factor, does not capture the risk variation between individuals adequately. Evidence for a less uniform variation in hazard function is provided by the breast cancer recurrence data of Harvey and Brinton [HB85]. Risk of contralateral breast cancer following an initial primary appears to be independent of the time since the initial primary was diagnosed (figure 5.18). This is at odds with the proportional hazards description. If the population incidence is to be monotonic increasing, then under the proportional hazards assumption, every patient must have a hazard that increases monotonically with time, regardless of their age at initial primary onset. An explanation put forward to explain the constant risk of cancer in the opposite breast after initial breast cancer diagnosis is that predisposition may be mediated through some of the multiple steps leading to cancer but not others [Fra04c]. So, perhaps breast cancer patients are predisposed to progress quickly through all except one stage in breast cancer development. Then, by the time of a patients first breast cancer diagnosis, he or she will have many lineages that have already passed through all except the final stage of tumorigenesis. Their risk of cancer will hence be constant with time. The finding that monozygotic twins, mothers and sisters of breast cancer patients all also have a high and roughly constant incidence of breast cancer after they attain the

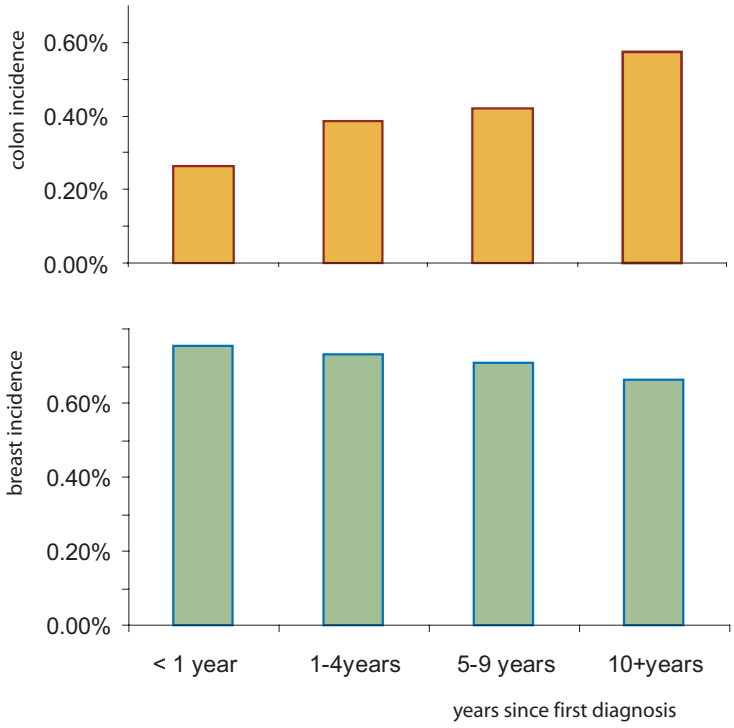

Figure 5.18: Colon cancer shows a rising risk of second primary with time since initial diagnosis, as predicted by the lognormal relative hazards model. Breast cancer, by contrast, shows a stable incidence with time since initial diagnosis.

index patients age at diagnosis supports this view [PM00]. The relationship between risk in individuals and population incidence is an interesting avenue for further study. The above argument for breast cancer could be quantified with a simple multistage model in which some rates of mutation or transition between stages vary widely in the population but others do not. In general, data on second primary incidence is a useful source of information on the hazard in high-risk individuals and multistage theory is likely to prove useful in generating hypotheses concerning the connection between the risk profiles of high susceptibility patients and the population incidence.

Chapter 6

Discussion

This thesis began with a sceptical review of the literature on multistage-modelling of cancer incidence. Multistage models are often referred to as 'quasi-mechanistic'. This phrase can be seen as an acknowledgement of the speculative nature of multistage theory and / or an acknowledgement of the overly simplistic picture of cancer which the theory describes. Multistage models are primarily stochastic time-to-event-models. They attempt to depict the distribution of the time until cancer in patients and hence the population incidence of cancer and its dependence on patient age. There are many unknowns in the development of cancer, a complex process involving multiple cell lineages interacting with each other and with the tissue micro-environment. In chapter two, the methods by which two prominent 'quasi-mechanistic' multistage models (Armitage and Doll's model and the two stage clonal expansion (TSCE) model) reduce the development of a tumour to a mathematically tractable abstraction were presented. A survey of some of the applications of these theories was then given, leading to the suggestion that multistage models may be more suitable in many situations for hypothesis generation rather than for hypothesis testing.

In Armitage and Doll's model, the time until cancer in a patient is seen as the end result of a sequence of mutations afflicting any of the cell lineages in a given tissue. Armitage and Doll viewed population incidence as an exact mirror of the hazard in an individual, a position that arises naturally from the assumption that the risk in every individual is the same. They reasoned that the risk of tumour in a patient, which rises as an integer power of age, should translate directly into a population incidence which rises with the same power of age. In the two stage clonal expansion (TSCE) model, an initial mutation causes a lineage to divide and proliferate, creating copies of

itself, a process referred to as 'clonal expansion'. A second mutation in one of these lineages then creates a malignancy. The authors of TSCE also assumed a homogeneous population and saw population incidence as a mirror of individual hazards. Clonal expansion is thought to be an important mechanism of tumorigenesis, increasing the number of dysfunctional target cell lineages that can be further transformed by the next in a sequence of mutations. Patterns of clonal expansion are very difficult to observe however. Comparing TSCE and Armitage and Doll's model it can be seen that both achieve a theoretical population incidence that rises with age but by different means. Increasing risk with age in Armitage and Doll's model results solely from the assumption of many mutations. In TSCE, increasing risk with age is dictated as much by clonal expansion as mutation numbers. So TSCE and Armitage and Doll's model embody two of the main themes around which explanations for the central observation of age-specific risk, that it rises sharply with age, have been based.

In chapter three a simple statistical exercise was used to show that the number of mutations implied by a multistage model depends upon the assumed clonal expansion pattern in that model. This result shows that statistical inference of aetiological detail from incidence data cannot be considered reliable, unless uncertainty over clonal expansion patterns and other details concerning the development and detection of a tumor are accounted for.

The concept of a rate-limiting step is central to the interplay between mutation and clonal expansion. In chapter three, it was shown that a gene mutation, necessary for the development of a malignant tumor and able to target many different lineages within a large clone, may happen very quickly and not have an appreciable effect on the time taken for the tumor to emerge. Such a mutation should therefore not be considered rate-limiting. Working from an incidence profile, it is not possible to discern the existence or effects of non-rate-limiting mutations. Therefore, greater insight into cancer aetiology may be gained by modelling the effect on incidence that a specific and identified mutation has, rather than trying to model the combined effects of an unknown number of unidentified mutations. This can be achieved by comparing cancer risk in individuals with and without a germline mutation in a known susceptibility allele. The fact that such an allele confers susceptibility suggests that its inactivation is a rate-limiting event. Further, the observed change in incidence caused by the presence of a germline

mutation in the allele can be more confidently attributed to the effect of a specific gene product.

In chapter four, bowel cancer risks in patients with and without a germline *APC* mutation were compared. *APC* mutation is thought to initiate bowel tumorigenesis. This makes the relationship between risk in cases with or without a germline *APC* mutation quite simple. A method for estimating the rate of *APC* mutation was described which exploits this simple relationship and bypasses the need for many of the difficult assumptions concerning clonal expansion that are required for a more complete model of cancer development.

Bowel cancer risks in patients with and without mutations in the human miss-match-repair (MMR) machinery were also compared in chapter 4. The aetiological relationship between these cancers is less clear and differences in their genetic pathways have been shown to exist. Also, the point in the sporadic genetic pathway at which MMR inactivation occurs is not certain. In view of these complicating factors, a more general analysis of incidence modulation through germline mutation was presented to generate hypotheses on the aetiological relationship between bowel cancer with or without a germline MMR mutation. To facilitate this analysis Franks measure, ΔLLA, was employed. ΔLLA is a measure of the change in the log-log gradient of the incidence curve. Observed ΔLLA owing to an MMR mutation was compared with ΔLLA derived under various theoretical scenarios. It was found that a strongly rising ΔLLA with age, as observed following germline MMR mutation, can be created by a slowing of some aspects of tumorigenesis, i.e. through extra stages or slower mutations coincident with a quickening of other stages of tumorigenesis i.e. through increased mutation rate. Changes in clonal expansion and their effect on ΔLLA were found not to match MMR mutation data.

An interesting aspect of germline MMR mutations is that they give rise to a humped incidence pattern; one that rises and falls, peaking between forty and sixty years of age. This observation has been repeated in three separate studies of bowel cancer risk in patients with MMR mutations [JBD$^+$06, DFC$^+$97, QVvH05] and a similar phenomenon has been observed for breast cancer in patients with mutations of *BRCA1* or *BRCA2*. Multistage theory does not predict a peaked incidence pattern so this behaviour complicates attempts to model ΔLLA due to germline *BRCA1*, *BRAC2*

or MMR mutation. On the other hand it is an interesting point of departure for new hypotheses to explain the peak. It has been argued that the peaking incidence in old age, observed in some cancers, results from population heterogeneity [HJTMF$^+$00], that elderly populations are purged of high risk members and so the incidence among them is lower than at younger ages. Incidence peaks in high penetrance mutation carriers may represent a similar phenomenon but shifted to an earlier age. In chapter four, Franks discrete model of population heterogeneity was used [Fra07], based on the assumption that only a subset of MMR mutation carriers had a raised penetrance. This creates a humped incidence pattern, provided the penetrance in each of the subsets of the population are suitably different. Suppose incidence in low risk individuals is given by $h_l(t)$ and in high risk individuals the hazard is $h_h(t)$. If a fraction, f, of these individuals are at high risk then the penetrance observed in the population is:

$$P[T \leq t] = f \cdot \left(1 - \exp\left[-\int_0^t h_h(s)ds\right]\right) + (1-f) \cdot \left(1 - \exp\left[-\int_0^t h_l(s)ds\right]\right).$$

The population incidence is

$$h_p(t) = \frac{f \cdot h_h(t) \exp\left[-\int_0^t h_h(s)ds\right] + (1-f) \cdot h_l(t) \exp\left[-\int_0^t h_l(s)ds\right]}{1 - P[T \leq t]}.$$

Figure 6.1 shows $h_p(t)$ assuming 30% of the population have elevated risk $h_h(t) = a \cdot t^b$ and the rest have the background incidence $h_l(t) = c \cdot t^d$.

The notion of a discrete risk heterogeneity such as that depicted by figure 6.1 is unconventional, but it is possible that the effect of an MMR germline mutation is only realised in the presence of another common but unidentified genetic modifier. In any case the causes of the incidence patterns in *BRCA* mutation carriers and MMR mutation carriers warrant further study. In chapter 5, a contrasting type of risk heterogeneity was investigated. A continuous model of variation in hazard functions was used to estimate how concentrated the cancer burden may be in high risk patients. Departure of the population incidence, from that observed in individuals was not so strong under the continuous model of hazard variation. But the model did predict a flattening in population incidence, relative to the assumed log-log linear individual hazards at older ages. Such a flattening is consistent with observation [HJTMF$^+$00]. A related aspect

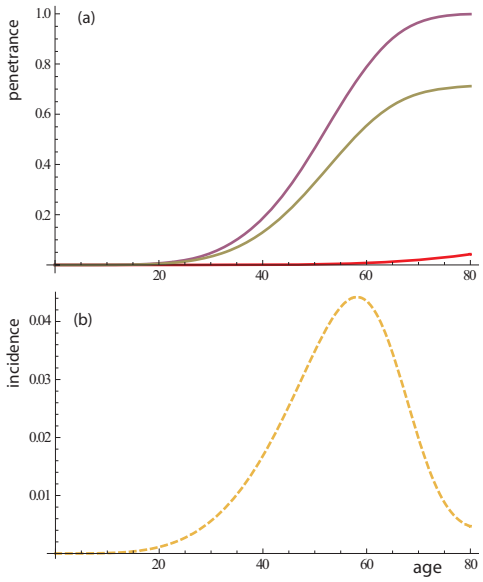

Figure 6.1: (a) Red line: background population penetrance of 5% ($c = 10^{-12}$, $d = 5$). Purple line: 30% of the population have complete penetrance by age 80 ($a = 10^{-8}$, $b = 4$. Green line: observed penetrance in the mixed population. (b) Incidence corresponding to the population penetrance.

of cancer biology is the potential for heterogeneity in the lineages which form a tissue. The models presented in this thesis have assumed that the target tissue contains a homogeneous population of target cell lineages each equally at risk of malignant transformation. However, in a tissue such as the colonic epithelium, exposure to carcinogens and other relevant risk factors may not be uniformly distributed. A non-uniform distribution of risk across lineages is suggested by the different spatial distributions within the bowel obeyed by different types of colorectal cancer.

As a simple, discrete theoretical model of this situation consider two subpopulations of cell lineages within a tissue with distinct risk profiles. The risk of cancer in the first population could be given by $h_1(t)$ and $h_2(t)$ could dictate cancer risk in the second population. The combined hazard is then simply $h_1(t) + h_2(t)$. Figure 6.2 shows how such a composite hazard could lead to a modulating acceleration pattern even when the two subpopulations have log-log-linear hazard.

Consider another situation where every lineage progresses along the same pathway via the same number of stages but where each lineage passes through these stages at a different rate. In this case the penetrance of disease in the tissue is dictated by a few

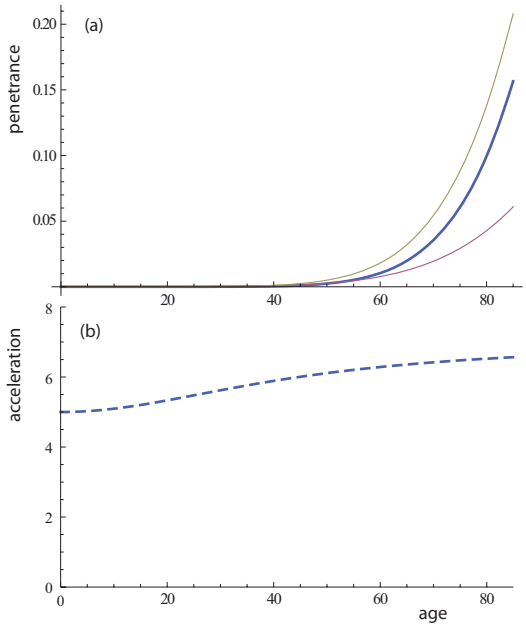

Figure 6.2: (a) Green line: higher penetrance in some lineages disposed to a pathway with greater acceleration of 7. Red line: lower penetrance in cells disposed to a pathway with a lower acceleration of 5. Blue line: the composite penetrance has a modulating acceleration which starts at 5 and rises to 7.

lineages with high mutation rates. Working with Armitage and Doll's formula and a truncated lognormal distribution of mutation rate (so that the maximum rate per annum is 0.01), the penetrance arising from a variable population of lineages can be expressed by:

$$\text{penetrance} = 1 - \left(1 - \frac{\int_0^{0.01} f_L(l) \left(1 - \exp[-l \cdot ut] \sum_{i=0}^{n-1} \frac{(l \cdot ut)^i}{i!} \right) dl}{\int_0^{0.01} f_L(l) dl} \right)^N , \quad (6.1)$$

where $f_L(l)$ is the lognormal PDF for the factor l which multiplies the mutation rate, u is the mutation rate, n is the number of stages and N is the number of cell lineages. This expression can be used to show that the mutation rate implied by an observed penetrance depends on the variance in mutability of the various cell lineages in the tissue of interest. The wider the variance, the lower the mean mutation rate per lineage implied by the same risk profile. Figure 6.3 shows penetrance calculated from equation (6.1) with $f_L(l)$ a lognormal PDF with mean and variance both equal to one and

with $u = 2 \times 10^{-4}$, $n = 6$ and $N = 10^8$. The penetrance is equivalent to that arising from a homogeneous population of cell lineages with common mutation rate 10^{-3} per annum.

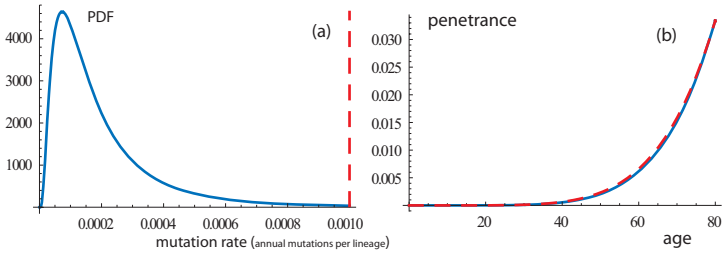

Figure 6.3: (b) Penetrance arising from two different distributions of mutation rate in target lineages within a tissue. The dashed red line is calculated assuming every lineage has a mutation rate of 0.001 per annum. The blue line is calculated assuming a lognormal distribution of mutation rate as shown in (a).

Heterogeneity in risk between individuals can produce departures in observed population incidence from the hazard experienced by an individual. Such heterogeneity can be observed through recurrence data, revealing a tendency for cancer to target certain individuals or certain families. Risk variation within the lineages of a tissue is more difficult to measure, but is reflected to an extent by spatial clustering of cases affecting a particular epithelial surface for example. Overall variability in the aetiology of cancer between patients and also at different locations within the same organ requires further study and is an important component of the age-onset pattern.

The hope that quantitative analysis of tumour risk may help to resolve the complexities of cancer aetiology still seems justified. Success will require a greater understanding of clonal expansion patterns which control the relative effects of different mutations, determining, for example, those that are rate-limiting and those that are not. Mendelian cancer syndromes suggest germline variation at a rate-limiting locus and closer inspection of the incidence shift in a particular syndrome may reveal further aspects of the underlying allele's aetiological role. Finally it must be acknowledged that age-specific onset data are not an exact reflection of individual hazards but rather an aggregate of the various risks experienced by a cohort. Likewise, individual hazards are dominated by the lineages within a tissue that are most easily drawn into the ways

of malignancy. Multistage models may ultimately be essential not only to understanding how somatic DNA aberrations combine to produce a particular hazard curve in an individual, but also how environmental factors and germline variation in susceptibility alleles act to modify this hazard and produce the population incidence of cancer.

6.1 Directions for further work

When work on this thesis began the field of quasi-mechanistic cancer incidence modelling was dominated by the TSCE model in various forms. Implementations of TSCE were being used to predict the nature of all the rate limiting steps in bowel cancer [LM02], elucidate the mechanisms through which smoking causes cancer [HL01], decipher the point during tumorigenesis at which the genome may become unstable [LW03] and understand the interaction between radiation and carcinogenesis [KZH$^+$03]. As has been outlined above and in more detail in chapters two and three, it can be argued that many of these aims are too high for TSCE. However, this argument was almost never made in the literature and so the work of these chapters contributes significantly to formalizing a quantitative framework through which to judge multistage modelling [HPT07]. A key development documented in chapter three is a novel quantitative definition of 'rate-limiting', providing a measure of the observability of a carcinogenic event through age onset statistics. This definition and its properties, particularly its interaction with the size of the registry collecting the age-onset data, could benefit from further investigation however. Another key concept arising from the critique of multistage models is model uncertainty:- the problem of different but equally plausible models of tumorigenesis producing different conclusions when applied to the same question. This was formally demonstrated in chapter three through estimates of the number of mutations in a cancer. Statistical approaches for dealing with model uncertainty, although readily available, are seldom if ever applied to multistage modelling. Going forward, practitioners in this area may be discouraged from techniques such as Bayesian model averaging due to computational intractability; however progress in understanding the limitations of multistage models could equally be made by systematically testing models against simulated data. For example synthetic incidence data, generated from a known clonal expansion structure, could be used to test a given multistage models efficacy in delineating this structure etc..

Despite reservations expressed in chapters two and three, chapter four uses a multi-stage approach but in a novel way more recently advocated [Fra05] to address questions of cancer aetiology. The approach advocated is distinguished from the TSCE model approaches referenced above by a focus on situations that require fewer arbitrary biological assumptions. Repeating an earlier comment, it seems likely that greater insight into cancer aetiology may be gained by modelling the effect on incidence that a specific and identified mutation has, rather than trying to model the combined effects of an unknown number of unidentified mutations. This approach was vindicated in chapter four with a stable estimate of the rate of mutation of the APC gene [HPT08] via an original computational technique. There is a scarcity of data around in-vivo mutation rates as they are difficult to calculate and so the estimate in chapter four can be viewed as significant.

Comparing cancer risk in individuals with and without a germline mutation in a known susceptibility allele was originally made famous by Knudson in his celebrated study of retinoblastoma [Knu71]. A handful of other 'copy-cat' studies have appeared since [KS72a, KS72b, MYFS90] but in general attention in this type of comparative study waned after Knudson's original success. The approach has been championed again more recently by Frank [Fra07] who has developed new computational machinery to analyse the incidence patterns of hereditary cancer syndromes. Applying these techniques to HNPCC in chapter four showed for the first time that HNPCC has a rising ΔLLA with respect to sporadic MSI+ CRC, consistent with its role in replication fidelity. Despite some difficulties with data quality and some ambiguity associated with smoothing parameters etc.., more work in this area certainly seems justified and many hereditary syndromes exist as potential targets for these types of investigation. An observation arising from chapter four, which may prove to be important, is that incidence of HNPCC is biphasic. Franks hypothesis for the similar nature of BRCA1/2 breast cancer incidence is outlined above in figure 6.1. It would certainly be interesting to look for this pattern in other hereditary cancer syndromes and to speculate further on its origin.

The work done in chapter five is based on the idea of polygenic susceptibility to cancer. Previous quantitative work has shown how multiple common susceptibility loci lead naturally to a log-normal distribution of risk in a given population [PAB$^+$02].

However, methods for quantifying the variance of this distribution have typically been based on studies of risk in relatives of cancer patients. These methods are unreliable because they confuse environmental and genetic causes of familial clustering. The central contribution of chapter five is the development of a novel quantitative method for estimating the variance in risk arising from the combined influence of genes and environment. This method does not rely on twin or family data. Instead, data on the risk of second cancers to cancer patients themselves is used. The approach was successfully applied to bowel and breast cancer but further work could apply the same method to other cancers for which data on second primary tumours exist. An abundance of data of this kind has been collected by Flannery et al. [FBD$^+$85]. The value of quantifying variance in cancer risk is at least two fold. From a medical perspective, cancers which are shown to cluster in families or correlate strongly with environmental exposures can be mitigated through changes in lifestyle or benefit from targeted prophylactic care. Second, as described above, any variance in cancer risk impacts age-onset patterns and complicates the relationship between individual and population age-related risk profiles. This is also true for variation in cancer risk between cell lineages within a patient (see figure 6.2 and figure 6.3). For these reasons, continuing to estimate variance in susceptibilities is a particularly valid avenue for futher study.

Bibliography

[ACP+08] AC Antoniou, AP Cunningham, J Peto, DG Evans, F Lalloo, SA Narod, HA Risch, JE Eyfjord, JL Hopper, MC Southey, H Olsson, O Johannsson, A Borg, B Pasini, P Radice, and S Manoukian. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *Br J Cancer*, 98:1457–1466, 2008.

[AD54] P Armitage and R Doll. The age distribution of cancer and a multistage theory of carcinogenesis. *Br J Cancer*, 8:1–12, 1954.

[AGZ+05] D J Araten, D W Golde, R H Zhang, H T Thaler, L Gargiulo, R Notaro, and L Luzzatto. A Quantitative Measurement of the Human Somatic Mutation Rate. *Cancer Res*, 65:8111–8117, 2005.

[AJD06] W Anderson, I Jatoi, and S Devesa. Assessing the impact of screening mammography: breast cancer incidence and mortality rates in Connecticut (1943-2002). *Breast Cancer Research and Treatment*, 99:333–340, 2006.

[AM04] R P Araujo and D L S McElawin. A history of the study of solid tumour growth: the contributions of mathematical modeling. *Bull Math Biol*, 66:1039–1091, 2004.

[AMA+95] M Aarnio, J Mecklin, L A Aaltonen, M Nystrom-Lahti, and H J Jarvinen. Life-time risk of different cancers in hereditary non-polyposis colorectal cancer (HNPCC) syndrome. *Int J Cancer*, 64:430–433, 1995.

[and08] F Bunz and. *Principles of cancer genetics*. Springer, New York, 2008.

[Arm85] P Armitage. Multistage Models of Carcinogenesis. *Environ Health Perspect*, 63:195–201, 1985.

[Ash69] D J B Ashley. Colonic Cancer Arising in Polyposis Coli. *J Med Genet*, 6:376–378, 1969.

[ASK+98] L A Aaltonen, R Salvovaara, P Kristo, F Canzian, A Hemminki, P Peltomaki, R Chadwick, H Kaarianinen, M Eskelinen, H Jarvinen, J P Mecklin, and A de la Chapelle. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *N Engl J Med*, 338:1481–1487, 1998.

[ASP+99] M Aarnio, R Sankila, E Pukkala, R Salovaara, L A Aaltonen, A de la Chapelle, P Peltomaki, J P Mecklin, and H J Jarvinen. Cancer risk in mutation carriers of DNA-mismatch-repair genes. *Int J Cancer*, 81:214–218, 1999.

[AUAY05] K G Arbeev, S V Ukraintseva, L S Arbeeva, and A I Yashin. Decline in human cancer incidence rates at old ages: Age-period-cohort considerations. *Demographic Res*, 12:273–300, 2005.

[BA98] K P Burnham and D Anderson. *Model selection and multimodel inference: a practical information-theoretic approach.* Springer, New York, 1998.

[BBK04] E C Butcher, E L Berg, and E J Kunkel. Systems biology in drug discovery. *Nat Biotech*, 22:1253–1259, 2004.

[BDR+07] M Bettstetter, S Dechant, P Ruemmele, M Grabowski, G Keller, E Holinski-Feder, A Hartmann, F Hofstaedter, and W Dietmaier. Distinction of hereditary nonpolyposis colorectal cancer and sporadic microsatellite-unstable colorectal cancer through quantification of MLH1 methylation by real-time PCR. *Clin Cancer Res*, 13:3221–3228, 2007.

[BFB+94] M L Bisgaard, K Fenger, S Bulow, E Niebuhr, and J Mohr. Familial adenomatous polyposis (FAP): frequency, penetrance and mutation rate. *Hum Mutat*, 3:121–125, 1994.

[BHP06] A S Butterworth, J P T Higgins, and P Pharoah. Relative and absolute risk of colorectal cancer for individuals with a family history: A meta-analysis. *Eur J Cancer*, 42:216–227, 2006.

[Bru60] A M Brues. Critique of mutational theories of carcinogenesis. *Acta Unio Int Contra Cancrum*, 16:415–417, 1960.

[BTH+98] C R Boland, S N Thibodeau, S R Hamilton, D Sidransky, J R Eshleman, R W Burt, S J Meltzer, M A Rodriguez-Bigas, R Fodde, G N Ranzani, and S Srivastava. A National Cancer Institute workshop on microsatellite instability for cancer detection and familial predisposition: Development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res*, 58:5248–5257, 1998.

[BTH05] P Burton, M Tobin, and J Hopper. Key concepts in genetic epidemiology. *Lancet*, 336:941–951, 2005.

[CBP04] J Carayol and C Bonaiti-Pellie. Estimating Penetrance From Family Data Using a Retrospective Likelihood When Ascertainment Depends on Genotype and Age of Onset. *Genet Epidemiol*, 27:109–117, 2004.

[CLH02] K Czene, P Lichtenstein, and K Hemminki. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int J Cancer*, 99:260–266, 2002.

[CMJ+05] P Calabrese, J Mecklin, H J Jarvinen, L A Aaltonen, S Tavare, and D Shibata. Numbers of mutations to different types of colorectal cancer. *BMC Cancer*, 5:126, 2005.

[Col06] G A Colditz. Epidemiology - identifying the causes and preventability of cancer? *Nat Rev Cancer*, 6:75–82, 2006.

[CTC+94] K C Chu, R E Tarone, W H Chow, B F Hankey, and L A G Ries. Temporal Patterns in Colorectal Cancer Incidence, Survival, and Mortality From 1950 Through 1990. *J Natl Cancer Inst*, 86:997–1006, 1994.

[CTS04] P Calabrese, S Tavare, and D Shibata. Pretumor Progression: Clonal Evolution of Human Stem Cell Populations. *Am J Pathol*, 164:1337–1346, 2004.

[DBC+04] G Deng, I Bell, S Crawley, J Gum, J P Terdiman, B A Allen, B Truta, M H Sleisenger, and Y S Kim. BRAF Mutation Is Frequently Present in Sporadic Colorectal Cancer with Methylated hMLH1, But Not in Hereditary Nonpolyposis Colorectal Cancer. *Clinical Cancer Research*, 10:191–195, 2004.

[DFC+97] M G Dunlop, S M Farrington, A D Carothers, A H Wyllie, L Sharp, J Burn, B Liu, K W Kinzler, and B Vogelstein. Cancer risk associated with germline DNA mismatch repair gene mutations. *Hum Mol Genet*, 6:105–110, 1997.

[DFN+00] M G Dunlop, S M Farrington, I Nicholl, L Aaltonen, G Petersen, M Porteous, and A Carothers. Population carrier frequency of hMSH2 and hMLH1 mutations. *Br J Cancer*, 83:1643–1645, 2000.

[DH72] R DeMars and K R Held. The spontaneous azaguanine-resistant mutants of diploid human fibroblasts. *Humangenetik*, 16:87–110, 1972.

[dlC03] A de la Chapelle. Microsatellite Instability. *N Engl J Med*, 349:209–210, 2003.

[dlC04] A de la Chapelle. Genetic Predisposition to Colorectal Cancer. *Nat Rev Cancer*, 4:769–780, 2004.

[DPBS04] R Doll, R Peto, J Boreham, and I Sutherland. Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ*, 328:1519, 2004.

[DPW66] R Doll, P Payne, and J Waterhouse. *Cancer Incidence in Five Continents. Vol. 1.* Union Internationale contre le Cancer., Geneva, 1966.

[eaY02] H. et al Yamamoto. Differential involvement of the hypermethylator phenotype in hereditary and sporadic colorectal cancers with high frequency microsatellite instability. *Genes Chromosomes Cancer*, 33:322–325, 2002.

[Els81] R C Elston. Segregation Analysis. *Adv Hum Genet*, 11:63–120, 1981.

[FBD+85] J T Flannery, J D Boice, S S Devesa, R A Kleinerman, R E Curtis, and J F Fraumeni. Cancer Registration in Connecticut and the Study of Multiple Primary Cancers, 1935-1982. *Natl Cancer Inst Monogr*, 68:13–24, 1985.

[Fis01] R Fishel. The Selection for Mismatch Repair Defects in Hereditary Nonpolyposis Colorectal Cancer: Revising the Mutator Hypothesis. *Cancer Res*, 61:7369–7374, 2001.

[Fou08] W D Foulkes. Inherited Susceptibility to Common Cancers. *N Engl J Med*, 359:2143–2153, 2008.

[FPNO+05] A M Fernandez-Peralta, N Nejda, S Oliart, V Medina, M M Azcoita, and J J Gonzalez-Aguilera. Signigicance of mutations in TGFBR2 and BAX in neoplastic progression and patient outcome in sporadic colorectal tumors with high-frequency microsatellite instability. *Cancer Genetics and Cytogenetics*, 157:18–24, 2005.

[Fra04a] S A Frank. A multistage theory of age-specific acceleration in human mortality. *BMC Biology*, 2:16–16, 2004.

[Fra04b] S A Frank. Age-Specific Acceleration of Cancer. *Curr Biol*, 14:242–246, 2004.

[Fra04c] S A Frank. Genetic predisposition to cancer - insights from population genetics. *Nat Rev Genet*, 5:764–772, 2004.

[Fra05] S A Frank. Age-specific incidence of inherited versus sporadic cancers: A test of the multistage theory of carcinogenesis. *Proc Natl Acad Sci USA*, 102:1071–1075, 2005.

[Fra07] S A Frank. *Dynamics of cancer: Incidence, inheritance, and evolution*. Princeton University Press, Princeton, 2007.

[FSW+98] T Fujiwara, J M Stolker, T Watanabe, A Rashid, P Longo, J R Eshleman, S Booker, H T Lynch, J R Jass, J S Green, H Kim, J Jen, B Vogelstein, and S R Hamilton. Accumulated Clonal Genetic Alterations in Familial and Sporadic Colorectal Carcinomas with Widespread Instability in Microsatellite Sequences. *Am J Pathol*, 153:1063–1078, 1998.

[FWB02] N S Fearnhead, J L Wilding, and W F Bodmer. Genetics of colorectal cancer: hereditary aspects and overview of colorectal tumorigenesis. *Br Med Bull*, 64:27–43, 2002.

[GSS+07] C Greenman, P Stephens, R Smith, G L Dalgliesh, C Hunter, G Bignell, H Davies, J Teague, A Butler, C Stevens, S Edkins, S O'Meara, I Vastrik, E E Schmidt, T Avis, and S Barthorpe. Patterns of somatic mutation in human cancer genomes. *Nature*, 446:153–158, 2007.

[Hae61] W Haenszel. Cancer mortality among the foreign born in the United states. *J Natl Cancer Inst*, 26:37–132, 1961.

[Hay65] L Hayflick. The limited in vitro lifetime of human diploid cell strains. *Exp Cell Res*, 37:614–636, 1965.

[HB85] E B Harvey and L A Brinton. Second cancer following cancer of the breast in Connecticut, 1935-1982. *Natl Cancer Inst Monogr*, 68:99–112, 1985.

[HBF06] K Hemminki, J L Bermejo, and A Forsti. The balance between heritable and environmental aetiology of human disease. *Nat Rev Genet*, 7:958–965, 2006.

[HC02] K Hemminki and K Czene. Attributable risks of familial cancer from the family-cancer database. *Cancer Epidemiol Biomarkers Prev*, 11:1638–1644, 2002.

[HCM05] W D Hazelton, M S Clements, and S H Moolgavkar. Multistage Carcinogenesis and Lung Cancer Mortality in Three Cohorts. *Cancer Epidemiol Biomarkers Prev*, 14:1171–1181, 2005.

[HDV01] K Hemminki, C Dong, and P Vaittinen. Cancer Risks to Spouses and Offspring in the Family-Cancer Database. *Genet Epidemiol*, 20:247–257, 2001.

[HJTMF+00] P Herrero-Jimenez, A Tomita-Mitchell, E E Furth, S Morgenthaler, and W G Thilly. Population risk and physiological rate parameters for colon cancer. The union of an explicit model for carcinogenesis with the public health records of the United States. *Mutat Res*, 447:73–116, 2000.

[HL01] W D Hazelton and E G Luebeck. Analysis of a historical cohort of Chinese tin miners with arsenic, radon, cigarette smoke, and pepe smoke exposures using the biologically based two-stage clonal expansion model. *Radiat. Res*, 156:78–, 2001.

[HMRV99] J A Hoeting, D M Madigan, A E Raftery, and C T Volinsky. Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14:382–417, 1999.

[HPLW08] C Harding, F Pompei, E E Lee, and R Wilson. Cancer suppression at old age. *Cancer Res*, 68:4465–4478, 2008.

[HPT07] C Hornsby, K M Page, and I P Tomlinson. What can we learn from the population incidence of cancer? Armitage and Doll revisited. *Lancet Oncol*, 8:1030–1038, 2007.

[HPT08] C P Hornsby, K M Page, and I Tomlinson. The in vivo rate of somatic APC mutation. *Am J Pathol*, 172:1062–1068, 2008.

[HW00] D Hanahan and R A Weinberg. The hallmarks of cancer. *Cell*, 100:57–70, 2000.

[HWB+85] S K Hoar, J Wilson, W J Blot, J K McLaughlin, D M Winn, and A F Kantor. Second cancer following cancer of the digestive system in Connecticut, 1935-1982. *Natl Cancer Inst Monogr*, 68:49–82, 1985.

[ISTB99] M Ilyas, J Straub, I P M Tomlinson, and W F Bodmer. Genetic Pathways in Colorectal and other Cancers. *Eur J Cancer*, 35:335–351, 1999.

[Iwa01] T Iwama. Somatic Mutation Rate of the APC Gene. *Jpn J Clin Oncol*, 31:185–187, 2001.

[JBD+06] M A Jenkins, L Baglietto, J G Dowty, C M van Vliet, L Smith, L J Mead, F A Macrae, D James, B St John, J R Jass, G G Giles, J L Hopper, and M C Southey. Cancer Risks for Mismatch Repair Gene Mutation Carriers: A Population-Based Early Onset Case-Family Study. *Clin Gastroenterol Hepatol*, 4:489–498, 2006.

[JBF+03] J R Jass, M Barker, L Fraser, M D Walsh, V L J Whitehall, B Gabrielli, J Young, and B A Leggett. APC mutation and tumour budding in colorectal cancer. *J Clin Pathol*, 56:69–73, 2003.

[JH01] L E Johns and R S Houlston. A Systematic Review and Meta-Analysis of Familial Colorectal Cancer Risk. *Am J Gastroenterol*, 96:2992–3003, 2001.

[JLC+05] V Johnson, L R Lipton, C Cummings, A T Eftekhar-Sadat, L Izatt, S V Hodgson, I C Talbot, H J W Thomas, A J R Silver, and I P M Tomlinson. Analysis of somatic molecular changes, clinicopathological features, family history, and germline mutations in colorectal cancer families: evidence of distinct groups of non-HNPCC families. *J Med Genet*, 42:756–762, 2005.

[Jon08] D Jones. Pathways to cancer therapy. *Nat Rev Drug Disc*, 7:875–876, 2008.

[JSD+06] B Jung, E J Smith, R T Doctolero, P Gervaz, J C Alonso, K Miyai, T Keku, R S Sandler, and J M Carethers. Influence of target gene muta-

tions on survival, stage and histology in sporadic microsatellite unstable colon cancers. *Int J Cancer*, 118:2509–2513, 2006.

[JVH+05] V Johnson, E Volikos, S E Halford, E T Eftekhar-Sadat, S Popat, I Talbot, K Truninger, J Martin, J Jass, R Houlston, W Atkin, I P M Tomlinson, and A R J Silver. Exon 3 beta-catenin mutations are specifically associated with colorectal carcinomas in hereditary non-polyposis colorectal cancer syndrome. *Gut*, 54:264–267, 2005.

[KB00] T A Kunkel and K Bebenek. DNA replication fidelity. *Annu Rev Biochem*, 69:497–529, 2000.

[KF88] W S Kendal and P Frost. Pitfalls and Practice of Luria-Delbruck Fluctuation Analysis: A Review. *Cancer Res*, 48:1060–1065, 1988.

[KFFM00] H J Kim, M P Fay, E J Feuer, and D N Midthune. Permutation tests for joinpoint regression with applications to cancer rates. *Statist Med*, 19:335–351, 2000.

[KKYT+96] M Konishi, R Kikuchi-Yanoshita, K Tanaka, M Muraoka, A Onda, Y Okumura, N Kishi, T Iwama, T Mori, M Koike, K Ushio, M Chiba, S Nomizu, F Konishi, J Utsunomiya, and M Miyaki. Molecular Nature of Colon Tumors in Hereditary Nonpolyposis Colon Cancer, Familial Polyposis, and Sporadic Colon Cancer. *Gastroenterology*, 111:307–317, 1996.

[Knu71] A G Knudson. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proc Natl Acad Sci USA*, 68:820–823, 1971.

[KS72a] A G Knudson and L C Strong. Mutation and Cancer: A Model for Wilms' Tumor of the Kidney. *J Natl Cancer Inst*, 48:313–324, 1972.

[KS72b] A G Knudson and L C Strong. Mutation and Cancer: Neuroblastoma and Pheochromocytoma. *Am J Hum Genet*, 24:514–532, 1972.

[KSKK98] L A Kunz-Schughart, M Kreutz, and R Knuechel. Multicellular spheroids: a three-dimensional in vitro culture system to study tumour biology. *Int J Exp Pathol*, 79:1–23, 1998.

[KSW+04] T Kambara, L A Simms, V L J Whitehall, K J Spring, C V A Wynter, M D Walsh, M A Barker, S Arnold, A McGivern, N Matsubara, N Tanaka, T Higuchi, J Young, J R Jass, and B A Leggett. BRAF mutation is associated with DNA methlylation in serrated polyps and cancers of the colorectum. *Gut*, 53:1137–1144, 2004.

[KT00] P Kraft and D C Thomas. Bias and Efficiency in Family-Based Gene-Characterization Studies: Conditional, Prospective, and Joint Likelihoods. *Am J Hum Genet*, 66:1119–1131, 2000.

[KZH+03] D Krewski, J M Zielinski, W D Hazelton, M J Garner, and S H Moolgavkar. The use of biologically based cancer risk models in radiation epidemiology. *Radiat Prot Dosimetry*, 104:367–376, 2003.

[LdlC03] H T Lynch and A de la Chapelle. Hereditary colorectal cancer. *N Engl J Med*, 348:919–932, 2003.

[LdLJ+97] L Losi, M Ponz de Leon, J Jiricny, C D Gregorio, P Benatti, A Percesepe, R Fante, L Roncucci, M Pedroni, and J Benhattar. K-ras and p53 mutations in hereditary non-polyposis colorectal cancers. *Int J Cancer*, 74:94–96, 1997.

[LH89] M Lichten and J E Haber. Position Effects in Ectopic and Allelic Mitotic Recombination in Saccharomyces cerevisiae. *Genetics*, 123:261–268, 1989.

[LHV+00] P Lichtenstein, N V Holm, P K Verkasalo, A Iliadou, J Kaprio, M Koskenvuo, E Pukkala, A Skytthe, and K Hemminki. Environmental and heritable factors in the causation of cancer - analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med*, 343:78–85, 2000.

[LIR+99] H Lamlum, M Ilyas, A Rowan, S Clark, V Johnson, J Bell, I Frayling, J Efstathiou, K Pack, S Payne, R Roylance, P Gorman, D Sheer, K Neale, R Phillips, and I Talbot. The type of somatic mutation at APC in familial adenomatous polyposis is determined by the site of the germline mutation: a new facet to Knudson's 'two-hit' hypothesis. *Nat Med*, 5:1071–1075, 1999.

[Lit96] M P Little. Generalisations of the two-mutation and classical multi-stage models of carcinogenesis fitted to the Japanese atomic bomb survivor data. *J Radiol Prot*, 16:7–24, 1996.

[LKV98] C Lengauer, K W Kinzler, and B Vogelstein. Genetic instabilities in human cancers. *Nature*, 386:643–649, 1998.

[LM02] G Luebeck and S H Moolgavkar. Multistage carcinogenesis and the incidence of colorectal cancer. *Proc Natl Acad Sci USA*, 99:15095–15100, 2002.

[LPBS98] P Laurent-Puig, C Beroud, and T Soussi. APC gene: database of germline and somatic mutations in human tumors and cell lines. *Nucleic Acids Res*, 26:269–270, 1998.

[LRLY07] I Locatelli, A Rosina, P Lichtenstein, and A I Yashin. A correlated frailty model with long-term survivors for estimating the heritability of breast cancer. *Statist Med*, 26:3722–3734, 2007.

[LSA01] E Limpert, W A Stahel, and M Abbt. Log-normal Distributions across the Sciences: Keys and Clues. *BioScience*, 51:341–352, 2001.

[LW03] M P Little and E G Wright. A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data. *Math Biosci*, 183:111–134, 2003.

[MDS80] S H Moolgavkar, N E Day, and R G Stevens. Two-stage model for carcinogenesis: Epidemiology of breast cancer in females. *J Natl Cancer Inst*, 65:559–69, 1980.

[MDV88] S H Moolgavkar, A Dewanji, and D Venzon. A Stochastic Two-Stage Model for Cancer Risk Assessment. I. The Hazard Function and the Probability of Tumor. *Risk Anal*, 8:383–392, 1988.

[MFDC02] R J Michell, S M Farrington, M G Dunlop, and H Campbell. Human Repair Genes hMLH1 and hMSH2 and Colorectal Cancer: A HuGE Review. *American Journal of Epidemiology*, 156:885–902, 2002.

[MIK+99] M Miyaki, T Iijima, J Kimura, M Yasuno, T Mori, Y Hayashi, M Koike, N Shitara, T Iwama, and T Kuroki. Frequent Mutation of beta-Catenin and APC Genes in Primary Colorectal Tumors from Patients with Hereditary Nonpolyposis Colorectal Cancer. *Cancer Res*, 59:4506–4509, 1999.

[MIN06] F Michor, Y Iwasa, and M A Nowak. The age incidence of chronic myeloid leukemia can be explained by a one-mutation model. *Proc Natl Acad Sci USA*, 103:14931–14934, 2006.

[MM30] R S McCombs and R P McCombs. A Hypothesis on the Causation of Cancer. *Science*, 72:423–424, 1930.

[MVSC+07] M L Maestro, M Vidaurreta, M T Sanz-Casla, S Rafael, S Veganzones, A Martinez, C Aguilera, M D Herranz, J Cerdan, and M Arroyo. Role of the BRAF Mutations in the Microsatellite Instability Genetic Pathway in Sporadic Colorectal Cancer. *Ann Surg Oncol*, 14:1229–1236, 2007.

[MWW+04] A McGivern, C V A Wynter, V L J Whitehall, T Kambara, K J Spring, M D Walsh, M A Barker, S Arnold, L A Simms, B A Leggett, J Young, and J R Jass. Promoter hypermethylation frequency and BRAF mutations distinguish hereditary non-polyposis colon cancer from sporadic MSI-H colon cancer. *Familial Cancer*, 3:101–107, 2004.

[MYFS90] E R Maher, J R W Yates, and M A Ferguson-Smith. Statistical analysis of the two stage mutation model in von Hippel-Lindau disease, and in sporadic cerebellar haemangioblastoma and renal cell carcinoma. *J Med Genet*, 27:311–314, 1990.

[Nor53] C O Nordling. A new theory on the cancer-inducing mechanism. *Br J Cancer*, 7:68–72, 1953.

[PA07] J Pellettieri and A S Alvarado. Cell Turnover and Adult Tissue Homeostasis: From Humans to Planarians. *Annu Rev Genet*, 41:83–105, 2007.

[PAB+02] P D P Pharoah, A Antoniou, M Bobrow, R L Zimmern, and D F Easton. Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet*, 31:33–36, 2002.

[Par06] D M Parkin. Evolution of the population-based cancer registry. *Nat Rev Cancer*, 6:603–612, 2006.

[PBH03] C S Potten, C Both, and D Hargreaves. The small intestine as a model for evaluating adult tissue stem cell drug targets. *Cell Prolif*, 36:115–129, 2003.

[PDD+97] P D P Pharoah, N E Day, S Duffy, E F Easton, and B A J Ponder. Family history and the risk of breast cancer: a systematic review and meta-analysis. *Int J Cancer*, 71:800–809, 1997.

[Pet01] J Peto. Cancer epidemiology in the last century and the next decade. *Nature*, 411:390–395, 2001.

[PM95] M K B Parmar and D Machin. *Survival Analysis: A Practical Approach*. Wiley, Chichester, 1995.

[PM00] J Peto and T M Mack. High constant incidence in twins and other relatives of women with breast cancer. *Nat Genet*, 26:411–414, 2000.

[Pri58] C H Price. Primary bone-forming tumours and their relationship to skeletal growth. *J Bone Joint Surg. Br.*, 40:574–593, 1958.

[PV04] P Peltomaki and H Vasen. Mutations associated with HNPCC predisposition - Update of ICG-HNPCC/InSiGHT mutation database. *Dis Markers*, 20:269–276, 2004.

[PW01] F Pompei and R Wilson. Age Distribution of Cancer: The Incidence Turnover at Old Age. *HERA*, 7:1619–1650, 2001.

[PWF+97] D M Parkin, S L Whelan, J Ferlay, L Raymond, and J Young. *Cancer Incidence in Five Continents, Vol. VII*. IARC, Lyon, 1997.

[PWF+02] D M Parkin, S L Whelan, J Ferlay, L Teppo, and D B Thomas. *Cancer Incidence in Five Continents, Vol. VIII*. IARC, Lyon, 2002.

[PWFS05] D M Parkin, S L Whelan, J Ferlay, and H Storm. *Cancer incidence in five continents: Volumes I to VIII*. IARC, Lyon, 2005.

[PWO07] F Prall, V Weirich, and C Ostwald. Phenotypes of invasion in sporadic colorectal carcinomas related to aberrations of the adenomatous polyposis coli (APC) gene. *Histopathology*, 50:318–330, 2007.

[QB02] M Quinn and P Babb. Patterns and trends in prostate cancer incidence, survival, prevalence and mortality, Part II individual countries. *BJU Int*, 90:174–184, 2002.

[QVvH05] F Quehenberger, H F A Vasen, and H C van Houwelingen. Risk of colorectal and endometrial cancer for carriers of mutations of the hMLH1 and hMSH2 gene: correction for ascertainment. *J Med Genet*, 42:491–496, 2005.

[RHG+05] A Rowan, S Halford, M Gaasenbeek, Z Kemp, O Sieber, E Volikos, E Douglas, H Fiegler, N Carter, I Talbot, A Silver, and I Tomlinson. Refining Molecular Analysis in the Pathways of Colorectal Carcinogenesis. *Clin Gastroenterol Hepatol*, 3:1115–1123, 2005.

[RLI+00] A J Rowan, H Lamlum, M Ilyas, J Wheeler, J Straub, A Papadopoulou, D Bicknell, W F Bodmer, and I P M Tomlinson. APC mutations in sporadic colorectal tumors: A mutational hotspot and interdependence of the two hits. *Proc Natl Acad Sci USA*, 97:3352–3357, 2000.

[RNVL03] H Rajagopalan, M A Nowak, B Vogelstein, and C Lengauer. The significance of unstable chromosomes in colorectal cancer. *Nat Rev Cancer*, 3:695–701, 2003.

[RYI+97] N Rampino, H Yamamoto, Y Ionov, Y Li, H Sawai, J C Reed, and M Perucho. Somatic Frameshift Mutations in the BAX Gene in Colon Cancers of the Microsatellite Mutator Phenotype. *Science*, 275:967–969, 1997.

[SCS97] M A Smith, T Chen, and R Simon. Age-Specific Incidence of Acute Lymphoblastic Leukemia in U.S. Children: In Utero Initiation Model. *J Natl Cancer Inst*, 89:1542–1544, 1997.

[SHT03] O M Sieber, K Heinimann, and I P M Tomlinson. Genomic instability - the engine of tumorigenesis? *Nat Rev Cancer*, 3:701–708, 2003.

[SHW+97] J P Struewing, P Hartge, S Wacholder, S M Baker, M Berlin, M McAdams, M M Timmerman, L C Brody, and M A Tucker. The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Askenazi jews. *N Engl J Med*, 336:1401–1408, 1997.

[Siv96] D S Sivia. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, New York, 1996.

[SJW+06] T Sjoblom, S Jones, L D Wood, D W Parsons, J Lin, T D Barber, D Mandelker, R J Leary, J Ptak, N Silliman, S Szabo, P Buckhaults, C Farrell, P Meeh, S D Markowitz, and J Willis. The Consensus Coding Sequences of Human Breast and Colorectal Cancers. . *Science*, 314:268274, 2006.

[SKP+99] S Salahshor, U Kressner, L Pahlman, B Glimelius, G Lindmark, and A Lindblom. Colorectal Cancer With and Without Microsatellite Instability Involves Different Genes. *Genes Chromosomes Cancer*, 26, 1999.

[SKT+87] R Seshadri, R J Kutlaca, K Trainor, C Matthews, and A A Morley. Mutation Rate of Normal and Malignant Human Lymphocytes. *Cancer Res*, 47:407–409, 1987.

[SLK+00] R Salovaara, A Loukola, P Kristo, H Kaarianinen, H Ahtola, M Eskelinen, N Harkonen, R Julkunen, E Kangas, S Ojala, J Tulikoura, E Valkamo, H Jarvinen, J Mecklin, L A Aaltonen, and A de la Chapelle. Population-Based Molecular Detection of Hereditary Nonpolyposis Colorectal Cancer. *Journal of Clinical Oncology*, 18:2193–2200, 2000.

[SMB+06] H Schollnberger, M Manuguerra, H Bijwaard, H Boshuizen, H P Altenburg, S M Rispens, M J P Brugmans, and P Vineis. Analysis of epidemiological cohort data on smoking effects and lung cancer with a multi-stage cancer model. *Carcinogenesis*, 27:1432–1444, 2006.

[SMT+06] E Svensson, T A Moger, S Tretli, O O Aalen, and T Grotmol. Frailty modelling of colorectal cancer incidence in Norway: Indications that individual heterogeneity in risk is related to birth cohort. *Eur J Epidemiol*, 21:587–593, 2006.

[SRV+06] N Suraweera, J Robinson, E Volikos, T Guenther, I Talbot, I Tomlinson, and A Silver. Mutations within Wnt pathway genes in sporadic colorectal cancers and cell lines. *Int J Cancer*, 119:1837–1842, 2006.

[ST06] S Segditsas and I Tomlinson. Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene*, 25:7531–7537, 2006.

[STK+07] L Shen, M Toyota, Y Kondo, E Lin, L Zhang, Y Guo, N S Hernandez, X Chen, S Ahmed, K Konishi, S R Hamilton, and J P J Issa. Intergrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *PNAS*, 104:18654–18659, 2007.

[TB95] I P M Tomlinson and W F Bodmer. Failure of programmed cell death and differentiation as causes of tumors: Some simple mathematical models. *PNAS*, 92:11130–11134, 1995.

[TSO+01] G Togo, Y Shiratori, M Okamoto, Y Yamaji, M Matsumura, T Sano, T Motojima, and M Omata. Relationship Between Grade of Microsatel-

lite Instability and Target Genes of Mismatch Repair Pathways in Sporadic Colorectal Carcinoma. *Dig Dis Sci*, 46:1615–1622, 2001.

[Tur90] T Turanyi. Sensitivity Analysis of Complex Kinetic Systems. Tools and Applications. *J Math Chem*, 5:203–248, 1990.

[TWCC+07] I Tomlinson, E Webb, L Carvajal-Carmona, P Broderick, Z Kemp, S Spain, S Penegar, I Chandler, M Gorman, E Wood, E Barclay, S Lubbe, L Martin, G Sellick, E Jaeger, and R Hubner. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet*, 39:984–988, 2007.

[Vea65] A M O Veale. *Intestinal Polyposis. Eugenics Laboratory Memoirs, No. 40.* Cambridge University Press, Cambridge, 1965.

[VWM+96] H F A Vasen, J T Wijnen, F H Menko, J H Kleibeuker, B G Taal, G Griffioen, F M Nagengast, E H Meijers-Heijboer, L Bertario, L Varesco, M L Bisgaard, J Mohr, R Fodde, and P M Khan. Cancer Risk in Families With Hereditary Nonpolyposis Colorectal Cancer Diagnosed by Mutation Analysis. *Gastroenterology*, 110:1020–1027, 1996.

[WSC+06] D J Weisenberger, K D Siegmund, M Campan, J Young, T I Long, M A Faasse, G H Kang, M Widschwendter, D Weener, D Buchanan, H Koh, L Simms, M Barker, B Leggett, J Levine, and M Kim. CpG island methylator phenotype underlies sporadic mircrosatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet*, 38:787–793, 2006.

[YAN+98] O K Yagi, Y Akiyama, T Nomizu, T Iwama, M Endo, and Y Yuasa. Proapoptotic Gene BAX Is Frequently Mutated in Hereditary Nonpolyposis Colorectal Cancers but Not in Adenomas. *Gastroenterology*, 114:268–274, 1998.

[Yan00] M C K Yang. *Introduction to statistical methods in modern genetics*. Gordon and Breach Science Publishers, Singapore, 2000.

[YIM+06] T Yamaguchi, T Iijima, T Mori, K Takahashi, H Matsumoto, H Miyamoto, T Hishima, and M Miyaki. Accumulation Profile of Frameshift Mutations During Development and Progression of Colorectal Cancer From Patients With Hereditary Nonpolyposis Colorectal Cancer. *Dis Colon Rectum*, 49:399–406, 2006.

[YK90] L W Yuan and R L Keil. Distance-Independence of Mitotic Intrachromosomal Recombination in Saccharomyces cerevisiae. *Genetics*, 124:263–273, 1990.

[YSB+01] J Young, L A Simms, K G Biden, C Wynter, V Whitehall, R Karamatic, J George, J Goldblatt, I Walpole, S A Robin, M M Borden, R Stitz, J Searle, D McKeone, L Frazier, and D R Purdie. Features of Colorectal Cancers with High-Level Microsatellite Instability Occurring in Familial and Sporadic Settings. *Am J Pathol*, 159:2107–2116, 2001.

[YSW+98] H Yamamoto, H Sawai, T K Weber, M A Rodriguez-Bigas, and M Perucho. Somatic Frameshift Mutations in DNA Mismatch Repair and Proapoptosis Genes in Hereditary Nonpolyposis Colorectal Cancer. *Cancer Res*, 58:997–1003, 1998.

[YTS01] Y Yatabe, S Tavare, and D Shibata. Investigating stem cells in human colon by using methylation patterns. *Proc Natl Acad Sci USA*, 98:10839–10844, 2001.

[YVI95] A I Yashin, J W Vaupel, and I A Iachine. Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies*, 5:145–159, 1995.

[ZS05] X Zhang and R Simon. Estimating the number of rate-limiting genomic changes for human breast cancer. *Breast Cancer Res Treat*, 91:121–124, 2005.

Appendix A

In section 2.3.1 it was claimed in equation (2.16) that if y, $x_1, x_2, \ldots$ and x_n all map from a vector space, V, into the real numbers and y is such that:

$$y(a+b) = y(a) \cdot y(b), \forall a, b \in V,$$

then

$$\sum_r y(r) \sum_{\substack{r_1,r_2,\ldots,r_n \\ r_1+r_2+\cdots+r_n=r}} \prod_{i=1}^{n} x_i(r_i) = \prod_{i=1}^{n} \left(\sum_r y(r) x_i(r) \right). \tag{A1}$$

To prove (A1) first note that, $\forall n$,

$$\begin{aligned} \sum_{r_1,r_2,\ldots,r_n} y\left(\sum_{i=1}^{n} r_i\right) \prod_{i=1}^{n} x_i(r_i) &= \sum_{r_1,r_2,\ldots,r_n} \prod_{i=1}^{n} y(r_i) x_i(r_i) \\ &= \prod_{i=1}^{n} \left(\sum_r y(r) x_i(r) \right). \end{aligned} \tag{A2}$$

The LHS of (A1) can then be rewritten as:

$$
\begin{aligned}
&\sum_r y(r) \sum_{\substack{r_1,r_2,\ldots,r_n \\ r_1+r_2+\cdots+r_n=r}} \prod_{i=1}^{n} x_i(r_i) \\
&\quad = \sum_r \sum_{\substack{r_1,r_2,\ldots,r_n \\ r_1+r_2+\cdots+r_n=r}} y(r) \prod_{i=1}^{n} x_i(r_i) \\
&\quad = \sum_r \sum_{\substack{r_1,r_2,\ldots,r_n \\ r_1+r_2+\cdots+r_n=r}} y\left(r-\sum_{i=1}^{n-1} r_i\right) y\left(\sum_{i=1}^{n-1} r_i\right) \prod_{i=1}^{n} x_i(r_i) \\
&\quad = \sum_r \sum_{r_1,r_2,\ldots,r_{n-1}} y\left(r-\sum_{i=1}^{n-1} r_i\right) y\left(\sum_{i=1}^{n-1} r_i\right) \left(\prod_{i=1}^{n-1} x_i(r_i)\right) x_n\left(r-\sum_{i=1}^{n-1} r_i\right) \\
&\quad = \sum_{r_1,r_2,\ldots,r_{n-1}} y\left(\sum_{i=1}^{n-1} r_i\right) \prod_{i=1}^{n-1} x_i(r_i) \sum_r y\left(r-\sum_{i=1}^{n-1} r_i\right) x_n\left(r-\sum_{i=1}^{n-1} r_i\right) \\
&\quad = \sum_{r_1,r_2,\ldots,r_{n-1}} y\left(\sum_{i=1}^{n-1} r_i\right) \prod_{i=1}^{n-1} x_i(r_i) \left[\sum_r y(r) x_n(r)\right] \\
&\quad = \prod_{i=1}^{n} \left(\sum_r y(r) x_i(r)\right), \text{ by (A2).}
\end{aligned}
$$

Appendix B

Lognormal distribution

The lognormal distribution has density:

$$f(x;\mu_n,\sigma_n) = \frac{1}{x\sigma_n\sqrt{2\pi}}e^{-\frac{(\log(x)-\mu_n)^2}{2\sigma_n^2}},$$

where μ_n and σ_n are the mean and variance of the associated normal-distribution. The mean and variance, μ_l and σ_l^2 of a lognormal variable with density given above are:

$$\mu_l = e^{\mu_n+\frac{\sigma_n^2}{2}},$$

and

$$\sigma_l^2 = (e^{\sigma_n^2}-1)e^{2\mu_n+\sigma_n^2}.$$

Conversely,

$$\mu_n = \log(\mu_l) - \frac{1}{2}\log\left(1+\frac{\sigma_l^2}{\mu_l^2}\right)$$

and

$$\sigma_n^2 = \log\left(1+\frac{\sigma_l^2}{\mu_l^2}\right).$$

Hence, if μ_l is set equal to one then $\mu_n = -\frac{1}{2}\sigma_n^2$.

Bivariate lognormal distribution

The density of the bivariate lognormal distribution for (X, Y) when X and Y have the same variance is:

$$f(x,y) = \frac{1}{2\pi\sigma_n^2(1-\rho_n^2)^{\frac{1}{2}}} e^{-q/2},$$

$$q = \frac{1}{1-\rho_n^2}\left[\left(\frac{\log(x)-\mu_n}{\sigma_n}\right)^2 - 2\rho_n\left(\frac{\log(x)-\mu_n}{\sigma_n}\right)\left(\frac{\log(y)-\mu_n}{\sigma_n}\right) + \left(\frac{\log(y)-\mu_n}{\sigma_n}\right)^2\right].$$

where

$$\rho_n = \frac{1}{\sigma_n^2}\log\left(1 + \left(e^{\sigma_n^2} - 1\right)\rho_l\right).$$

Here ρ_l is the correlation coefficient of the lognormal bivariates X and Y. μ_n, σ_n and ρ_n are the mean, variance and correlation of the associated multinormal pair. Conversely,

$$\rho_l = \frac{e^{\sigma_n^2\rho_n} - 1}{e^{\sigma_n^2} - 1}.$$

Glossary of terms and abbreviations

FAP	familial adenomatous polyposis coli, a Mendelian cancer syndrome
HNPCC	hereditary non-polyposis colorectal cancer, a Mendelian cancer syndrome
APC	adenomatous polyposis coli, a tumour suppressor gene, gives rise to FAP
MMR	miss-match repair
RB	retinoblastoma, a tumour suppressor gene
hMLH1	human mut-L homolog 1, a caretaker gene involved in DNA repair that gives rise to HNPCC
hMSH2	human mut-S homolog, a caretaker gene involved in DNA repair that gives rise to HNPCC
BRCA1	breast cancer 1 caretaker gene, gives rise to hereditary breast and ovarian cancers
BRCA2	breast cancer 2 caretaker gene, gives rise to hereditary breast and ovarian cancers
TSCE	two stage clonal expansion model, a quantitative model of tumorigenesis involving two mutations separated by a clonal expansion

penetrance	the proportion of individuals with a particular genotype who express a particular phenotype
incidence	the rate at which a disease occurs in a population measured in disease counts per unit time per unit population
hazard function	closely related to incidence, a function of age giving the instantaneous rate of disease in a population or individual
LLA	log-log acceleration, the gradient of log(incidence) against log(age), provides a measure of the change in incidence of a disease with age
ΔLLA	the difference in acceleration measured in two diseases
Bayesian method	a procedure used to improve a statistical model in the light of observed data
likelihood function	a function central to Bayesian statistics, used to quantify the quality of a statistical model
prior distribution	used to represent knowledge of a parameter or other attribute of a statistical model in advance of observing data pertinent to that parameter or attribute
posterior distribution	used to represent knowledge of a given parameter or attribute in the light of observed data