What can we learn from the population incidence of cancer? Armitage and Doll revisited

Chris Hornsby, Karen M Page, Ian PM Tomlinson

Lancet Oncol 2007; 8: 1030–38

Department of Computer Science (C Hornsby URCS) **and Department of Mathematics** (K M Page DPhil)**, University College London, London, UK; Molecular and Population Genetics Laboratory, London Research Institute, Cancer Research UK, London, UK** (Prof I P M Tomlinson DPhil)

Correspondence to:
Mr Chris Hornsby, Department of Computer Science, University College London, WC1E 6BT, UK
c.hornsby@ucl.ac.uk

Most cancers occur with the same characteristic pattern of incidence. The simplicity of this pattern is in contrast to the perceived complexity of carcinogenesis. Therefore, age-onset statistics represent a tempting 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. Although their basic notion 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. Why have we been unable to elucidate the quantitative dependence of cancer incidence on the molecular processes that feature in its aetiology? In this review we aim to provide answers for this question.

Introduction

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 20th century. The advent of population-based cancer registries (PBCR) led to the first reliable statistics on rates of cancer by age at diagnosis and site. A population-based cancer registries obtains these data by recording every new case of cancer in a defined population—usually people living within a specified geographical area. Beginning in Europe in 1927 and in North America in 1940, population-based cancer registration has developed into a worldwide activity. The International Association of Cancer Registries currently has 449 member registeries worldwide, covering more than 20% of the world's population.

The rise of population-based cancer registration was motivated by the desire to compare cancer prevalence between different places and over time.[1] Such comparisons have uncovered potential carcinogens through the identification of environmental factors that modify cancer risk. On the basis of the finding that cancer incidence in migrants often matches that of their new country,[2] the conclusion was made in the early 1980s that large disparities in cancer burden between the UK and the USA and other countries were attributable to differences in diet (figure 1), smoking, reproductive behaviour, sexual behaviour, infection, and occupational exposures.[3] The existence and extent of these associations have been confirmed in subsequent epidemiological studies.[3] Accumulated registry data have also been put to use in many aspects of cancer control, from planning to the assessment of screening and treatment programmes.[4]

Figure 1: **Large international disparities in cancer burden are attributable to diet and other environmental factors**

An alternative branch of cancer epidemiology developed in parallel to the standard study of incidence. In 1954, Armitage and Doll[5] published a landmark study on the age distribution of cancer. Mortality statistics (viewed as a good indicator of incidence) recorded in several developed countries had shown an intriguing dependence of cancer on age.[6] The number of deaths in a specified age group, recorded over a year, was roughly proportional to the "nth" power of age, with "n" being about five or six for many cancers, including the common carcinomas. We now know this to be true of incidence as well. The incidence of cancer is described as log–log linear, because it appears as a straight line when plotted against age on double logarithmic paper. Figure 2 shows this relation for three cancers of distinct histological origin. In addition to an exponential character, the incidence of leukaemias and sarcomas show small peaks in early childhood and adolescence, respectively. These peaks could be consistent with periods of intense proliferation in the cancer target cells.

Armitage and Doll proposed a multistage theory to account for the log–log linear observation. They showed that if about six rare cellular changes led to cancer (figure 3A), then its age distribution would have a shape that is roughly consistent with the actual observed incidence (figure 4A). Their proposed cellular changes can be equated to gene (epigenetic) mutations.

The key to Armitage and Doll's formulation is to assume that cancer arises in a susceptible target of

asymmetrically dividing stem cells. Each stem cell and its lineal descendants can 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 straightforward interpretation. It can be interpreted as the probability that at least one of the susceptible stem-cell lineages that make up the organ has acquired the necessary number of mutations by the age given. A crude expression for this probability can be written in terms of the number of lineages at risk, "N", the number of mutations needed, "n", and also the probability of mutation per year at each locus, "μ".

As an illustrative test of the model, if the incidence of colon cancer (figure 4B) is fitted with the assumption that 10^8 stem-cell lineages (N) are at risk of malignant conversion in the average colon,[9] then the implication is made that n=6 and $\mu=8\times10^{-4}$. Data used in this fit were recorded from Finnish females between 1959 and 1961.[1] We used a **Bayesian method** with uniform **prior distributions:** $2\leq n\leq 9$ and $10^{-8}\leq\mu\leq10^{-3}$. The **likelihood function** was constructed according to Luebeck and Moolgavkar,[10] from a generalised multistage model hazard derived by Little,[11] with all growth and death rates set to zero. The estimate for μ is several orders of magnitude higher than those made in human-cell cultures.[12] This difference is most probably because the model does not take account of selection and clonal growth; two mechanisms which can accelerate the multistage process despite low gene mutations.

Understanding non-log–log linear cancer incidence

Since at least the 1930s, the suggestion has been made that cancer might arise through mutations in the hereditary material of a somatic cell.[13] Despite this notion, 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. A popular theory was that cancer arose from embryonic cells that had failed to differentiate and persisted in adult tissues. Even as late as 1960, substantial doubt still existed regarding mutational theories.[14] In this context, the pertinent insight of Armitage and Doll was that steep increases in the incidence of cancer with age are indicative of random, heritable, multiple, and rare causal cellular events. However, the simplifying assumptions of their original theory are insufficient for a more exact understanding of the age dependence of a particular cancer.

The incidence of every type of cancer deviates to some extent from log–log linearity. Moreover, a cell-centric model, which 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 is the two-stage clonal expansion (TSCE) model[15] and its derivatives.[10,16]

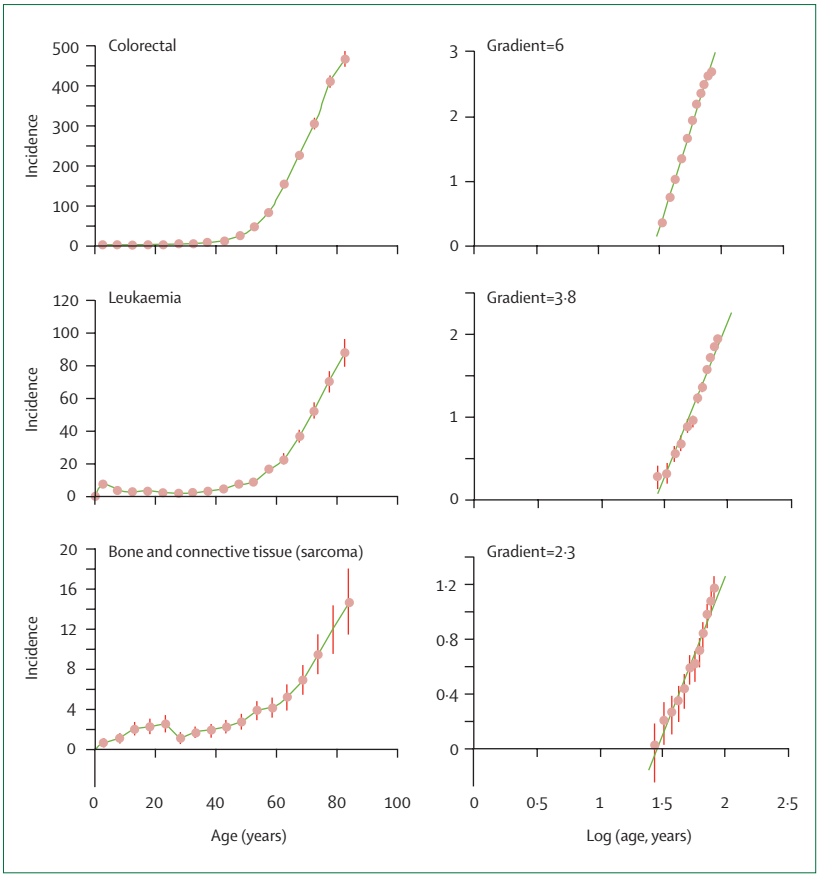

Figure 2: **Incidence of cancer against age (left) and log–log plots of this incidence (right)**
Gradients were calculated using a least squares method. Year of diagnosis=2003. Data from reference 7.

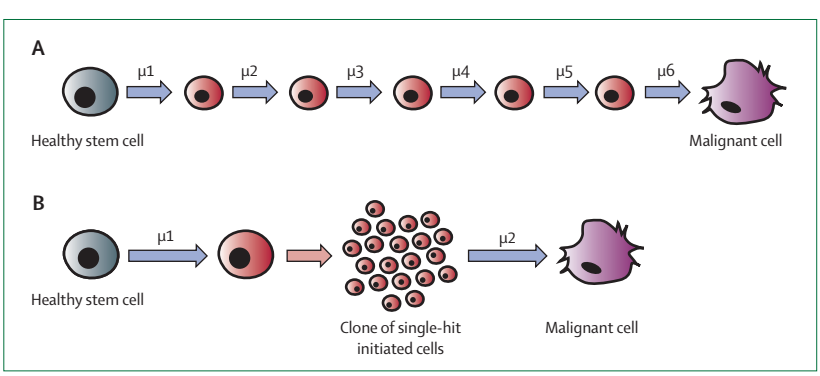

Figure 3: **Schematics of two contrasting multistage theories**
(A) Original multistage theory: a healthy stem-cell lineage becomes transformed through sequential mutations at different rates (μ1–μ6). (B) Two-stage clonal expansion model: an initial mutation (μ1) causes a clone to grow and any of the cells in the clone can then become malignant through a second mutation (μ2).

TSCE (figure 3B) shares the common basic assumptions of the original multistage theory. For example, it shares the assumption that 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. However, this model differs from the

Bayesian method
A procedure used to improve a statistical model in the light of observed data

Prior distributions
Used to represent knowledge of a model parameter before observing the data fit

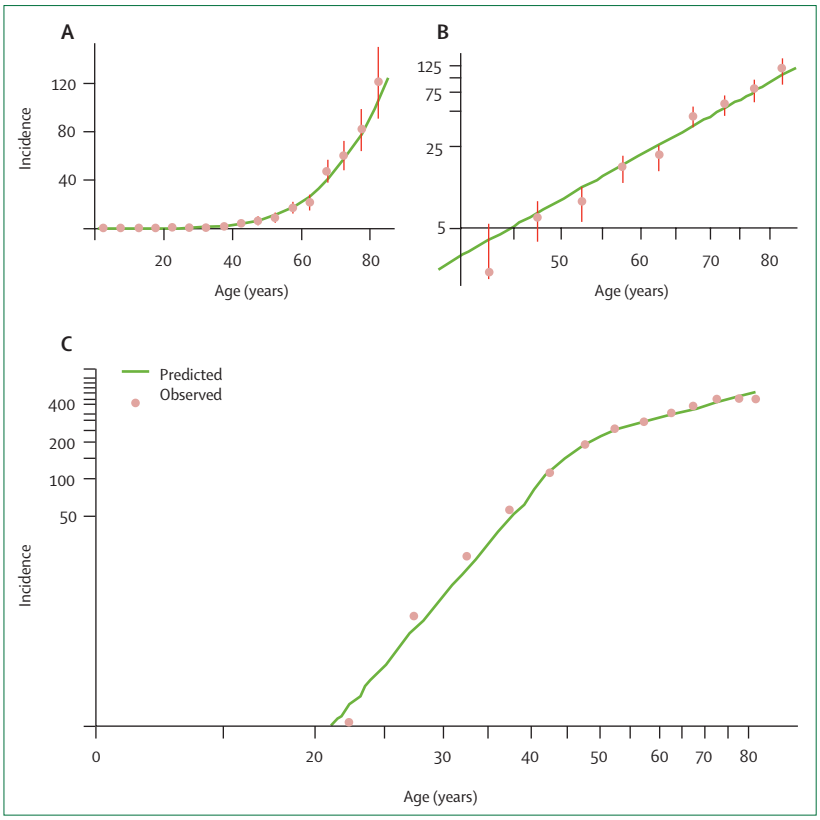

Figure 4: **Incidence patterns of colon and breast cancer**
(A) Fit of Armitage and Doll's model to incidence of colon cancer, with 95% CIs shown for observed data. Year of diagnosis=1959–1961.[1] (B) Log–log fit of Armitage and Doll's model, with 95% CIs shown for observed data. (C) Clemmesen's hook for breast-cancer incidence against age. Year of diagnosis=1993–1997.[8]

Likelihood function
A mathematical expression, central to Bayesian statistics, used to quantify the quality of a statistical model

original multistage model 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 rise 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.

TSCE has proved to be a versatile theory, which is able to synthesise several incidence patterns, both log–log linear and otherwise. From a technical perspective, its biggest triumph lies in its stochastic representation of clonal growth. Less accurate descriptions of clonal growth treat the expansion of an initiated lineage as inevitable once the lineage has arrived at a certain genotype. In TSCE, mutant clones can 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 restriction of TSCE is that it only allows for two rate-limiting stages. Although most, if not all cancers, contain more than two mutations, TSCE was designed to be consistent with the major rate-limiting steps identified in chemical tumorigenesis: initiation (first mutation); promotion (clonal growth); and progression (final mutation). Elegant generalisations of TSCE have removed this restriction[10,16] and can account for many sequential rounds of clonal expansion at different growth rates. Therefore, distinct functional implications of specific mutations can be considered within this framework.

Typical derivatives of TSCE and other multistage theories are obtained by allowing certain parameters to vary with time. For example, a mutation rate might change with age to indicate changing effects of the tissue microenvironment. Alternatively, the number of cell lineages at risk might increase with age to account for tissue growth between conception and adulthood. More subtle phenotypic effects, such as genome destabilisation, have also been quantified.[16] Applications of some incarnations of the multistage theory are presented below.

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 an incidence profile referred to as Clemmesen's hook, so named because of its appearance on log–log paper (figure 4C). Its shape is consistent with a rapid increase in risk starting in the third decade of life followed by a gentler rise during the end of the fifth decade and continuing into old age.

A derivative of TSCE has been used in a quantitative attempt to explain breast cancer incidence.[17] The basic TSCE model, with a stem-cell pool of constant size, provides a poor fit to data (figure 5; left column), but improvements are made with suitable modifications. First, the prediction for the risk of cancer at young ages can be improved by assuming that these susceptible target cells (breast stem cells) grow in numbers that are consist with the development of the breast during puberty (figure 5; middle column). The excess risk associated 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 indicate involution of the breast in old age (figure 5; right column).

Smoking and lung cancer

Bronchial carcinoma arises in a classic log–log linear fashion in non-smokers and is substantially more prevalent in the smoking population. Multistage models have been used in attempts to explain the excess risk in smoking cohorts, with the goal of developing an understanding of the mechanism through which tobacco smoke exerts its carcinogenic effect. Typically, a model of lung cancer is formulated for non-smokers. This model is then adapted for a given smoking cohort by modyifing parameters (eg, mutation rates) from their basal level during the years in which the members of the cohort have used cigarettes. The magnitude of these modifications depends on the smoking level of the cohort members, measured in cigarettes per day.

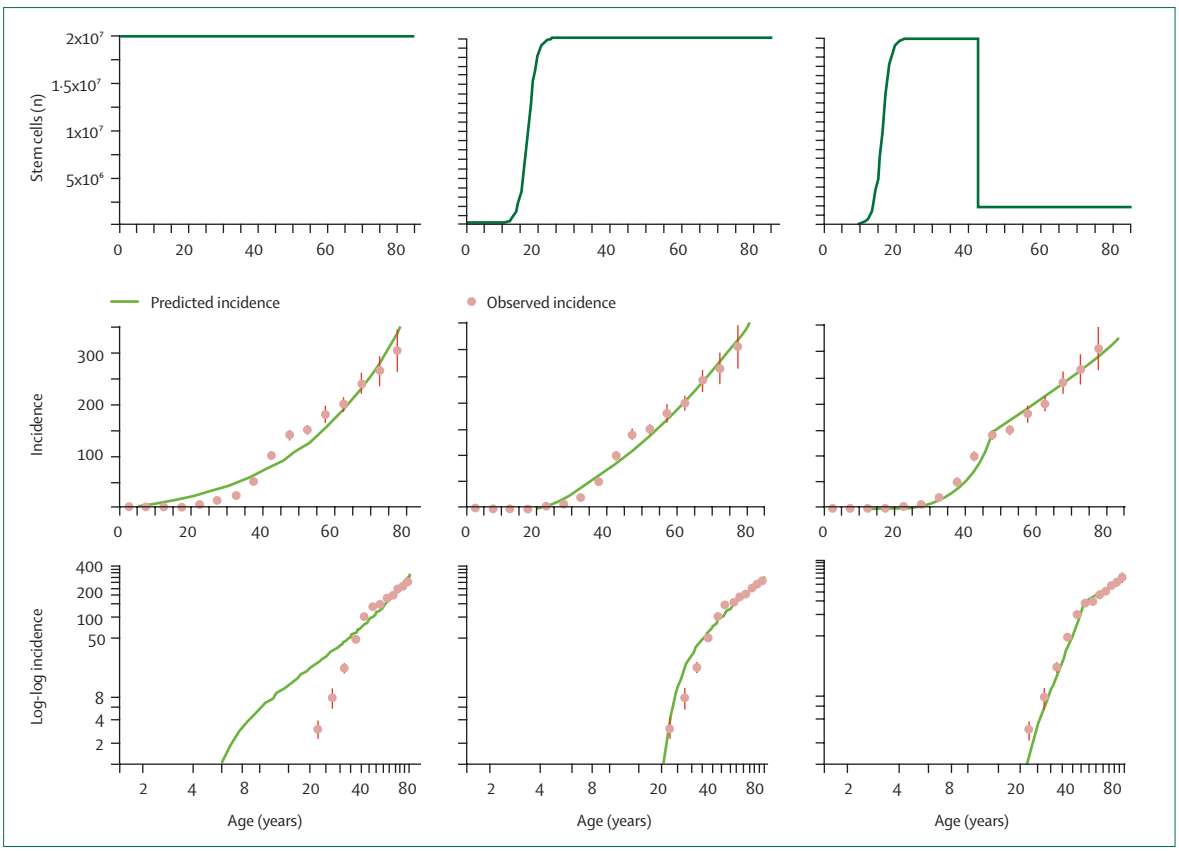

Figure 5: **Modifications of TSCE model for stem-cell growth patterns in breast cancer[17]**
All fits done using a maximum likelihood method. 95% CIs shown for observed data. (Left column) Model based on assumption that size of stem-cell pool remains constant. (Middle column) Model based on assumption that size of stem-cell pool changes to account for development of breast during puberty. (Right column) Model based on assumption that size of stem-cell pool changes to account for both development of breast during puberty and involution of breast during old age.

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 substantially affected in the smoking cohort. Unfortunately, different studies that have used slightly different methods or datasets have yielded different conclusions (figure 6). For example, Hazelton and co-workers[18] and Schollnberger and colleagues[19] both used TSCE, with dose responsive parameters, to model the incidence of lung cancer in the British Doctors smoking cohort[18] (a cohort of 34439 male doctors who replied to a postal questionnaire regarding smoking habits in 1951). Hazelton and co-workers assume that mutation rates and rates of clonal growth in smokers differed from those of non-smokers according to a general power law. Therefore, a smoker of "d" cigarettes per day, has a mutation rate, "μ_s", related to that of a non-smoker, "μ_n", by $\mu_s=\mu_n(1+ad^b)$, where "a" and "b" are free parameters, inferred from data, and used to calibrate the model. By contrast, Schollnberger and colleagues assume $\mu_s=\mu_n(1+f[d;a,b])$ where $f(d;a,b)=b(1-\exp[-(a/b)d])$. The dose responses of TSCE parameters, as predicted in the two studies, are shown in figure 6. Despite the use of a similar method, Hazelton and co-workers emphasise the effect of smoking on initiation and promotion, whereas Schollnberger's method downplays the relative effect of smoking on the promotion phase. A general theme of multistage modelling is that robust conclusions are difficult to formulate. Of interest, however, is Hazelton and co-worker's ability to predict risk in ex-smokers. The suggestion has long been made that the absence of an abrupt fall in risk after quitting, indicates that the final event triggering clonal expansion of a fully malignant bronchial cell is unaffected by smoking.[20,21] The absence of dose sensitivity in Hazelton and co-workers' progression rate (figure 6) lends some support to this hypothesis.

Cancer 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 previously discussed, this is an idealisation and many cancers stray from the log–log linear relation. Such cancers have an incidence with a changing, rather

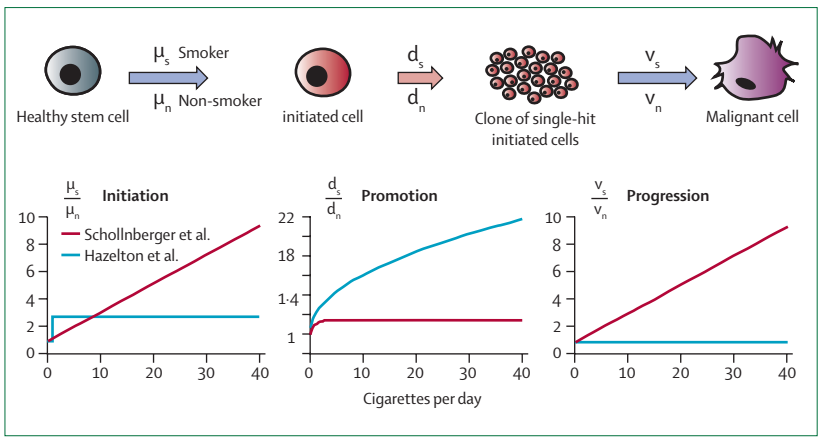

Figure 6: **Incidence of lung cancer in the British Doctors smoking cohort[18] modelled by TSCE with dose-responsive parameters**
Suggested dose responses shown with scheme used by Hazelton and co-workers[18] (blue line) and with scheme used by Schollnberger and colleagues[19] (red line).

than constant, gradient on log–log paper. The gradient at any particular age is referred to as age-specific cancer acceleration.

For example, an idealised log–log linear cancer has a constant acceleration with age. By contrast, breast cancer, in the paradigm of Clemmesen's hook, has a roughly constant acceleration until late in the fifth decade of life 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 idea was exploited by Moolgavkar and colleagues[17] in their model of breast cancer incidence. Prostate cancer incidence is distinguished from that of other common epithelial cancers by a pronounced increase in acceleration before the age of 40 years and an equally pronounced decline thereafter (figure 7A). Frank[22] has used multistage arguments to show how the size and position of such a peak might depend qualitatively on the number, size, and speed of clonal expansions leading to the disease.

Posterior distribution
Indicates knowledge of a given parameter in the light of observed data

How many crucial mutations are in a cancer?

Multistage interpretations of prostate, breast, and lung cancer incidence show that mathematical models can play a useful role in generating plausible theories for qualitatively interesting features of age of onset. However, on the basis of incidence data alone, the selection of a single definitive model from a collection of reasonable alternatives is usually impossible.[23] Furthermore, no single theory can tractably account for every mechanism that might contribute to the disease. In practice, when designing a model, choices should be made about which features are important so that negligible details can be excluded for simplification. For example, although the importance of extracellular factors for establishing the clonal development of a tumour is clear, models used to interpret incidence data have tended to concentrate on heritable changes at the genomic level as drivers of this process. Microenvironmental selection parameters that control the relation between genotype and phenotype are typically accounted for only through fixed clonal growth profiles assumed to associate with a given combination of mutations. An understanding of how such simplifications affect inferences made with the resulting models is crucial. Estimates of the number of mutational stages that lead to cancer, for example, have been shown to be sensitive to the assumptions about clonal growth on which they are made.[10,24] A description of this finding is provided below, where two contrasting models are used to interpret the same data set with different conclusions.

Original multistage model

Armitage and Doll originally suggested that late-onset epithelial cancers contained about six mutations.[5] This estimate was formed by qualitative comparisons between the predicted incidence of the multistage model and the observed incidence. We reapplied the version of their model detailed in the introduction, using a Bayesian algorithm to infer the number of mutations. Data for the incidence of colon cancer were used, recorded from men in England and Wales who were diagnosed between 1960 and 1962.[1] A likelihood function was derived, as before, in terms of "N" (size of the target stem-cell pool), "n" (number of stages), and "μ" (common annual rate of each mutation per cell lineage). By fixing "N" at 10^8 and taking uniform priors of $2 \leq n \leq 9$ and $10^{-8} \leq \mu \leq 10^{-3}$, a **posterior distribution** on "n" (figure 7B) was produced from the prior distributions and likelihood function by integrating out μ as a nuisance parameter.[25] Six mutations was the overwhelming favourite.

Logistic clonal expansion

If clonal expansion is allowed to affect the probability of cancer, the question arises as to how the estimate of the number of mutations would be affected. We constructed a model which incorporated the number of initial cell changes (n_i) that cause a lineage to undergo a logistic clonal expansion and the number of further mutations (n_f) that make any lineage in the clone malignant. Thus, although the initial number of target stem cells is set at 10^8, this number is effectively increased as mutant lineages begin to undergo clonal growth. The distribution of the time until a cell lineage becomes malignant was computed as the sum of two waiting times: the waiting time for the n_i mutations (derived from the generalised multistage model hazard[11]); and the time taken for n_f mutations to arise in any one cell of a logistically growing clone. An expression for this waiting time (t) can also be derived from the generalised multistage model, with the growth profile set at:

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

The composite waiting time was used to construct a hazard function. From this hazard function an expression for the likelihood function was derived according to the

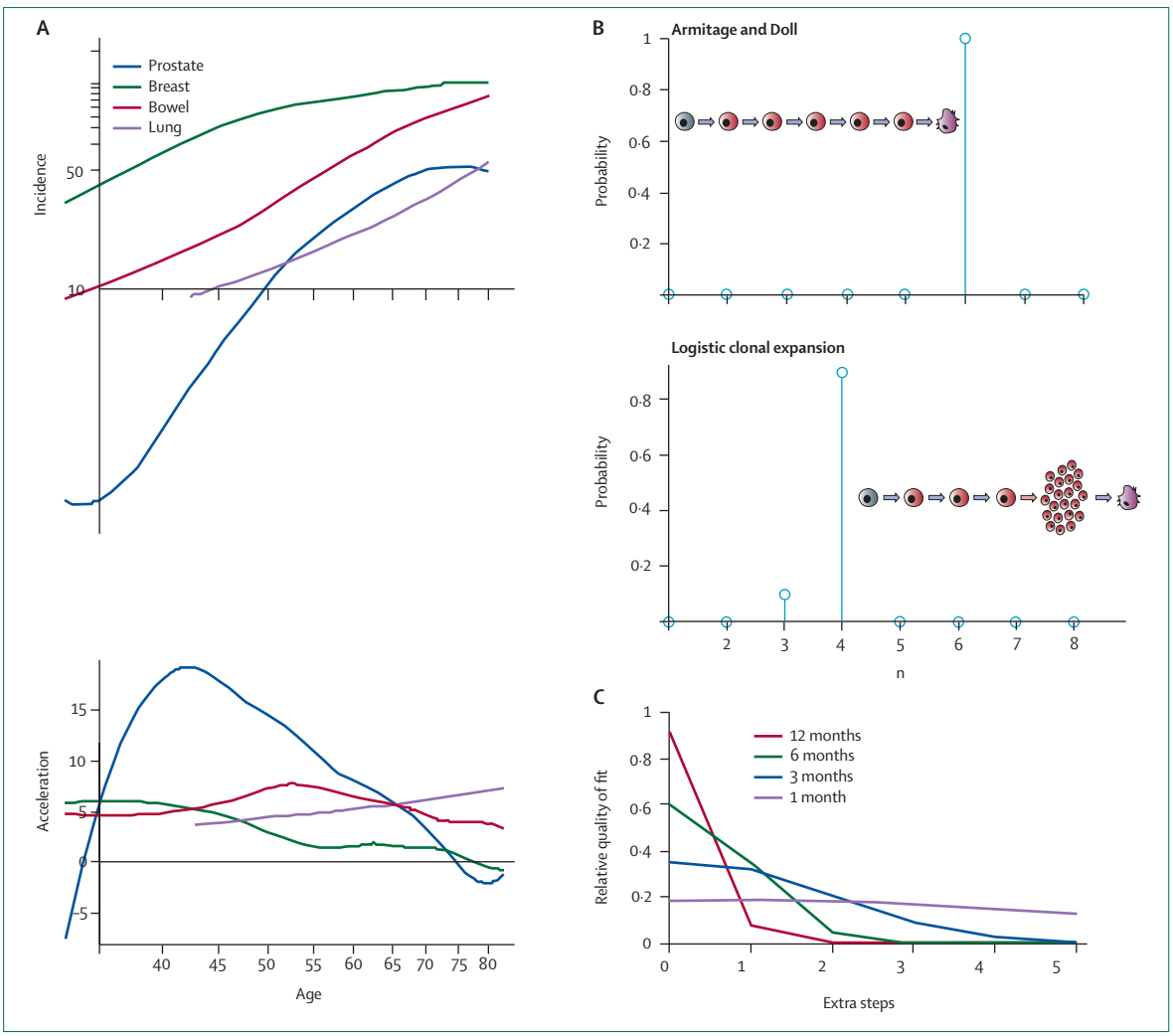

***Figure 7*: Cancer acceleration, mutation numbers, and rate-limiting events**
(A) Incidence and acceleration of the common epithelial cancers. Lung data taken from the CPS2 cohort,[19] all other data from SEER database.[8] (Year of diagnosis=1993–1997). (B) Probability of different numbers of mutations (n) implied by observed bowel-cancer incidence data according to the two models. (C) Relative quality of fit achieved by adding extra stages to an optimised multistage model of bowel cancer.

method of Luebeck and Moolgavkar,[10] and depended on "n_i", "n_f" , "μ" (mutation rate at each locus), "N" (initial size of the stem-cell pool), and two further parameters, "K" and "r" describing the logistic growth of the clone. These growth parameters were fixed at $K=10^6$ and $r=0{\cdot}1$ to produce a slow growing clone taking 50 years to reach a size of 150 proliferating stem cells. "N" was fixed at 10^8 and uniform priors were taken on $1\leq n_i,n_f\leq 8$, and $10^{-8}\leq\mu\leq 10^{-3}$. The posterior distribution at a given parameter vector (n_i,n_f), was calculated by integrating out μ as a nuisance parameter.[21] The posterior distribution at "n" (total number of hits), was calculated as the sum of the posterior distribution at each pair (n_i,n_f) for which $n_i+n_f=n$. The results strongly favour three or four mutations altogether, with only one mutation occurring after the clonal expansion in both cases (figure 7B). This is a moderate change in conclusion to arise from a seemingly minor alteration in assumptions.

Accounting for uncertainty

If more mechanistic details were included in the model, the uncertainty surrounding 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 both have a crucial role in determining incidence.[26] At the tissue level, the number and dynamics of target cells are important modulators of incidence (figure 5). At the population level, genetic heterogeneity in patients can create subpopulations with distinctive risk patterns.[27] When many confounders are present, isolating the effects of mutation numbers is difficult. This fact is hidden by the standard approach to fitting quasi-mechanistic models of carcinogenesis to data, which treats a single, predefined clonal-growth structure as if it were definitive, and takes the optimised state of this

model (ie, the parameter values which give the best fit) as a starting point for making inferences. The result is often to exaggerate the specificity of the conclusion that can be drawn from the data. In their attempts to estimate the number of rate-limiting stages in breast and colorectal cancer, respectively, Zhang and Simon[28] and Luebeck and Moolgavkar[10] both use models that depend on a plausible but narrowly constrained description of clonal growth and mutation. We believe 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 distribution 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.[29]

Evidence from cancer genome projects

Multistage model predictions seem to be inconsistent with evidence generated through the systematic study of cancer genomes. Quantitative analyses of incidence data have conventionally suggested that less than 10 mutational stages exist in, for example, human breast and colorectal cancer. However, evidence produced by a screen of about 13 000 genes in cell lines and xenographs derived from these tumour types suggests that about 14 and 20 genes could be changed by selected mutations in the average colorectal and breast cancer, respectively.[30] These are, however, rough estimates because of the difficulties inherent in distinguishing genuine selected somatic mutations from passenger mutations or artifacts of sequencing and PCR. Nonetheless, a more recent study of about 500 protein-kinase genes,[31] across roughly 200 cancer types, using different methods to identify selected mutations, also suggests that a larger number of functionally changed genes than previously anticipated are operative in many human cancers. Up to now, the apparent discrepancy between low mutation numbers predicted by multistage models and the larger number of changes identified in cancer genomes has been given the standard explanation that only certain crucial mutations restrict the speed at which a cancer is formed. Although they make essential contributions to the cancer phenotype, other mutations occur more quickly than their crucial counterparts—eg, during the clonal development of an established cancer—and are not rate limiting.

Rate-limiting events

The concept of a rate-limiting step originates in the quantitative study of chemical reactions, wherein several precise mathematical definitions have been suggested to identify such a step.[32] Common among these definitions is the idea that changes in the speed of a rate-limiting step will have a substantial effect on the speed of the overall chain of events to which the rate-limiting step belongs. When applied to cancer modelling, the term has been used colloquially and without a strict meaning.

From an incidence-modelling perspective, a rate-limiting step could be defined as a step whose consequences can be seen by looking at age distributions. The relevant question in this context is how fast a mutational or transformational step needs to be before it ceases to be visible in the age-onset pattern. A rough approach to addressing this question is to build a simple model of tumorigenesis; fit this model to the incidence distribution of a specific cancer; augment the model by adding extra mutational stages and note the effect on quality of fit; and increase the common speed of the extra steps until the effect on quality of fit becomes negligible. Figure 7C shows the relative quality of fit to colon cancer incidence (recorded in males from England and Wales between 1960 and 1962) obtained by adding between zero and five extra stages. These stages had an expected duration of 1 year or less—ie, they proceeded at a rate of one or more mutations per cell lineage per year. The initial fit to colon data, before extra stages were added, was achieved using the generalised multistage model[11] with the same mutation rate at each locus and no clonal expansions. A likelihood function was constructed, in terms of "N" (fixed at 10^8), "n", "μ", and an extra parameter, "Ø", representing population heterogeneity. This parameter was included to account for over-dispersion[33] in the registry data, a factor that reduces the speed of a rate-limiting step under the definition we have given. The construction followed that of Luebeck and Moolgavkar[10] except that rather than representing the number of cases within each age group by a Poisson random variable, a variable based on the gamma distribution was assumed[34] so that the variance of cases within each age group was about Ø times the mean. The optimum parameters from this fit were n=6, $\mu=9\cdot3\times10^{-4}$ and Ø=4·89. Extra stages were added as extra steps, with a common fixed mutation rate of one or more mutations per cell lineage per year in the generalised multistage model. Relative quality of fit was measured by a posterior density on the number of these extra stages, n_e. The posterior was calculated by assuming a uniform prior on $0\leq n_e\leq5$ and with the optimum parameters of the initial fit. One or two extra stages of 3 months have a negligible effect, and five extra steps of 1 month also go unnoticed.

When is a mutation rate limiting?

At first it does not seem to matter much that, at the population level, the effects of an event that takes less than 6 months on average to occur are undetectable. 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 affect any target cells in a large population. For example, in

10^8 healthy target cells, if we assume 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 might arise in a substantial precursor lesion or an early-stage cancer. Therefore, how large does such a clone need to be before one or more sequential mutations cease to be rate limiting?

The table shows the expected time for mutations to occur in neoplastic clones of varying sizes. In a large clone of a billion cells (enough to constitute a clinically apparent tumour) two consecutive mutations need not be rate limiting if they occur with the probability of 10^{-6} per cell division or higher. Therefore, a reasonable assumption can be made that many non-rate-limiting mutations occur once the tumour mass has reached a substantial size. This notion 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 metastasising and widespread,[35] by implying that clinical stage depends on non-rate-limiting mutations (table) or other events that occur with high frequency after malignant transformation.

The concept of a rate-limiting step provides a convenient explanation for the discrepancy between multistage model predictions and the emerging picture of cancer genotypes. However, as cancer genome projects implicate larger numbers of genetic changes and this discrepancy becomes more remarkable, the distinction and relation between rate-limiting and non-rate-limiting events is becoming more important. A new challenge for developers of multistage models is to create a quantitative model, in accordance with observed age incidence, which can explicitly incorporate the inactivation, loss, or transformation of 30–40 alleles,[30] with reasonable mutation rates, and rates of clonal growth. To our knowledge, this has, so far, not been achieved.

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. Therefore, the question arises as to how further clarification of age incidence in relation to underlying cell and molecular biology might be achieved. One approach, available for cancers with a well-defined hereditary counterpart, is to compare the incidence of sporadic and familial forms of the disease. For example, retinoblastoma (RB) is a rare childhood cancer of the nervous system initiated by *RB1* inactivation in the developing retina. Before the identification of *RB1*, Knudson[36] showed that children with familial (bilateral) retinoblastoma develop tumours one hit faster than patients with sporadic (unilateral) disease—a finding which strongly suggests a tumour suppressing function for the putative susceptibility gene. A weaker implication, conveyed by the timing of diagnoses, was that retinoblastoma needs only two mutations. However, further genetic defects now seem to have a role,[37,38] although they might not be rate limiting.

Clone size, number of cells	Hits*, n	Mutation rate†‡		
		10^{-5}	10^{-6}	10^{-7}
10^7	1	10^{-4}	10^{-3}	10^{-2}
	2	0·40	3·96	39·64
	3	7·55	75·48	754·78
10^8	1	10^{-5}	10^{-4}	10^{-3}
	2	0·13	1·25	12·53
	3	3·5	34·99	349·94
10^9	1	10^{-6}	10^{-5}	10^{-4}
	2	0·04	0·40	3·96
	3	1·62	16·23	162·34

*The number of specific gene mutations that occur in any one cell of the clone. †The mutation rate is defined as number of mutations per cell division, assuming 100 divisions per year. ‡Because we are using a continuous model of mutation, mutations can occur at anytime and are not limited to fixed points in the cell cycle.

Table: **Expected time lapse in years before one, two, or three specific mutations occur in clones of target cells**

In his study of retinoblastoma incidence, Knudson's primary success was to understand how risk is modulated by gene-carrier status. He successfully applied his method to other childhood cancer syndromes in which the underlying gene defect features as a somatic change in the associated sporadic cancer.[39,40] Because the change in risk caused by a germline mutation is consistent with the role of the functionally deficient or oncogenic protein that is created, additional quantitative investigation could yield useful methods for correlating risk modulation with function. For example, the relative penetrance of colonic carcinoma arising in the context of familial adenomatous polyposis coli, hereditary non-polyposis colorectal cancer, and Li-Fraumeni syndrome, provides interesting evidence regarding the contrasting consequences and relative stage specificity of *APC*, mismatch repair, and P53 mutation, respectively.[35,41,42]

Conclusion

Attempts to fit multistage models to age distributions are highly sensitive to the assumptions about cancer on which they are based. Therefore, if incidence data are used naively, a false sense of confidence is created over the specificity of conclusion that can be drawn. Care needs to be taken to ensure that inferences made are consistent with our current uncertainty of cancer biology and our understanding of it.

We are yet to attain a detailed picture of the connection between incidence and aetiology. A major barrier to progress lies in the fact that the collection of rate-limiting

Search strategy and selection criteria

Data for this personal view were identified by searches of Medline, Current Contents, PubMed, and references from relevant articles using the search terms "multistage cancer", "cancer incidence", and "age distribution". Only papers published in English between 1950 and 2007 were included.

mutations in a given cancer is often a subset of those that are actually crucial to the malignant phenotype. Because the number of such crucial (or at least selected) mutations per cancer genome now seems larger than previously thought, we will need to work harder to understand the relation between rate-limiting and non-rate-limiting genetics. As more genetic data are accumulated in population-based registries, quantitative analysis of risk modulation through germline mutation might prove a useful instrument in helping to decipher this relation.

In the meantime, quasimechanistic models of carcinogenesis need to be carefully applied, and the conclusions they yield should be treated with caution.

Conflicts of interest

The authors declared no conflicts of interest.

Acknowledgments

This work was funded by the Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (London, UK) and a Medical Research Council studentship.

References

1 Doll R, Payne P, Waterhouse J. Cancer incidence in five continents, volume.1. Geneva: Union Internationale contre le Cancer; 1966.
2 Haenszel W. Cancer mortality among the foreign-born in the United States. *J Natl Cancer Inst* 1961; **26:** 37–132.
3 Colditz GA. Epidemiology—identifying the causes and preventability of cancer? *Nat Rev Cancer* 2006; **6:** 75–82.
4 Parkin DM. Evolution of the population-based cancer registry. *Nat Rev Cancer* 2006; **6:** 603–12.
5 Armitage P, Doll R. The age distribution of cancer and a multistage theory of carcinogenesis. *Br J Cancer* 1954; **8:** 1–12.
6 Nordling CO. A new theory on the cancer-inducing mechanism. *Br J Cancer* 1953;**7:** 68–72.
7 Cancer Research UK. CancerStats. http://info.cancerresearchuk.org/cancerstats (accessed Sept 20, 2007).
8 National Cancer Institute. Surveillance, Epidemiology and End Results. http://seer.cancer.gov/ (accessed Sept 20, 2007).
9 Potten CS, Booth C, Hargreaves D. The small intestine as a model for evaluating adult tissue stem cell drug targets. *Cell Prolif* 2003; **36:** 115–29.
10 Luebeck G, Moolgavkar SH. Multistage carcinogenesis and the incidence of colorectal cancer. *Proc Natl Acad Sci USA* 2002; **99:** 15095–100.
11 Little MP. Generalisations of the two-mutation and classical multi-stage models of carcinogenesis fitted to the Japanese atomic bomb survivor data. *J Radiol Prot* 1996; **16:** 7–24.
12 Seshadri R, Kutlaca RJ, Trainor K, Matthews C, Morely AA. Mutation rate of normal and malignant human lymphocytes. *Cancer Res* 1987; **47:** 407–09.
13 McCombs RS, McCombs RP. A hypothesis on the causation of cancer. *Science* 1930; **72:** 423–24.
14 Brues AM. Critique of mutational theories of carcinogenesis. *Acta Unio Int Contra Cancrum* 1960; **16:** 415–17.
15 Moolgavkar SH, Dewanji A, Venzon D. A stochastic two-stage model for cancer risk assessment. I. The hazard function and the probability of tumor. *Risk Anal* 1988; **8:** 383–392.
16 Little MP, Wright EG. A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data. *Math Biosci* 2003; **183:** 111–34.
17 Moolgavkar SH, Day NE, Stevens RG. Two-stage model for carcinogenesis: Epidemiology of breast cancer in females. *J Natl Cancer Inst* 1980; **65:** 559–69.
18 Hazelton WD, Clements MS, Moolgavkar SH. Multistage carcinogenesis and lung cancer mortality in three cohorts. *Cancer Epidemiol Biomarkers Prev* 2005; **14:** 1171–81.
19 Schollnberger H, Manuguerra M, Bijwaard H, et al. Analysis of epidemiological cohort data on smoking effects and lung cancer with a multi-stage cancer model. *Carcinogenesis* 2006; **27:** 1432–44.
20 Armitage P. Multistage models of carcinogenesis. *Environ Health Perspect* 1985; **63:** 195–201.
21 Peto J. Cancer epidemiology in the last century and the next decade. *Nature* 2001; **411:** 390–95.
22 Frank SA. Age-specific acceleration of cancer. *Curr Biol* 2004; **14:** 242–46.
23 Michor F, Iwasa Y, Nowak MA. The age incidence of chronic myeloid leukemia can be explained by a one-mutation model. *Proc Natl Acad Sci USA* 2006; **103:** 14931–934.
24 Calabrese P, Tavare S, Shibata D. Pretumor progression. *Am J Pathol* 2004; **164:** 1337–46.
25 Sivia DS. Data analysis: a Bayesian tutorial. Oxford: Oxford University Press, 1996.
26 Ilyas M, Straub J, Tomlinson IPM, Bodmer WF. Genetic pathways in colorectal and other cancers. *Eur J Cancer* 1999; **35:** 335–51.
27 Peto J, Mack TM. High constant incidence in twins and other relatives of women with breast cancer. *Nat Genet* 2000; **26:** 411–14.
28 Zhang X, Simon R. Estimating the number of rate limiting genomic changes for human breast cancer. *Breast Cancer Res Treat* 2005; **91:** 121–24.
29 Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Stat Sci* 1999; **14:** 382–417.
30 Sjoblom T, Jones S, Wood LD, et al. The consensus coding sequences of human breast and colorectal cancers. *Science* 2006; **314:** 268–74.
31 Greenman C, Stephens P, Smith R, et al. Patterns of somatic mutation in human cancer genomes. *Nature* 2007; **446:** 153–58.
32 Turanyi T. Sensitivity analysis of complex kinetic systems. Tools and applications. *J Math Chem* 1990; **5:** 203–48.
33 Little MP, Li G. Stochastic modelling of colon cancer: is there a role for genomic instability? *Carcinogenesis* 2007; **28:** 479–87.
34 Fay MP, Feuer EJ. A semi-parametric estimate of extra-Poisson variation for vital rates. *Stat Med* 1997; **16:** 2389–401.
35 Calabrese P, Mecklin J, Jarvinen HJ, Aaltonen LA, Tavare S, Shibata D. Numbers of mutations to different types of colorectal cancer. *BMC Cancer* 2005; **5:** 126.
36 Knudson AG. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA* 1971; **68:** 820–23.
37 Chen D, Pajovic S, Duckett A, Brown VD, Squire JA, Gallie BL. Genomic amplification in retinoblastoma narrowed to 0.6 megabase on chromosome 6p containing a kinesin-like gene, RBKIN. *Cancer Res* 2002; **62:** 967–71.
38 Lillington DM, Kingston JE, Coen PG, et al. Comparative genomic hybridization of 49 primary retinoblastoma tumors identifies chromosomal regions associated with histopathology, progression, and patient outcome. *Genes Chromosomes Cancer* 2003; **36:** 121–28.
39 Knudson AG, Strong LC. Mutation and cancer: a model for Wilms' tumor of the kidney. *J Natl Cancer Inst* 1972; **48:** 313–24.
40 Knudson AG, Strong LC. Mutation and cancer: neuroblastoma and pheochromocytoma. *Am J Hum Genet* 1972; **24:** 514–32.
41 Kinzler KW, Vogelstein B. Landscaping the cancer terrain. *Science* 1998; **280:** 1036–37.
42 Frank SA. Age-specific incidence of inherited versus sporadic cancers: a test of the multistage theory of carcinogenesis. *Proc Natl Acad Sci USA* 2005; **102:** 1071–75.