Models of cancer incidence

Vibishan B.

July 8, 2018

1 Introduction

Accumulation of somatic mutations is a central process in carcinogenesis, and at least five distinct mutations are required to initiate a potential malignancy. The average somatic mutation rate is of the order of 10^{-6} per cell per generation [1–3], making it highly unlikley that any five cancerous mutations co-occur in the cell by random chance ($(10^{-6})^5 = 10^{-30}$). Given this remoteness, it follows that cancer must be a relatively rare occurence, with a fixed threshold age above which cancer is certain to occur. Both these predictions can be proven wrong based on actual incidence of cancer in the human population in recent years [4]. An elevated mutation rate alone is therefore insufficient to completely explain cancer incidence.

It is more realistic for mutations to accumulate sequentially through a process of clonal evolution, in which a mutation with a positive growth advantage would allow the mutant cells to grow and replace non-mutants. With sufficient expansion, this mutant population could then similarly acquire enough mutations for cancer to occur. This is the current paradigm of **somatic cancer evolution** [5], and significant progress has been made to develop this basic view of intra-cellular competition into a more substantive evolutionary account of cancer processes [6–8]. These perspectives have greatly improved the study of cancer, and opened new avenues of research focused on applying existing evolutionary theory to cancer biology.

However, the conventional somatic evolution picture has consistently ignored an important aspect of cancer incidence that concerns population heterogeneity. By assuming clonality between individuals, the hypothesis concludes that any given somatic mutation must have the same all-time advantage wherever it occurs, leading inevitably to cancer. But this assumption is not necessary from a physiological viewpoint; tissues and micro-environmental factors could vary widely between individuals of the same species, depending on genetic and life style differences. This implies that the exact outcome of an identical mutation need not necessarily be identical amongst members of a population. Selection on somatic mutations could instead be conditional on factors that reach *beyond* the mutation itself, to the local micro-environment that sets the stage for cellular competition. The extra-cellular matrix (ECM), for instance, is a complex combination of proteins and carbohydrates that is secreted by almost all somatic cells. By nature, it is highly dynamic, and has been shown to regulate a diverse array of cellular functions through feedback signaling. More importantly, the exact composition of the ECM differs between members of the same species, mostly in response to environmental factors, which could potentially result in a heterogeneous population. Given these confounding observations, it remains unclear whether or not the assumptions of the clonal selection theory can be accepted without a doubt.

Somewhat tangentially, it is also being recognized that somatic evolutionary processes can be diverse in modality [], and four such general modalities have been putatively identified, with the understanding that none of them is necessarily operating in isolation within a tumour. These are the linear, branching, neutral and punctuated modes of tumour evolution, and as such, represent different dynamics of mutation accumulation. It is therefore also interesting to explore how basic parameters like the mutation rate and cell number, as well as nuanced parameters like population heterogeneity, vary in their effects depending on the evolution mode.

To that end, we first develop a simple simulation of somatic mutagenesis, in which cells within an individual grow and accumulate mutations sequentially, ultimately leading to cancer. Since the argument here concerns incidence of cancer at the population level, we carry out the individual-level stochastic mutation process in a population of 100000 individuals to generate predictions that can be compared with epidemiological data [4]. Through this comparison, we are able to show that an elevated mutation rate and clonal selection are insufficient to completely explain cancer incidence, argue that a complete model of carcinogenesis must explicitly account for heterogeneity in mutational outcomes in the population, and then continue to build predictions for other parameters for other modes of evolution.

2 Methods

2.1 The linear process

The general linear process considers mutation accumulation to occur sequentially; one cell from the initial non-mutant cell population gets a mutation that, by chance, allows the mutant to outcompete its non-mutant neighbours. This advantage leads to complete competitive exclusion of the non-mutants, with mutants replacing them in the entire niche. As this replacement progresses, one of the single mutant cells gets a second mutation that in turn, allows it to outcompete the single-mutant population, leading to a second cycle of replacement by the double-mutant cells. Over time, this process ultimately produces cells with enough accumulated mutations to become cancerous.

We begin by considering a collection of 100000 individuals, each with a poulation of n stem cells at steady state logistic growth. If p is the genome-wide mutation rate per cell per generation, then the probability that at least one mutation occurs per generation in the population of n cells, based on simple probability, is given by:

$$p_{mut} = 1 - (1 - p)^n \tag{1}$$

The above probability is used to simulate a stochastic mutation process; if the value p_{mut} exceeds a randomly-generated value between 0 and 1, a mutation is considered to have occurred. This gives rise to a new mutant population of size $m = 1$, which continues to grow logistically at a rate g relative to the normal stem cells. Once a mutant population is initiated, p_{mut} is calculated iteratively for the mutant population by replacing n with m in equation 1. At some point, one of the cells in this population stochastically acquires a second mutation, which creates a new doubly mutant population. The growth of these cells is now followed by setting the value of m back to 1, and allowing them to grow at a rate, g. It must be noted here that this growth rate

is now relative to the single mutation population rather than the normal stem cell population. For simplicity, we assume that the carrying capacity for all cell populations is the initial steady state cell number, n.

Arbitrarily, we set a mutation threshold of five; five such mutation events are considered to constitute a case of cancer, which leads to the death of the individual. We intentionally neglect post-diagnosis clinical progression as they are relatively independent of the dynamics of mutation accumulation. We limit the natural lifespan to 90 years per individual, with 100 cell division cycles per year []; this limit reflects availability of epidemiological data of cancer incidence across age groups [4]. Each simulated death therefore occurs when five mutations are accumulated, or at the age of 90 years, whichever happens first. We then repeat the entire process for the next, starting from the zeroth generation. We record each instance of cancer and the corresponding age at which it occurs, to get an age-wise count of cancer incidence.

For the purposes of this simulation, we define cancer incidence in terms of both the crude rate, or the normalized fraction, of cancer in the population per 100000 individuals at each age, as well as the age-adjusted rate of incidence. This calculation is explained in further detail in Section 2.1.3.

Codes for all simulations, and most data analysis and plotting are written in Python, and executed on the Jupyter Notebook platform. Data is exported and stored in Excel format, and record keeping is managed on Github.

2.1.1 Randomizing p and n

The description above is one way of running the simulation whose primary output is the curve of crude rate vs age, and the age-adjusted incidence rate calculated for the population based on the crude rates (Section 2.1.3). It is also possible to randomize the values of p and n within a pre-defined range across the population, in order to clearly demosntrate the effects of p and n on the dynamics of mutation accumulation.

We start with uniform distributions for both p and n as no data exist to suggest *a priori*, a specific distribution for either. p is randomized in the range $[10^{-9}, 10^{-6}]$, while n, in the range $[10^{7}, 10^{10}]$ []. Following randomization, we run each inidividual simulation according to the mutation threshold or lifespan specified above. When cancer does occur in an individual, we record the corresponding age of incidence along with that individual's mutation rate and cell number, and look for correlations between them. This correlation carries information about the effect of a variable on the temporal progression of mutation accumulation; faster progression leads to faster accumulation of mutations, and therefore cancer incidence at an earlier age. For instance, higher p could accelerate mutation accumulation, and would therefore have a negative correlation with age of cancer incidence.

A general feature about the effects of n and p can be expected *a priori*, stemming from equation 1 which is used in stochastic mutagenesis. The quantity, p_{mut} in equation 1 has a saturating trend with both n and p; beyond some threshold value, p no longer has an effect on p_{mut}, and the same is true of n. We can therefore expect a similar saturating relationship between p and n in the incidence parameters in the model predictions.

2.1.2 Determination of g

Since we use g as the growth rate of mutant cells relative to their non-mutant neighbours, it is essentially a proxy for selection acting on mutations in an individual. Whether this selection is assumed to be identical in the population, or heterogeneity in mutation effects is allowed has important consequences for model predictions. g is therefore an excellent parameter to test the context dependence of somatic mutations.

Two distinct possibilities arise in this case:

- Context-independent incidence

  Under this assumption, individuals in the population are clonal, and similar mutations would hence have the same selection advantage. Therefore, in any given run of the simulation, we set each individual to have the same value of g for every mutation, across all 100000 individuals. As each individual starts off with the same n and p, the first mutation ocurring in each individual is essentially identical. Each subsequent mutation has the same relative growth rate given by g, as mentioned above. This homogeneity in the population is important as it is expected to make g the key variable in determining cancer onset in the population. Since all individuals are identical, a higher g leads to faster cell growth and progression, and earlier cancer onset.

- Context-dependent incidence

  In this case, we model g as a normally-distributed random variable, as this view explicitly accounts for a heterogeneous population. For each individual, a different value of g is chosen randomly based on the underlying distribution. Therefore, mutations with identical p_{mut} values can still differ in terms of their selective advantage. Much like with randomizing p and n therefore, g values leading to cancer in the population can also be correlated with the age at which cancer occurs. This is in keeping with the earlier expectation that higher g values lead to faster progression and earlier cancer onset. The randomization of g therefore gives two additional parameters in the model; the mean, μ_g, and the standard deviation, σ_g. We test these effects by varying μ_g over $[0, 2]$ while $\sigma_g = 1$, and varying σ_g over $[0.5, 4]$ while $\mu_g = 0.5$. Interestingly, when n, p and g are randomized together, it is also possible to consider their relative effects on the temporal features of cancer. We can therefore demonstrate by correlations of onset age with any parameters, and residuals of onset age with any other parameter, the extent to which each affects the onset age itself.

  At this stage, we choose to maintain the same relative growth rates for subsequent mutations occurring within an individual, without considering intra-tumour heterogeneity. Although this would be a simple modification to the simulation, it is not directly relevant to our argument of conditional selection at the population level, and we do not expect it to change our conclusions significantly. For the sake of completeness, we also use the uniform and triangular distributions for g to verify that the qualitative relationship between cancer incidence and age is not sensitive to the exact shape of the distribution (data not shown).

2.1.3 Normalized cancer incidence

Depending on the distribution of g and the sequence in which the values occur, the fraction of the population surviving to a particular age could vary between subsequent runs. Moreover, absolute

count of cases always under-estimates actual incidence at higher age groups as the suriving population size is continuously decreasing. To correct for this, we calculate the normalized cancer incidence for each age by diving the number of observed cases by the size of the suriving population at that age. This gives us the expected fraction of cancer for each age, and by multiplying this fraction by 100000, we obtain the crude rate per 100000 for that age.

It is also common practice in epidemiology to calculate the age-adjusted incidence rate for any given population [4]. This is a weighted average of the crude rate, with the weights coming from the demorgaphic age distribution of a standard population. Since we do not have access to the exact age distribution of the simulated population, we extrapolate cancer incidence in different populations on to a known population structure, which allows for comparisons between them. Essentially, this eliminates underlying population structure as a source of variation while comparing the cancer incidence of several populations.

2.1.4 Effect of age on g

It has been argued that somatic maintenance declines with age[9], and from a life history perspective, this does not seem entirely implausible. Evolutionary theories of ageing describe a steady decrease in selection pressure in the later stages of life beyond the age of peak fecundity. The "carcinogencitiy" of mutations could therefore increase with age as suppression mechanisms decline in efficiency. We model this by adding various monotonously increasing functions to the initial value of g at each age, such that the value of g increases continuously within the lifespan of every individual starting from the initial value obtained from the random number generator, as dictated by the age function.

3 Results

3.1 The linear model: context-independent case

3.1.1 Effect of g

We begin with the linear model's context-independent case, which makes the simplest set of assumptions among all the models. As Figure 1 shows, a clonal population leads to sharp transitions in cumulative incidence, which the sum of all crude rates up to each age. Since mutational processes are identical across individuals, cancer onset is also practically identical across the population. The age at which this transition occurs, and the rate of this transition, both depend on g; higher g leads to cancer earlier in life, as well as a sharper transition to the cancerous state. It is noteworthy that lower values of g lead to slower transitions, which could be an artifact of the discrete growth process. Given a certain length of time required for the mutation threshold to be exceeded, lower g cell populations approach this threshold with smaller step sizes. The lower step size makes the cell populations more sensitive to stochastic variation in the exact age at which the initiating mutation occurred. This then represents a general property of g and its effect on the cumulative incidence curve; higher g leads to earlier cancer onset, and earlier cancer onset is also accompanied by faster approach to the cancerous stage.

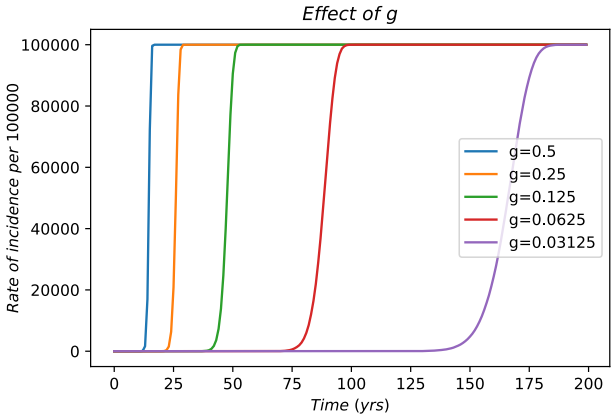


Figure 1: Effect of g on cumulative incidence of cancer in a linear process simulation, with $p =$ and $n =$ for the entire population. The curve reaching 100000 represents 100% incidence. These simulations were not constrained to the physiological lifespan of 90 years.

3.2 Effects of n and p

As suggested in Section 2.1.1, saturation effects of both n and p can be observed in the age-adjusted incidence (Figure 2), as well as crude and cumulative rates of incidence (Figure 3). In Figure 2, the lowest rows in the first two panels clearly show that the age-adjusted rate (AAR) for $n = 10^{10}$ does not depend upon the value of p. This apparently linear reponse of AAR with n becomes more confused across the panels. However, interpretation of this trend must also involve the age curves in Figure 3, which show that n and p mutually compensate for each other's effects even for a given value of g; there are either so many cells that the mutation rate becomes unimportant to mutation accumulation, or the mutation rate is high enough that mutation accumulation remains unaffected even when cell number is low. Interestingly, the strength of this reciprocal compensation is dependent on g; for instance, from the top row to the bottom in Figure 3(b), the cumulative incidence curve for $p = 10^{-6}$ changes much more significantly across a row for lower g.

This could again be explained based on the discrete logistic growth pattern; higher g would naturally lead to higher step size, but so would higher n, which is also the logistic carrying capacity. As with g, the larger step size shifts incidence to earlier ages. The same explanation can be applied to the changing effects of n with g; for high enough g, the cell populations would exceed the threshold size for a mutation to occur for most values of n; by how much this threshold is exceeded is immaterial.

Randomization of n and p within the population allows us to isolate their effects on the temporal progression of mutations within individuals. Figure 4 shows clearly that the value of n does not determine the age of onset of cancer, while p is significantly correlated to the age of onset. Even under the parsimonious linear fit, p explains more than 50% of the variance in the age of onset, supporting the notion that the mutation rate is at least as important as g in determining how fast cancer occurs in an individual, while cell number plays a marginal role, if any, in determining the age of onset.

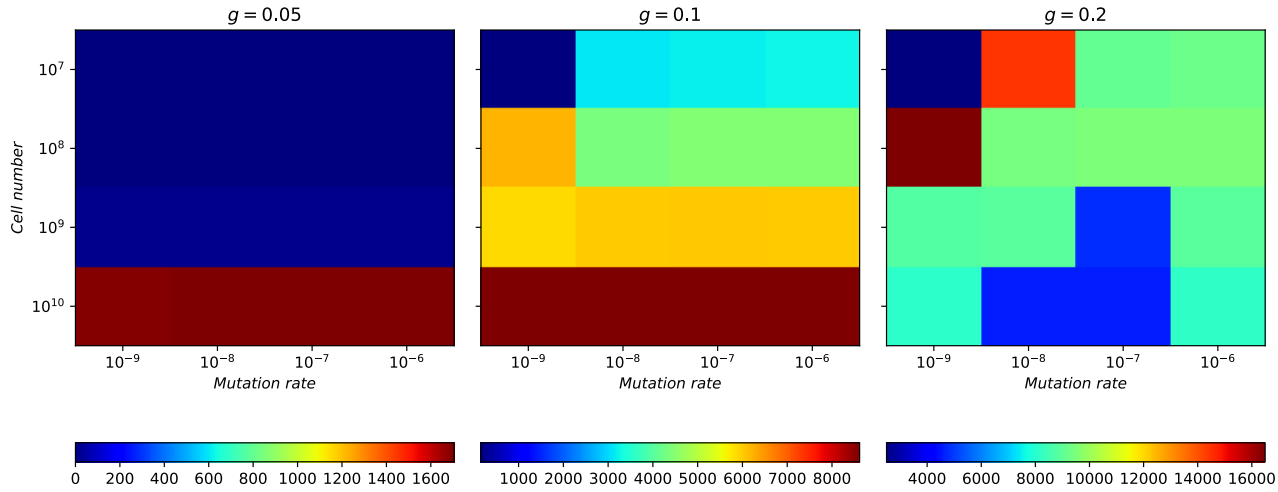


Figure 2: Age-adjusted incidence rates for the linear context-independent model. Each grid represents results for one value of g, p values on the horizontal axis increase from right to left, and n values on the vertical axis increase from top to bottom. Individual grids have their colourbars to scale.

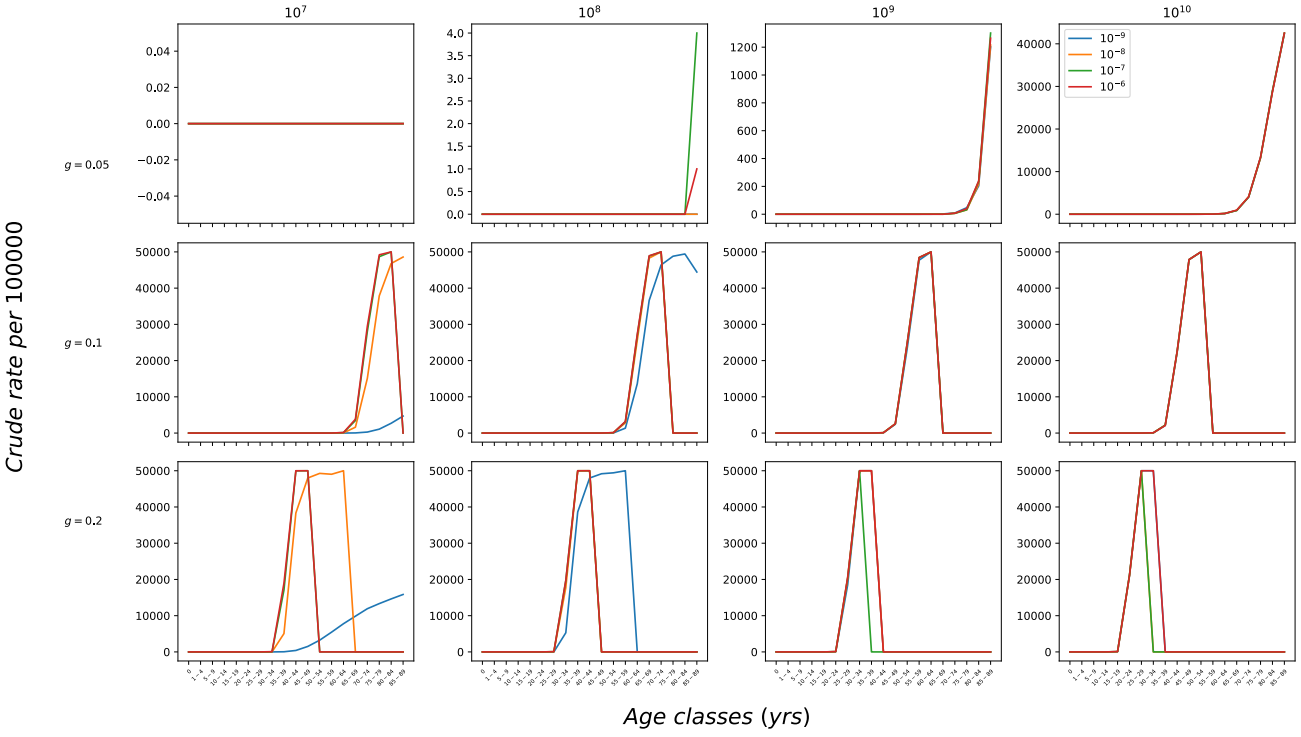


Figure 3a: Crude incidence rate per 100000; each panel contains crude rate curves for 4 values of p, each column represents one value of n, and each row, one value of g.

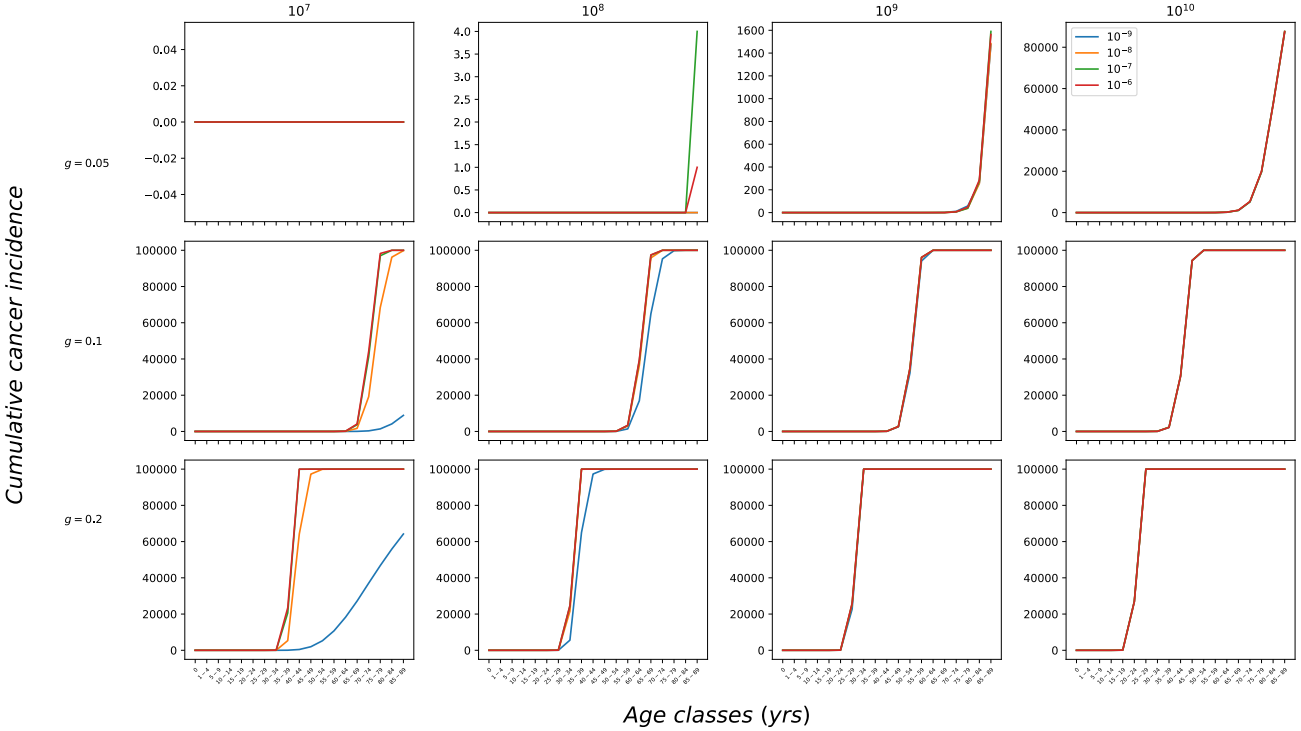
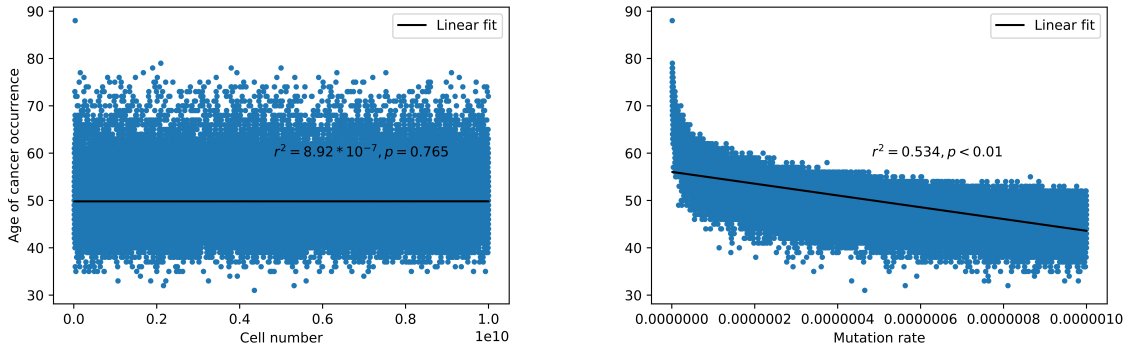


Figure 3b: Cumulative incidence rates corresponding to the crude rates in Figure 3.



(a) (b)

Figure 4: Correlations of age of cancer onset with (a) cell number, and (b) mutation rate, with corresponding linear fits. Ranges of n and p in the text.

3.3 The linear model: context-dependent case

An important difference between the context-dependent and -independent cases lies in the cumulative rate curve for the former, which does not always tend towards 100% incidence. This reflects the fact that while a clonal population will become completely cancerous given enough time, a heterogeneous population is free of such a requirement. As Figure 5(a), there can be pronounced decreases in the crude rate at later ages, along with gradual approaches to 100% incidence or saturation at lower incidence as in Figure 5(b). This difference can be attributed only to the addition of g as a normally-distributed random variable, as everything else between the two cases is the same.

However, there are also broad similarities between the context-dependent and -independent cases, beginning with the effects of n and p on each other. This mutual saturation threshold is clear from the crude and cumulative incidence curves, as well as from the AAR. These similarities notwithstanding, the effects of g can now be explored in much better through its distribution, both in terms of μ_g and σ_g.

As Figures 5(a) and (b) show, μ_g does not change the n-p interaction, but shifts it to a region of higher incidence; this is apparent from the similarity between a low n-high μ_g curve to a higher n-lower g curve, and the fact that the lowest p incidence curve begins as an outlier at n_1, and reaches saturation subsequently, for every value of μ_g. It is important to note that in Figure 5, the scale of the y-axis changes drastically across each row, which is explained by the gaps between the corresponding n values.

In terms of the effect on the n-p interaction, varying σ_g does not produce qualitatively different results from that in Figure 5. An interesting feature however, is the decreasing age of cancer onset for large values of σ_g (Figure 6). For large variance in g, cancer incidence not only increases, but shifts strongly to earlier ages. A potential explanation could be that the random variable g can have drastically large values for large σ_g, and as already been shown for the context-independent case (Figure 1), this will decrease the age of cancer onset.

Remarkably, however, this shift to earlier cancer onset is not accompanied by a rapid approach to a 100% cancerous state. Indeed, where for $g \in N(2, 1)$, the cumulative incidence rate closely approached 100000, for $g \in N(0.5, 4)$, it begins to show saturation around 50000, potentially indicating that large σ_g increases the likelihood of both large positive and large negative values of g, thus bringing down the total incidence.

Randomization of p and n along with g allows us to draw some inferences on the role each plays in the temporal dynamics of mutation accumulation. Figure 7 reiterates some of what was already known from the context-independent case in Figure 4; cell number is not signficantly associated with the age of cancer onset, but interestingly, the correlation on onset age with p has rather weakened compared to the context-independent case, explaining only about 1.2% of the variance in onset age. Instead, g has a strong negative association with age of cancer onset. Assuming a linear relationship, we calculate the residuals of onset age with g, E_g, which is the variance in onset age that is entirely explained by g. These residuals are better correlated to the mutation rate, p (Figure 7, although the correlation with n did not improve (data not shown). Although the assumption of linearity probably under-estimates the residual variance, the two associations (Figures 7(a) and (b)) show clearly that primarily g, and to some extent, p, are both key determinants of the age of cancer onset, with n playing almost no role at all.

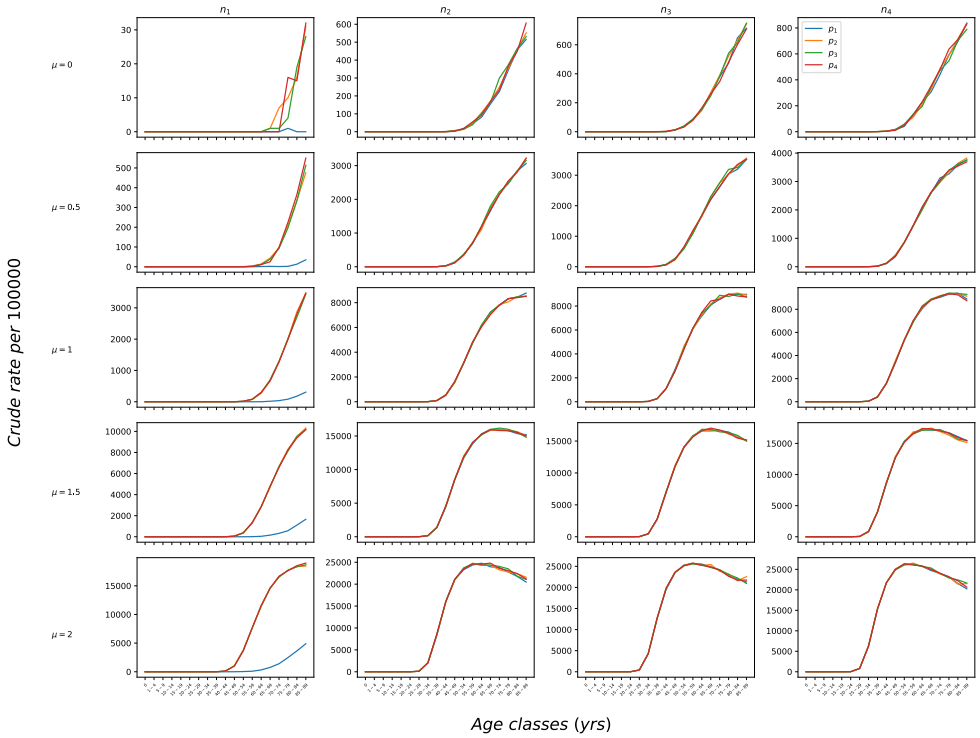


Figure 5a: Crude incidence rates in the linear context-dependent case. For all μ_g, $\sigma_g = 1$; $n_1 = 10^7$, $n_2 = 3 * 10^9$, $n_3 = 7 * 10^9$, $n_4 = 10^{10}$, $p_1 = 10^{-9}$, $p_2 = 3.3 * 10^{-7}$, $p_3 = 6.7 * 10^{-7}$, $p_4 = 10^{-6}$

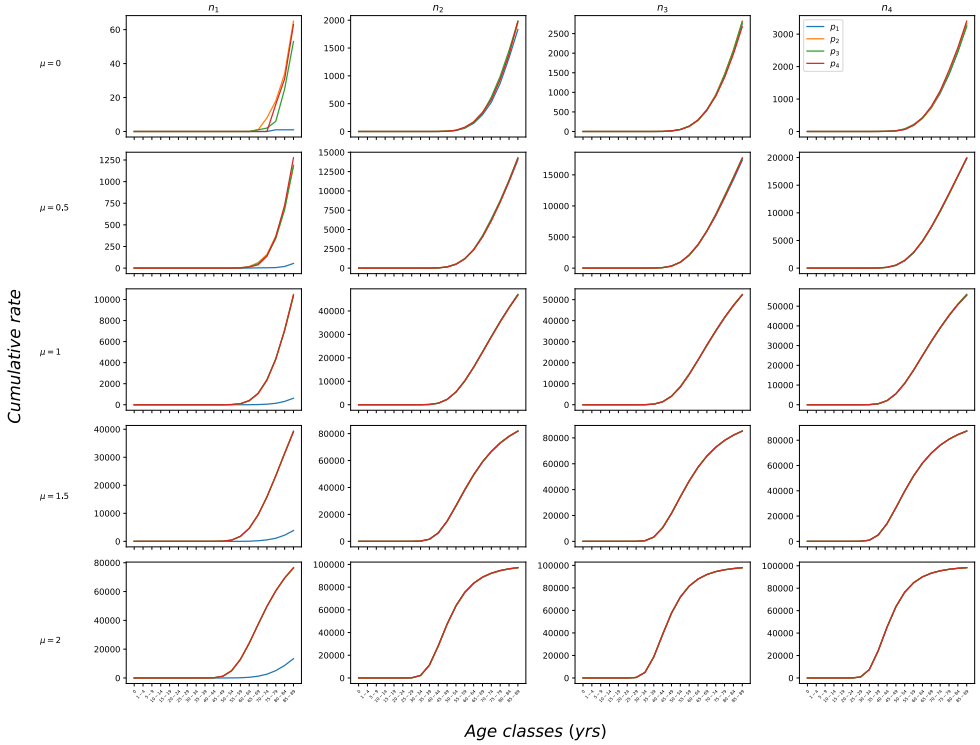


Figure 5b: Cumulative incidence curves correspoding to the crude rates in Figure 5(a). All other parameter values are the same.

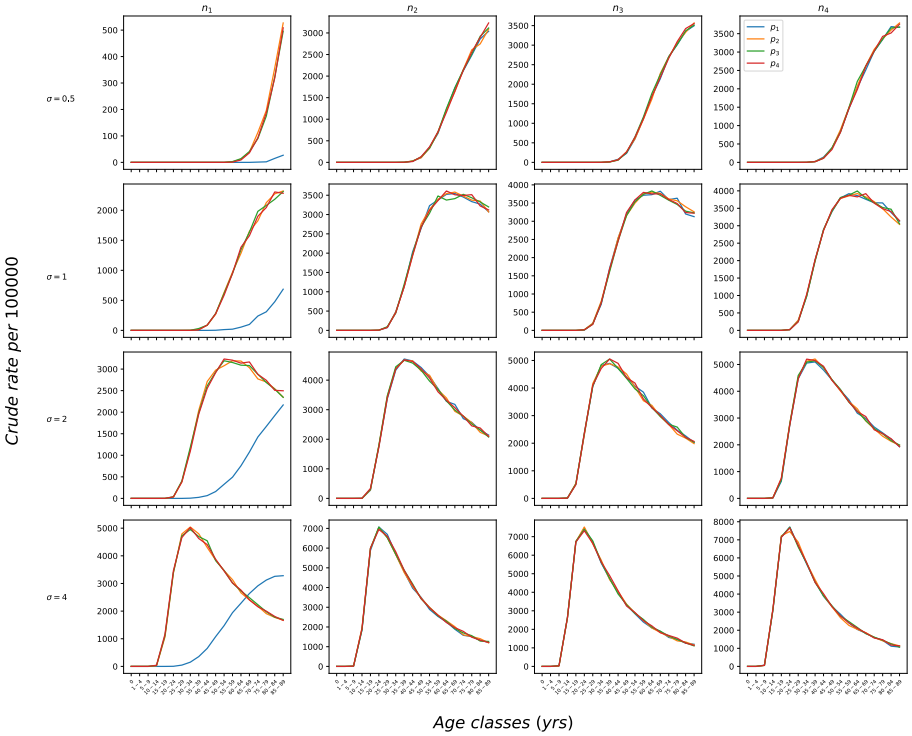


Figure 6a: Crude incidence rates in the linear context-dependent case. For all σ_g, $\mu_g = 0.5$; $n_1 = 10^7$, $n_2 = 3 * 10^9$, $n_3 = 7 * 10^9$, $n_4 = 10^{10}$, $p_1 = 10^{-9}$, $p_2 = 3.3 * 10^{-7}$, $p_3 = 6.7 * 10^{-7}$, $p_4 = 10^{-6}$

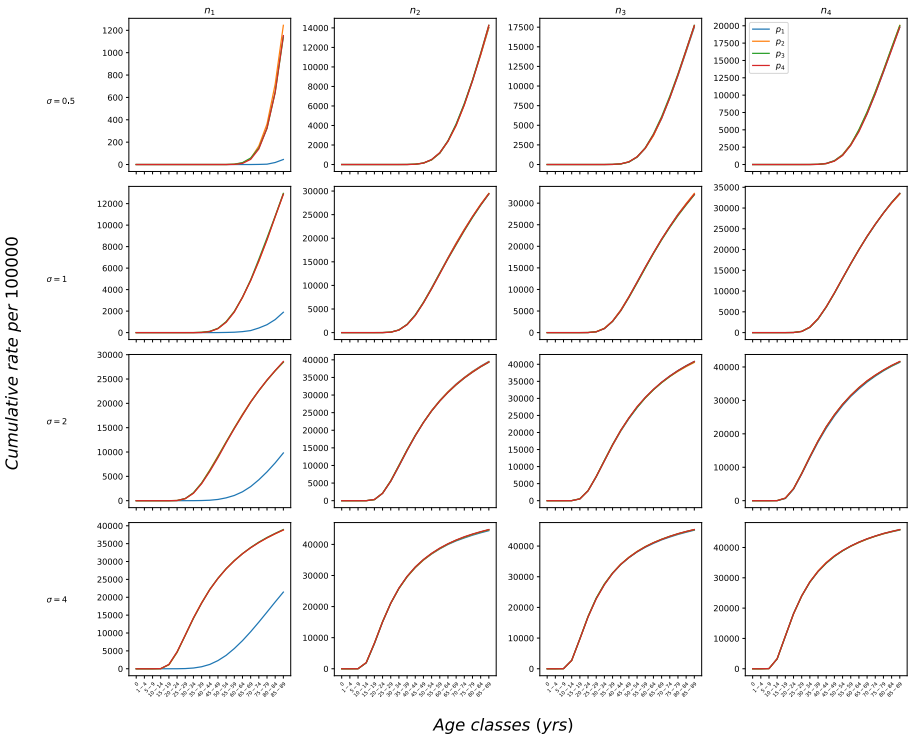


Figure 6b: Cumulative incidence curves correspoding to the crude rates in Figure 6(a). All other parameter values are the same.

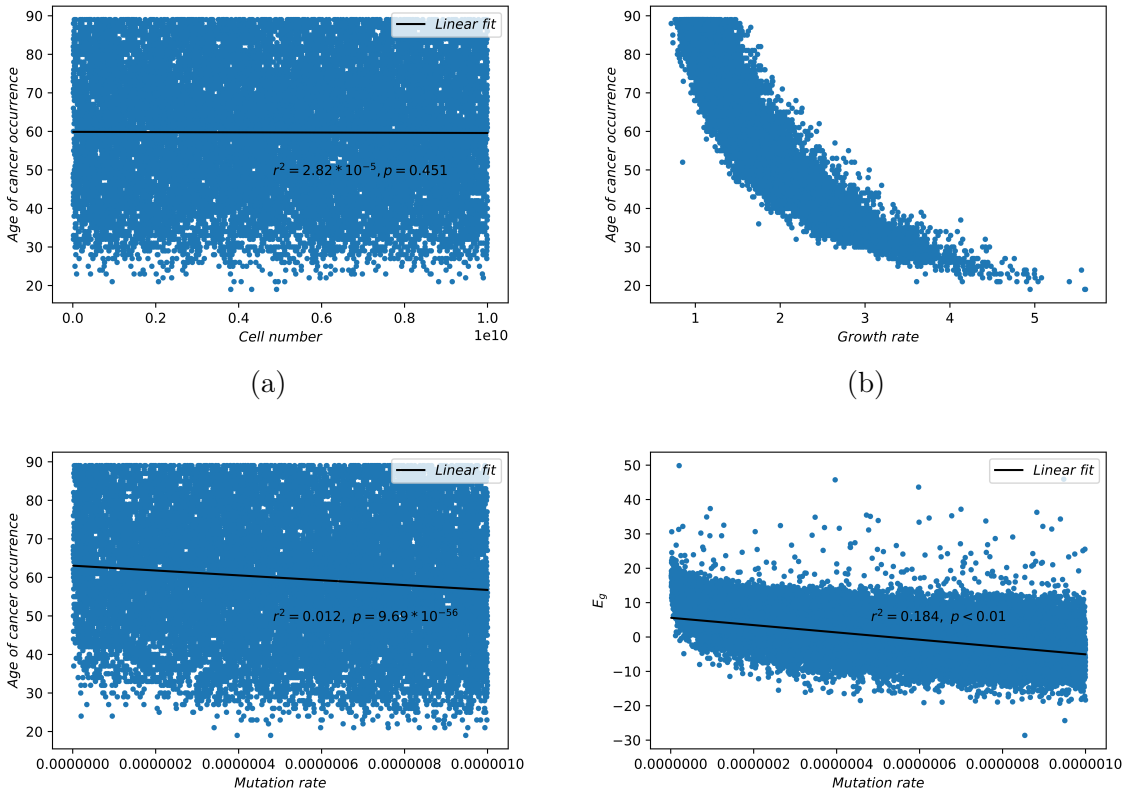


Figure 7: Correlations of age of cancer onset with (a) cell number, n, (b) growth rate, g, and (c) mutation rate, p, and (d) correlation of residuals from a linear fit with g of age of cancer onset, with mutation rate, p.

References

[1] C. Tomasetti and B. Vogelstein. "Variation in cancer risk among tissues can be explained by the number of stem cell divisions". In: *Science* 347.6217 (2015), pp. 78–81. arXiv: 1260825.

[2] Michael Lynch. "Evolution of the mutation rate". In: *Trends in Genetics* 26.8 (2010), pp. 345–352.

[3] Ram Seshadri et al. "Mutation Rate of Normal and Malignant Human Lymphocytes". In: *Cancer research* 47 (1987), pp. 407–409.

[4] American Cancer Society. "Cancer Facts & Figures 2016". In: *Cancer Facts & Figures 2016* (2016), pp. 1–9. arXiv: NIHMS150003.

[5] P. Nowell. "The clonal evolution of tumor cell populations". In: *Science* 194.4260 (Oct. 1976), pp. 23–28.

[6] C. Athena Aktipis and Randolph M. Nesse. "Evolutionary foundations for cancer biology". In: *Evolutionary Applications* 6.1 (2013), pp. 144–159.

[7] C Athena Aktipis et al. "Life history trade-offs in cancer evolution." In: *Nature reviews. Cancer* 13.12 (2013), pp. 883–92. arXiv: NIHMS150003.

[8] Marco Gerlinger et al. "Cancer: Evolution Within a Lifetime". In: *Annual Review of Genetics* 48.1 (Nov. 2014), pp. 215–236.

[9] Andrii I. Rozhok and James DeGregori. "The Evolution of Lifespan and Age-Dependent Cancer Risk". In: *Trends in Cancer* xx (2016), pp. 1–9.