AMATH 536 Project
Analysing the link between cancer risk and number of stem cell divisions through simulation

Katie Johnston

Spring 2020

1 Introduction

Tomasetti et al. observe a correlation between the number of stem cell divisions (lscd) and lifetime cancer risk in their 2015 paper [1]. The authors claim that this 0.8 correlation (in a log-log scale) means that 65% of cancer incidence is related to the random mutations of cells (intrinsic risk) rather than environmental or hereditary factors (extrinsic risk). In this paper, we will look at the data and the results found by Tomasetti et. al. and see if we can reproduce these results through simulations.

2 Summary of Tomasetti et al. Results

Tomasetti et al. are motivated by the fact that cancers of some tissue types are much more common that of other tissue types. There are some extrinsic (environmental and hereditary) factors associated with some types of cancer, but these identified factors only account for a small percentage of cancer types and some tissue types with similar extrinsic risk factors still have significantly different cancer risks. Tomasetti et al. explore the fact that different tissue types have different number of lifetime stem cell divisions (lscd) and finds a correlation between lscd and lifetime risk of cancer.

Tomasetti et al. start by gathering a lot of data through the literature about the number

of stem cells, number of divisions, and lifetime risk. They then calculate the lscd as

$$lscd = \sum_{n=1}^{\log_2(C)} 2^n + c\dot{d} = c(2+d) - 2 \tag{1}$$

where c is the number of stem cells and d is the number of lifetime divisions. This formula is reduced by using a partial sum of a geometric series because s might not be a power of 2. THe authors found a 0.8 correlation between the log of the lscd and the log of the lifetime cancer risk, which they claim means that 65% of cancer incidence is related to the random mutations of cells (intrinsic risk) rather than environmental or hereditary factors (extrinsic risk).

The authors also looked at what they called the Extra Risk Score (ERS) in order to determine which tumor types had larger contributions to extrinsic factors. From the ERS, the cancer types were broken into two groups using unsupervised machine learning, and there were some observations that cancer types with more known extrinsic risk factors were in the label D-tumors (deterministic).

3 Model

To explored the results found by Tomasetti and Vogelstein, we used a modified version of the model presented by Bozic et al. [2]. Our process will be a discrete time branching process, where the death rate is given by $d_j = \frac{(1-s)^j}{2}$, the birth rate is $(1-d_j)(1-u)$, and the mutation rate is given by u. Thus, we will have two parameters to determine for this model, s, the selective advantage of each additional mutation, and u the mutation rate. We will start with c cells, which will be different for each type of tissue, and T, the time between each step, will be determined by the number of divisions in a lifetime for that particular tissue type. A difference between the model we will use and the one presented by Bozic et al. [2] is how the lines start. In Bozic et al., the model starts with one cell of mutation type 1. However, we will start in a different way. We will starting with c cells of type 0 (no mutations), and

there will be d time steps for the d lifetime divisions per stem cell. These cells of type zero will have $d_0 = 0.5$, so the birth and death rates are equal for these type 0 cells, and the cells of type > 0 will be the same as the Bozic et al. model.

Next, we need to determine what we will classify as cancer. First, we need to determine the number of mutations. Although the number of mutations is not consistent across all tissue types [3], we will start with a simplified model where the number of mutations for cancer is 3. We will also claim that a person has cancer if they have one surviving cell lineage of type 3. The extinction rate of a lineage was calculated in Bozic et al. [2] to be

$$q_j = 1 - 2js \tag{2}$$

and thus, the survival rate is $1 - q_j = 2js$.

The next challenge is to pick our mutation rate u. Tomasetti et al. cited that the mutation rate across all human tissues is relatively constant [1, 4]. We will use a mutation rate of 3.4×10^{-5}, the mutation rate used in Bozic et al. [2]. We will also pick a selective advantage parameter identical to this paper at $s = 0.004$.

3.1 Running our Model

We first ran our model for Pancreatic islet (For the number of stem cells (7.4×10^9) and number of lifetime divisions (80)). This cancer is the one with the smallest Extra Risk Score (ERS), which Tomasetti et al. claim will have the least effects from extrinsic factors and therefore should fix our model the best. We calculate a 0.001 lifetime cancer incidence, while the observed lifetime risk is 0.0002. This means that we are overestimating cancer risk by one order of magnitude, so our model isn't fitting perfectly, but not completely unreasonable for this type of tissue.

However, when we test our model on small intestine cancer (which has 2920 lifetime divisions), we get 100% of test runs result in cancer, but the observed lifetime risk from

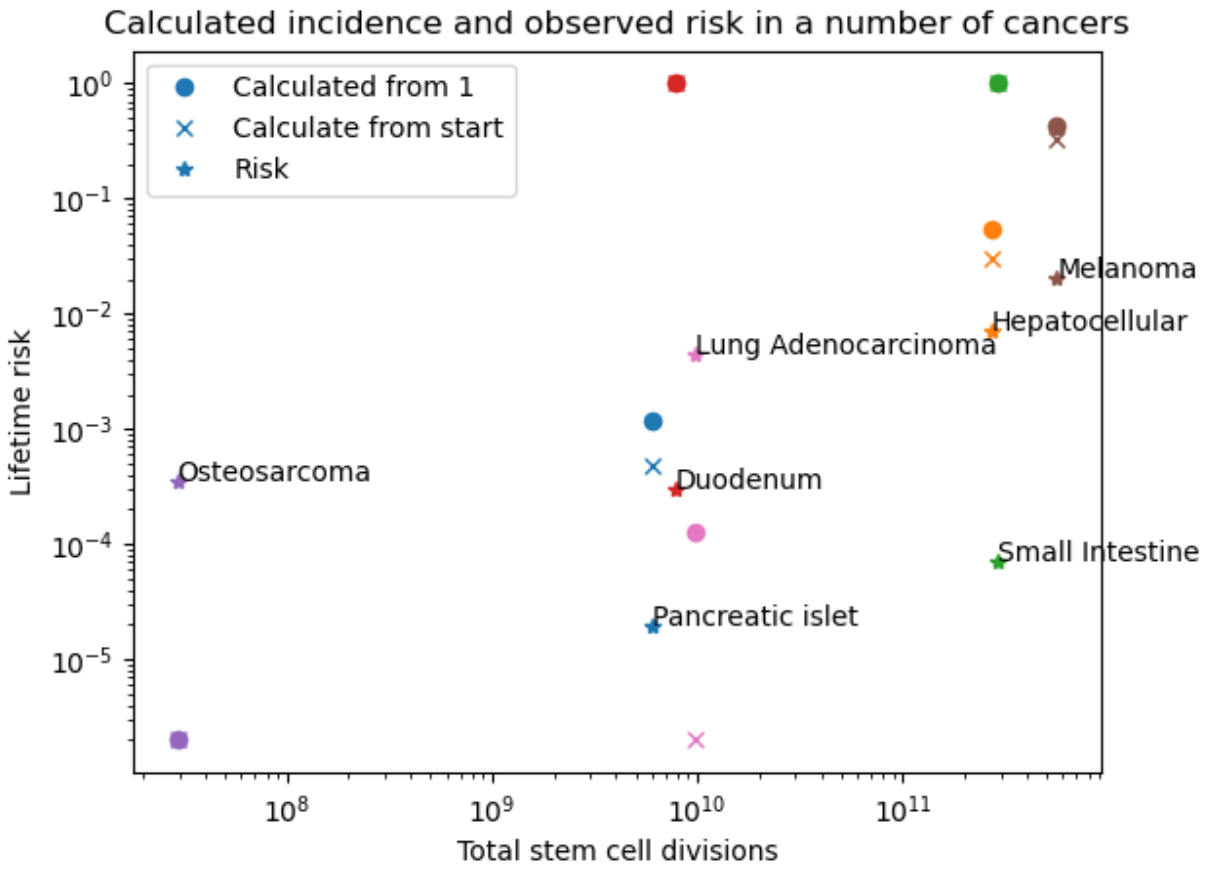


Figure 1: Simulation of results from Tomasetti et al. on an assortment of R-tumors. 'x's represent cancer risk from first form of the model starting at c cells. Dots represent second form of the model starting at 1 cell.

Tomasetti et al. is 0.0007. This is showing that this model does not fit this tissue type at all. Small intestine has a very similar ERS score to Pancreatic islet. Therefore, if the claims of Tomasetti et al. are true where the stochastic factors are the main factor in these cancers, then these two types of cancer should have similar results on our model.

Looking at the 'x' markers in Fig (1), you can see the simulated and observed lifetime risks of cancer for a variety of cancer types. You can see that for some cancer types, our model overestimates the risk, and for others, it underestimates the risk. There are also two types of cancers that have 100% incidence, and one type with 0% incidence. The types of cancers chosen were all classified as R-tumors by Tomasetti et al., which means that

the stochastic effects are most important and should ideally fit the model better than the D-tumors.

3.2 Modification to the model

Some cell types have a very few number of lifetime divisions per stem cell. Most extreme is Glioblastoma, which does not have any lifetime mutations per stem cell. Thus the cancer risk as calculated by our model will underestimate the cancer risk of these types of cancers with small numbers of divisions per cell because the number of divisions to get to c cells will be more significant. We will modify our model to start with 1 cell of type 0. We will have a birth rate of 100% for type 0 until we have c cells. Then we will proceed the same as before. This modification will allow for possible mutations to happen while the cells are growing to full capacity and make the calculated number of lifetime divisions equal to the number of divisions in our simulation. (Before we were underestimating the lscd.)

You can see the results of this modification in the dots of Figure (1). You can see that this modification raises the lifetime risk of Pancreatic, Lung, Hepatocellular, and slightly for Melanoma. This change does not make a significant change to the other types. The cancer types with the most significant change from this model are the ones with the lowest number of lifetime divisions per cell because these have the greatest effects of the growing cells.

4 Same lscd Varying Risk

From our model above, we were not able to reproduce the correlation found by Tomasetti et al. A flaw we can find in this correlation with our model is the fact that when we change the ratio of the number of stem cells to the number of cell divisions, the lscd will stay the same, yet the cancer incidence rate can drastically change. Figure (2) shows a test where the number of time steps was fixed, and the number of steps cells were varied to achieve a desired lscd. From this figure, you can observe that for a given lscd value, you can have 0%

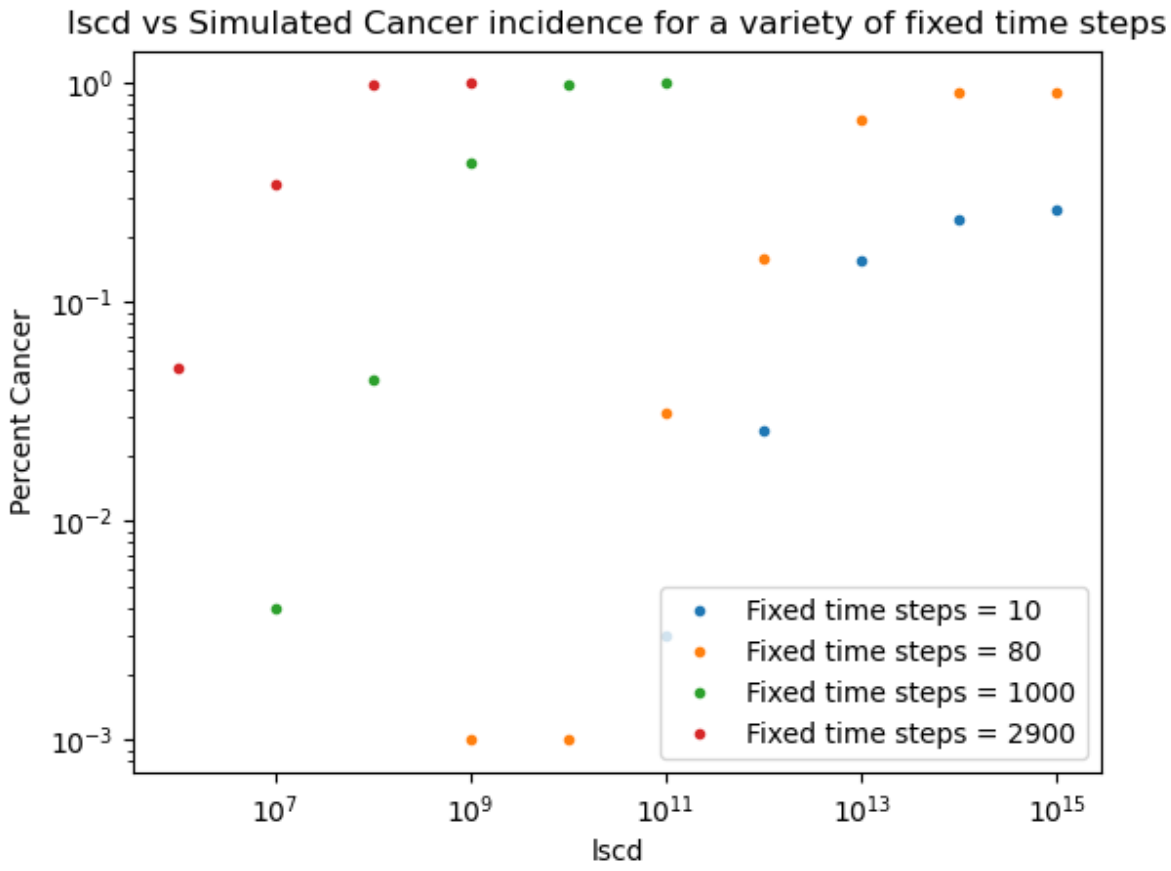
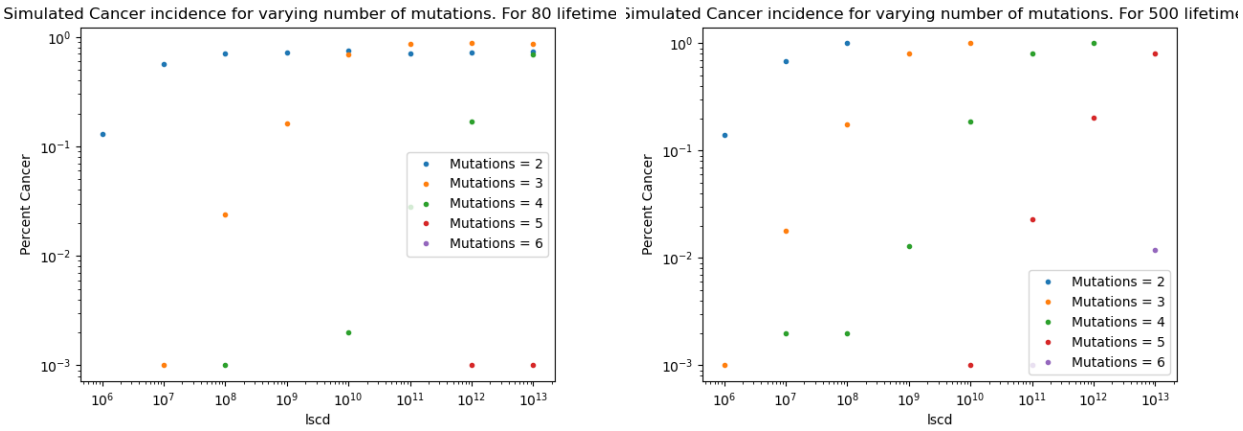


Figure 2: lifetime stem cell divisions vs lifetime cancer risk for varying ratios of divisions to stem cells. The number of time steps was held constant for each color and the number of stem cells was adjusted to achieve the desired lscd. Notice how the cancer risk ranges frmo 0-100% for a fixed lscd, based on the ratio of divisions to stem cells.

cancer or 100% cancer based on the ratio of time steps to stem cells. For example, when $lscd = 10^9$, if there are 10 time steps, there is a 0% chance of cancer. But if there are 2900 time steps, there is a 100% chance of cancer. For this simulation, we chose representative time step numbers, and lscd ranges to cover our complete range.

This observation that any cancer risk score can be achieved by any lscd by only changing the ratio of the number of time steps to number of stem cells shows a flaw in the correlation found by Tomasetti et al. It means that there is more going on than just the lscd, and that the systems have more factors contributing to cancer risk than solely the lscd score.



(a) (b)

Figure 3: Cancer risk vs lifetime stem cell divisions for varying mutations numbers. The number of time steps was kept constant at (a) 80 and (b) 500 and the number of stem cells was adjusted for the lscd. Notice how for a given lscd, the rate can vary between 0-100% based on the number of mutations required.

5 Number of Mutations

For their analysis, Tomasetti et al. assumed that all tissue types were the same, except the number of stem cells and lifetime divisions. However, one significant difference between tissue types is the number of mutations required before cancer. Anandakrishnan et al. [3] discuss the variety in the number of mutations for cancer. Most tissue types take between 3-5 mutation for cancer to form, but on the extreme ends, Glioblastoma only takes 2 mutations, while Liver hetatocellular carcinoma takes 8 mutations. This is a significant difference between tissue types that can play an significant role in cancer incidence rates.

In Fig (3), you can see the results of a simulation for varying number of mutations. Like the previous simulation, the number of time steps was kept constant (either at 80 or 500), and the number of stem cells were adjusted for the varying lscd. From this plot, you can see that based on the number of mutations, there could be between 0-100% cancer rate for a given lscd. This observation shows that the number of mutations is a critical component in the cancer incidence rate. Thus, the number of mutations for cancer cannot be a factor that is ignored nor assumed to be constant across tissue types.

Name	Risk	Stem Cell number	Lifetime divisions	Number of mutations	Calculated risk
Colorectal	0.048	$2 \cdot 10^8$	5840	3	1.0
Glioblastoma	0.00219	$1.35 \cdot 10^8$	0	2	0.0
Head & Neck	0.0138	$1.85 \cdot 10^7$	1720	4	0.0
Liver	0.0071	$3.01 \cdot 10^9$	88	8	0.0
Lung	0.0045	$1.22 \cdot 10^9$	5.6	3	0.0
Thyroid	0.01026	$6.5 \cdot 10^6$	7	5	0.0

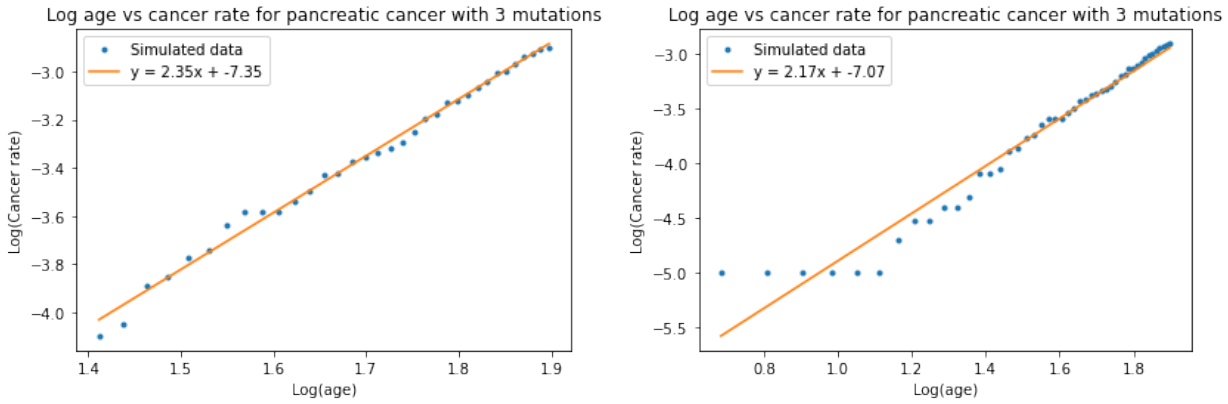
Table 1: Table displaying data used for simulation with varying number of mutations. This includes all tissue types used by both Tomasetti et al. [1] and Anandakrishnan et al. [3]. You can see that every tissue had either 100% or 0% cancer rate.

Next, we looked at the cancer types that were analysed by both Tomasetti et al. [1] and Anandakrishnan et al. [3]. There were 6 tissue types in common. You can few the data used in Table (1). However, when we ran our simulation accounting for these different number of mutation types, we got either all or none of the runs resulted in cancer, so we did not gain too insight from this simulation. A reason why this behavior was observed could be because the chosen tissue types either a had very small number of lifetime divisions or a large number of required mutations. Colorectal cancer on the other hand, had a very high number of lifetime divisions. These results help support the fact that there are many factor contributing to the lifetime cancer risk.

6 Comparison to Armitage and Doll

Tomasetti et al. also compared their results with results from Armitage and Doll. Tomasetti et al. say that their results agree with those of Armitage and Doll because they also found this exponential relationship. Armitage and Doll also concluded that this exponential relationship means suggests that stochastic factors are a significant factor in cancer incidence [5].

We decided to also compare our simulation results with those found by Armitage and Doll. To do this, we recorded the age at which cancer incidence occurred, which is slightly different than Armitage and Doll who looks at death age but the proportion should ideally



(a) (b)

Figure 4: Age vs cancer rate correlation with simulation. The orange line shows that best fit line on the loglog scale, and the blue dots show our simulated data. (a) starts at age 25 and this correlation holds. (b) starts at the first cancer incidence and the correlation does not hold for the entire range.

be the same. We used our model for pancreatic cancer the same as we set up in section 3 with 3 mutations. We plotted our results on a loglog scale and found a best fit line for our cancer incidence rate, shown in Figure (4).

Armitage and Doll found that there was a straight line correlation after the age of 25. We also found this to be true, there was a straight line correlation after a certain age, but not before. In Figure (4a), you can see the straight line correlation after the age of 25, and you can see that this does not hold before a certain age in Figure (4b). We also notice that the slope is 2.35, which is fairly close to our mutation number of 3. Armitage and Doll said that the slope of this line was the number of mutations it took to reach cancer, and we found through this simulation that seems fairly accurate for this example.

Armitage and Doll also discuss about how some types of cancers have this correlation between age and cancer rate is observed in some types of cancers but not all. They believe that this correlation occurs when stochastic factors are most important to a cancer and the correlation does not hold when extrinsic factors play a larger role. This simulation partial supports this because in this simulation, stochastic factors are the sole contributor, there are no extrinsic factors, and we are observing a correlation with a slope roughly equal to the

number of required mutations.

We also tried to expand the number of mutations to test this idea on a large area. However, this presented some challenges because more mutations require more time divisions until cancer. With only 80 time steps as in the above simulation, 4 mutations had $< 1/10000$ chance of cancer, and thus we need more time steps. However, if we increase the time steps such that the lifetime rate for some number of mutations is $\sim 100\%$, this correlation no longer holds, (we need small incidence numbers for this correlation). Thus, it becomes difficult to compare different mutation numbers hold all other factors equal.

7 Criticism from Literature

In the literature, there are many people criticizing and analyzing the methods and conclusions of Tomasetti et al. Many scientists criticize Tomasetti et al. for claiming that a majority of cancer is from "bad luck" and therefore most types of cancers are unavoidable . They discuss how labeling a majority of cancers as "bad luck" down plays the role of healthy lifestyles and measures taken to reduce cancer risk [6, 7, 8]. Also "bad luck" cancers do not describe how some cancers are more prevalent in some distinct populations and thus downplay the preventive measure that should be taken by certain populations at high risks of certain types of cancers [6, 8].

Some scientists also question how Tomasetti et al. data; such as, where they found some of their data and derived the lifetime cancer risk [9, 8]. The author's data set also only covers cancer types that make up 34% of cancer cases in the US [6], thus this correlation cannot be fully generalized to all types of cancer. Their data set is also skewed towards tumors types without known environmental factors, which can downplay the extrinsic risk factors of cancer, and also includes five types of Osteosarcoma, which is a significant percentage of this data set [10].

Similar to what we found in our analysis, some papers also talk about the fact that

Tomasetti et al. left out many important variables. For example the number of hits to reach cancer is very critical in the incidence rate [9]. Also there are variations in mutation rates of tissues which cannot just be ignored nor assumed to be equal [8].

Wu et al. [9] claim that this correlation cannot distinguish between intrinsic and extrinsic risk, meaning that this correlation does not imply the 65% contribution from intrinsic factors that Tomasetti et al. claim. Wu et al. propose that between 10-30% of cancer incidence difference can be contributed to intrinsic factors, while most cancer risk is attributed to extrinsic factors. Wu et al. back up their claims with both data driven and model driven methods. While Wu et al. also had criticism of their methods [11], (similar to criticism of Tomasetti et al.) it shows that cancer risk and what causes it are much more complicated than just a correlation between two numbers.

8 Conclusion

Overall, we found that Tomasetti and Vogelstein made a very bold claim in their 2015 paper. There is a correlation between the number of stem cell divisions and the risk of lifetime cancer, but there are also many more factors involved. There are more intrinsic factors that lead to cancer rates than just the number of stem cell divisions, such as number of mutations. There are also many extrinsic factors that affect cancer rates between populations.

Tomasetti et al. did make a good conclusion though: that many cancers are caused by "bad luck". They are correct in the fact that primary prevention measures such as vaccines and lifestyle can only limit cancer risk to a certain extent. This paper does highlight the need for secondary prevention, like early detection, and how many cancers are unpredictable in when and if they will occur.

References

[1] C. Tomasetti, B. Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science.* 2015 Jan 2; 347(6217): 78–81. doi: 10.1126/science.1260825

[2] Ivana Bozic, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W. Kinzler, Bert Vogelstein, Martin A. Nowak. Accumulation of driver and passenger mutations during tumor progression. *PNAS* October 26, 2010 107 (43) 18545-18550; https://doi.org/10.1073/pnas.1010978107

[3] Anandakrishnan R, Varghese RT, Kinney NA, Garner HR (2019) Estimating the number of genetic mutations (hits) required for carcinogenesis based on the distribution of somatic mutations. *PLOS Computational Biology* 15(3): e1006881. https://doi.org/10.1371/journal.pcbi.1006881

[4] Michael Lynch. Rate, molecular spectrum, and consequences of human mutation. PNAS January 19, 2010 107 (3) 961-968; https://doi.org/10.1073/pnas.0912629107

[5] Armitage, P., & Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. British journal of cancer, 8(1), 1–12. https://doi.org/10.1038/bjc.1954.1

[6] Christopher Wild, Paul Brennan, Martyn Plummer, Freddie Bray, Kurt Straif, Jiri Zavadil. Cancer risk: Role of chance overstated. Science 13 Feb 2015: Vol. 347, Issue 6223, pp. 728 DOI: 10.1126/science.aaa6799

[7] Jennifer Couzin-Frankel. The bad luck of cancer. Science 02 Jan 2015: Vol. 347, Issue 6217, pp. 12 DOI: 10.1126/science.347.6217.12 Article

[8] Michael O'Callaghan. Cancer risk: Accuracy of literature. Science 13 Feb 2015: Vol. 347, Issue 6223, pp. 729 DOI: 10.1126/science.aaa6212

[9] Wu, S., Powers, S., Zhu, W., & Hannun, Y. A. (2016). Substantial contribution of extrinsic risk factors to cancer development. Nature, 529(7584), 43–47. https://doi.org/10.1038/nature16166

[10] John D. Potter, Ross L. Prentice. Cancer risk: Tumors excluded. Science 13 Feb 2015: Vol. 347, Issue 6223, pp. 727 DOI: 10.1126/science.aaa6507

[11] Robert J Noble, Oliver Kaltz, Leonard Nunney, Michael E Hochberg. Overestimating the role of environment in cancers. OnlineFirst July 19, 2016 doi: 10.1158/1940-6207.CAPR-16-0126

A Python Codes

For all the codes used to run the simulations and make the displayed figures and more, please see the Github Repo:
`https://github.com/KatieJ16/AMATH536Project`.