Metodološki zvezki, Vol. 14, No. 2, 2017, 75–81

Bad Luck of Cancer – or Misinterpreted Statistics?

Janez Stare[1]

Abstract

A paper in Science (January 2015) claimed that the majority, 65% to be precise, of cancers is due to bad luck, so non-preventable. In this paper we show that the analyses, presented in the paper, give absolutely no grounds to make such a claim. Some of the arguments have in the meantime appeared elsewhere, but some have not. We also show that the authors' assumptions and their data can only support a claim of no more than 5% of cancers being random.

1 Introduction

In 2015 Tomasetti and Vogelstein (2015) published a paper in Science in which they analyzed association between the number of stem cell divisions of given tissues in a lifetime and probability of cancers of those tissues. They found a strong correlation of more than 0.8 (or $R^2 = 0.65$) between the logarithms of these variables and based on this concluded that a great majority of cancers, approximately two thirds of them, occurs randomly due to stem cell divisions, and that the other factors contribute only to the residual third of all cancers. Their work immediately met with some negative reactions, published mainly as letters in Science, but their results were essentially not disputed. Only a year later, Wu, Powers, Zhu, and Hannun (2016), in a paper published in Nature, showed that correlation cannot say anything about the proportion of random cancers.

In this paper we give a more detailed and more versatile criticism of the Tomasetti and Vogelstein analysis, and also show that using their assumptions one cannot claim more than 5% of all cancers, used in their analysis, being random.

There are different ways to show that correlation, or R^2, cannot come close to estimating the proportion of random cancers. In the first section we show in a direct way that such an estimation is impossible, and give an illustration which completely mimics the analysis by Tomasetti and Vogelstein, but on a different data set which makes the mistake more obvious. In the second section we show how understanding the properties of R^2 makes it obvious that something is wrong with the conclusion about the proportion of random cancers, and, finally, we show that Tomasetti and Vogelstein's assumptions and their data can only support a claim of no more than 5% of cancers being random.

[1] Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia; janez.stare@mf.uni-lj.si

2 A Direct Way of Showing that R^2 cannot Measure Randomness of Cancer

The model that Tomasetti and Vogelstein assume is

$$p_i = 10^a d_i^b$$

where p_i is the probability of cancer for tissue i, d_i is the number of stem cell divisions for that tissue, and a and b are coefficients to be determined from the data. Taking logarithms we get

$$\log p_i = a + b \log d_i$$

and by fitting a regression line to the points ($\log p_i$, $\log d_i$) we get the estimates of the coefficients and an R^2 equal to 0.65. If we say that cancers which occur regardless of people's life style, environment or similar, are random, and other cancers non random, then we can write every probability p_i as a sum of two probabilities, the probability of a cancer being random (or stochastic, denoted by s_i) and a probability of a cancer being non-random (or non-stochastic, denoted by ns_i). So let's write $p_i = s_i + ns_i$ and

$$\log(s_i + ns_i) = a + b \log d_i$$

If we knew all s_i then proportion of random cancers (PRC) among all cancers would be simply

$$PRC = \frac{\sum s_i}{\sum p_i}$$

The paper by Tomasetti and Vogelstein hints that this proportion is estimated by R^2 that they obtained, so

$$R^2 = \left(\sum s_i\right) / \left(\sum p_i\right) = 0.65.$$

It should be obvious that regression analysis will give the same results, and so the same R^2, regardless of how each p_i is decomposed into s_i and ns_i. This means that R^2 has *absolutely nothing* to do with PRC. This simple fact is illustrated in Figure 1 for subset of values in Tomasetti and Vogelstein paper. Proportion of random cancers is varied, but the total probabilities remain the same. The actual situation should, under the Tomasetti and Vogelstein assumption, look something like subfigure (c), but the values for randomness (dark areas) can be almost anything.

We illustrate this simple fact using another example, completely analogous to the above, but much more obvious.

Let us look at the European countries and record their areas and population sizes. Data can be, for example, found here http://bit.ly/1K2oosV. Of course, any set of countries will do. So d_i now represent areas, and p_i are proportions of each country's population in the total population of Europe. We use the same model as Tomasetti and Vogelstein (so logs of proportions and areas) and fit a regression line (Figure 2). We get an R^2 of 0.86. If, as an example, s_i and ns_i represent proportions of smokers and non-smokers of country i in the total of European population, can we say that there are 86% of smokers in Europe? Obviously not. But this is exactly the argument that Tomasetti and Vogelstein make. The point is, again, that s_i and ns_i can be any two numbers adding up to p_i, and we will always have the same R^2.

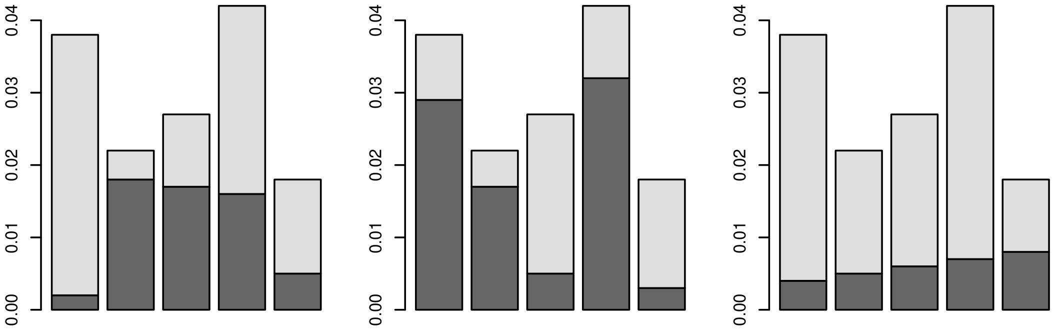

Figure 1: Different proportions of random cancers do not change the overall proportions

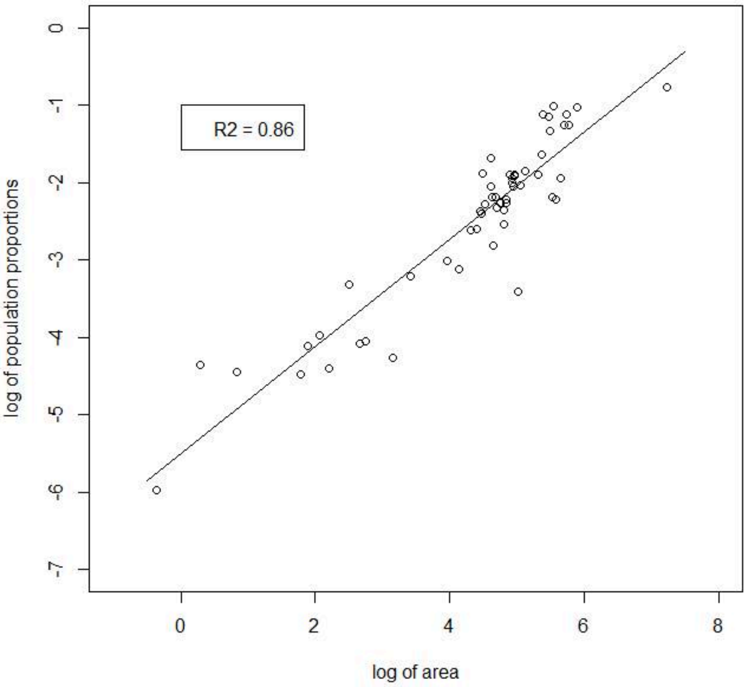

Figure 2: Regression of proportions of countries' populations in the total European population on countries' areas (log scales)

3 Indirect Ways of Showing that R^2 cannot Measure Randomness of Cancer

Tomasetti and Vogelstein calculate the correlation for a certain range of the number of stem cell divisions. If R^2 did indeed estimate a proportion of random cancers, for cancers included in their analysis, then we should get a similar result if we calculate this proportion for a subrange. For example, if we did this separately for the tissues with stem cell divisions below the median, and above the median, then we should be able to combine these results into the overall number. For example, if R^2 below the median was a, and above the median b (of course, any subdivision would do), then a weighted average of these two numbers should give us 0.65. In fact, the numbers that we get for R^2 are 0.43 and 0.21 (Figure 3). So, parts have less random cancers than the whole? On the other hand, if, in future, say, other cancers are included which have numbers of stem cell divisions lower than the present minimum, or higher than the present maximum, R^2 will increase.

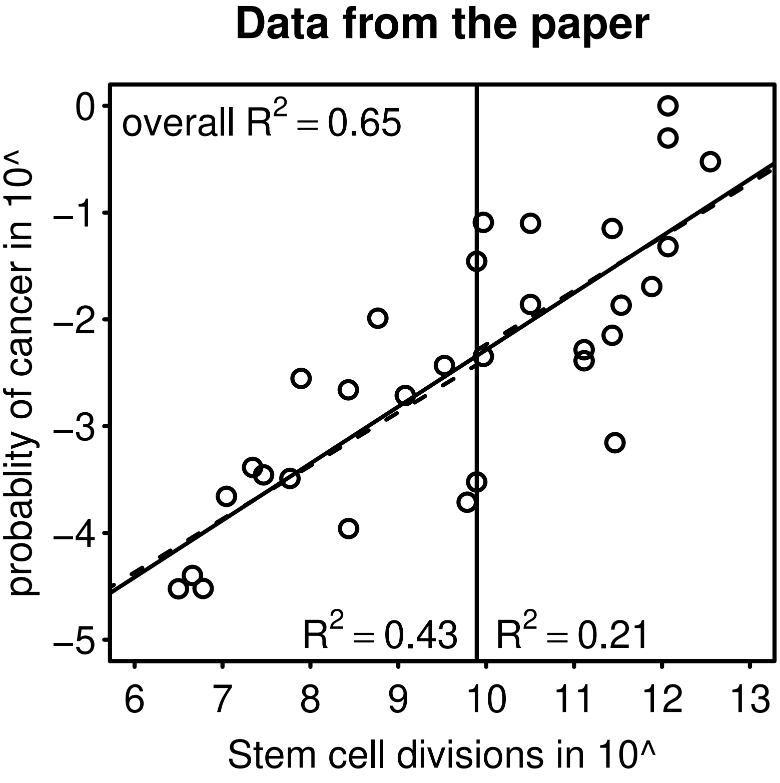

Figure 3: Proportion of explained variation depends on the chosen interval of the independent variable

Another way of showing that R^2 cannot estimate PRC is to assume that all probabilities in their data were multiplied by a certain factor. One can imagine a country which has more (less) risk due to some extra factor (or lack of some factor). Since the probabilities of random cancers cannot change, their proportion in the overall number of cancers will now be different, smaller or larger, depending on the factor. But R^2 will not change! This is illustrated in Figure 4.

4 A Note on Data

Data which Tomasetti and Vogelstein paper analyze contain some points that should not be there. For example, probabilities for lung cancer are given separately for smokers and nonsmokers. These are conditional probabilities, given values of some extra variable, and are certainly not the probabilities which one would predict based on the number of the

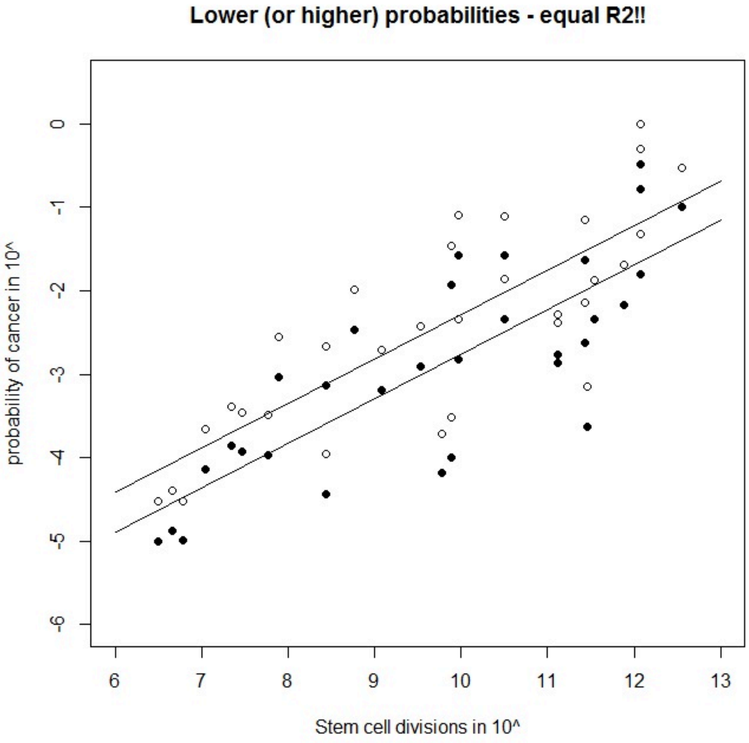

Figure 4: Increase or decrease of cancer probabilities by a give factor does not change the R^2

stem cell divisions. They argue that leaving them like this does not change the results of the analysis, but this is not a valid argument. We are interested in probabilities of cancers, given the number of stem cell divisions, nothing more. For our calculations in the next section we corrected these data, so that every number of stem cell divisions has just one corresponding probability of a cancer of that tissue. We used data from Tomasetti and Vogelstein supplementary material to do this. For example: lifetime risk of lung cancer for nonsmokers is 0.0045, and for smokers 0.081. Assuming (as Tomasetti and Vogelstein report in their supplementary material) that the proportion of smokers is 0.3, then the overall probability of lung cancer is $0.0045 \times 0.7 + 0.081 \times 0.3 = 0.02745$.

5 A Different Way of Estimating the Probability of Random Cancers

If we assume, as Tomasetti and Vogelstein do in their model, that probabilities of cancers only depend on the number of divisions, then the candidates for random cancers are those lying low on the graph. In Figure 5 we connected two low points on the graph and values on this line could be seen as (log of) probabilities for random cancers. Anything above them must be non random. Of course, assuming that those two points themselves represent probabilities of random cancer for those two tissues is probably overestimating the true probabilities. And still, the total probability of random cancer, calculated in this way, is very low.

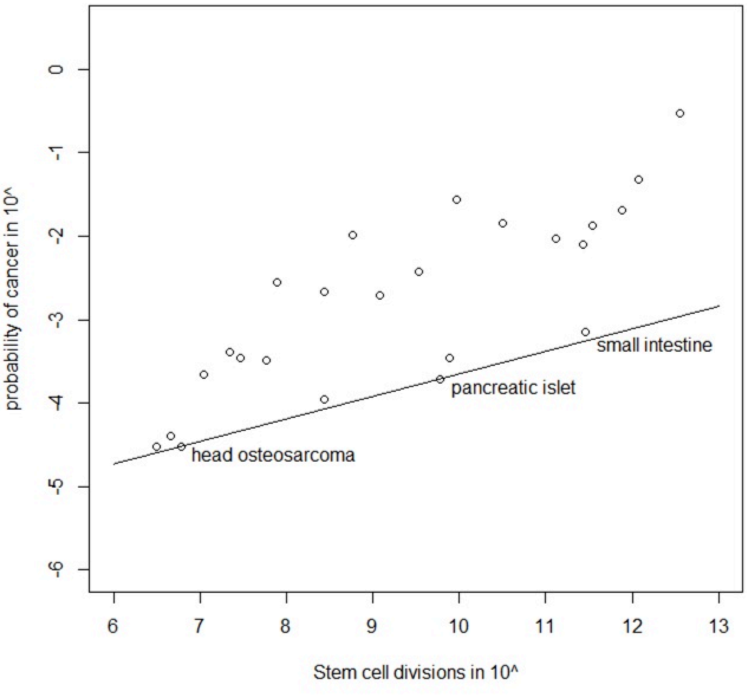

Figure 5: Regression line through two points representing possible random cancers

Another, end even better, way of calculating the overall probability of random cancers, is to do the following

1. Find cancer with the lowest probability per division.
2. Take this (or part of this) probability to be the probability of random cancer per division.
3. Multiply this probability by the number of divisions for each tissue.
4. This gives us probabilities of randomness per each tissue.

When we apply above to the Tomasetti and Vogelstein data, it turns out that the lowest probability per division is for small intestine cancer. Assuming (probably unreasonably) that all small intestine cancers are random, and continuing with points 2. to 4. above, and then summing up all the obtained probabilities, we get that the overall probability of random cancers is 1.6%! This translates into no more than 5% of all cancers in their analysis being random, depending on how much independence we want to assume among cancers. Of course we do not believe these numbers, we simply show that the claim of most cancers being random rests on a vary shaky ground.

6 Conclusion

There is no way one can claim any proportion of cancers being random, based on the analysis of Tomasetti and Vogelstein. In fact, their inherent assumption of all divisions being equally likely to produce cancer, yields a very low estimate of the proportion of random cancers.

References

[1] Tomasetti, C. and Vogelstein, B. (2015): Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, **347**, 78–81.

[2] Wu, S., Powers, S., Zhu, W., and Hannun, Y. A. (2016): Substantial contribution of extrinsic risk factors to cancer development. *Nature*, **529**, 43–47.