CANCER

Genes, environment, and "bad luck"

Explaining cancer risk in a statistical sense

By **Martin A. Nowak**[1] *and* **Bartlomiej Waclaw**[2]

It is a human trait to search for explanations for catastrophic events and rule out mere "chance" or "bad luck." When it comes to human cancer, the issue of natural causes versus bad luck was raised by Tomasetti and Vogelstein about 2 years ago (*1*). Their study, which was widely misinterpreted as saying that most cancers are due neither to genetic inheritance nor environmental factors but simply bad luck, sparked controversy. To date, a few hundred papers have been written in response, including (*2*–*6*), with some [e.g., (*2*)] coming to opposite conclusions. What is this controversy about? Tomasetti and Vogelstein concluded that 65% of the differences in the risk of certain cancers is linked to stem cell divisions in the various cancerous tissues examined (*1*). On page 1330 of this issue, Tomasetti *et al.* (*7*) provide further evidence that this is not specific to the United States.

In their earlier study, Tomasetti and Vogelstein estimated, for 31 cancers occurring in the United States, the number of stem cells N present in the tissue where that cancer originates and the rate b at which those cells divide. Then, for each cancer, they plotted the lifetime risk R versus the total number of stem cell divisions $D = NbT$, where T is the estimated human life span. The authors found a strong correlation ($r = 0.81$) between log R and log D and concluded that "65% of the differences in cancer risk among different tissues can be explained by the total number of stem cell divisions in those tissues" (*1*).

In the current study, Tomasetti *et al.* looked beyond the United States. They analyzed cancer incidence rates in 69 countries, representing two-thirds of the world's population, and show that the median correlation coefficient is very close ($r = 0.80$) to what they found for the United States.

> ***"The findings point to a clear need for a precise mathematical understanding of cancer."***

The discovery of such a strong correlation between cancer incidence and stem cell divisions raises a variety of scientific and epidemiologic questions. The strength of the linear correlation indicates how well the value of log R can be predicted from log D alone. If the correlation was perfect ($r = 1$), the average risk could be predicted with certainty from the number of divisions as the only variable, yet one would not necessarily know anything about the link between the observed correlation and the biological mechanism. Thus, the correlation "explains" the data in the statistical but not in a biological sense.

In particular, one cannot use this relationship to differentiate between contributions from replication, hereditary, and environmental factors to cancer risk. Importantly, Tomasetti and Vogelstein never claimed such a possibility in their earlier or the current studies. Wu *et al.* (*2*) provided a simple argument [also discussed in (*7*)] why this is not possible. A hypothetical environmental substance that increases the risk of all cancers in the same proportion would not affect the correlation, even though this substance would increase the fraction of environment-induced cancers.

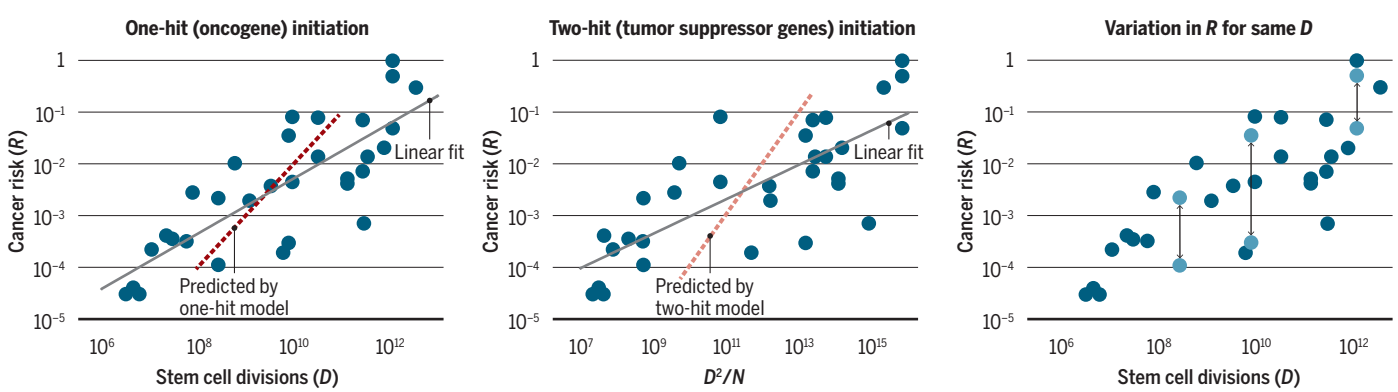
However, Wu *et al.* (*2*) went further: They reanalyzed the data of Tomasetti and Vogelstein and concluded that the vast majority of cancers are caused by extrinsic risk factors. Although their use of mathematical models is welcome, there are two problems with their analysis. One is their assumption that if two cancers have the same D value, then the variation in R is only caused by extrinsic risk factors, but this assumption is unwarranted because, as we explain below, R is affected by many other factors, which may vary in different cancers. Another problem with their analysis is the formulation of a mathematical model of cancer that ignores clonal expansion. Based on this model, Wu *et al.* conclude that the normal mutation rate of somatic cell division is too low to explain cancer incidence, but this conclusion rests on the unrealistic assumption that there is no clonal expansion.

Mathematical modeling can help to better explain factors that can affect the correlation between R and D (*8*–*12*). Consider a simple model (see the box) in which the lifetime

[1]*Program for Evolutionary Dynamics, Harvard University, Cambridge, MA 02138, USA.* [2]*School of Physics and Astronomy, University of Edinburgh, Edinburgh, UK. Email: martin_nowak@harvard.edu*

Cancer risk

Lifetime cancer risk R [data from (*1*)] compared to model predictions are shown (left and middle panels). Variation in R for the same D can be explained by a number of different factors, but only some of them are environmental (right panel).



GRAPHIC: J. YOU/SCIENCE

cancer risk R factorizes into the product of the probability of cancer initiation P and progression Q. Let N and b denote the number of stem cells and their rate of division. First, consider one-hit initiation via oncogenes (that is, a single mutation causes cancer initiation). Let μ be the probability for activating an oncogene multiplied by the number of possible oncogenes in this tissue and ρ be the average fixation probability of the oncogenic mutation, which depends on the selective advantage of the mutant and the local geometry of the tissue (*13*). The probability that cancer has been initiated in a tissue by time t is $P \cong Nb\mu\rho t$, which therefore increases linearly in time. Because $D = Nbt$, the lifetime risk of cancer $R = PQ = \rho\mu DQ$ is also linear in the total number of stem cell divisions D. Plotting log R versus log D yields a straight line with a slope of 1 (see the figure). Consider, instead, a two-hit model for cancer initiation (thus, it requires two mutations to initiate cancer) (*14*). Here, for example, if two alleles of a tumor suppressor gene are inactivated, the probability of cancer initiation increases as the square of time, and linearly as a function of the variable D^2/N. This again gives a straight line in the log-log scale, with a slope of 1 (see the figure). Yet the data of Tomasetti and Vogelstein (*1*) have much lower slopes in both cases: 0.53 and 0.33, respectively. This has received some attention so far (*4, 5, 15*), but not as much as it deserves, because it prompts the question, which biological mechanism could generate a slope of less than 1?

The model suggests two possible ways. One is that progression Q can be less likely in tissues that have large D. In this case, evolution could have generated additional checkpoints—for example, requiring a larger number of subsequent driver mutations (those that spur cancer progression), or better immune surveillance, in tissues that have more stem cell divisions. Here, Q would have to decrease with D. The other possibility is that, among the total number of stem cell divisions D, there may be, for reasons of tissue geometry (*8, 13*), a smaller number of effective stem cell divisions, and only such cell divisions contribute to the risk for cancer initiation. Both explanations are intriguing, and a combination of them could apply. A recent study argues that calculating the correlation coefficient for all cancers is problematic, and only cancers with similar features (tissue type) should be compared (*15*).

The mathematical model also reveals factors that lead to variation in R given the same D (see the figure). These factors include (i) the number of target genes that lead to cancer initiation (and whether they are oncogenes or tumor suppressors); (ii) the number of additional hits that are required for progression; (iii) different rates of cell division, levels of apoptosis, or immune surveillance during cancer progression; and (iv) exposure to environmental agents that increase the mutation rate or rate of cell division during cancer initiation or progression. Thus, very different factors lead to variation in R given the same D, and only some of them are "extrinsic"—that is, environmental or hereditary.

Can we say anything about contributions of various factors (replication, environmental, and hereditary) to cancer risk? Tomasetti *et al.* propose a method based on comparing cancer sequencing and epidemiological data to estimate the fraction of mutations that come from replication, environmental, and hereditary factors. They interpret their statistical analysis as showing that as much as 66% of driver mutations are due to replication. This does not stand in contradiction with many cancers being preventable. Tomasetti *et al.* give a simple reason for this apparent contradiction: If more than a single mutation causes cancer, all but one can be due to replication, and yet cancer may be entirely preventable if the last mutation is due to environmental factors.

Tomasetti *et al.* show that a large portion of the variation in cancer risk among tissues can be explained (in the statistical sense) by the number of stem cell divisions. An understanding of cancer risk that did not take bad luck into account would be as inappropriate as one that did not take environmental or hereditary factors into account. The earlier analysis by Tomasetti and Vogelstein has already stimulated much discussion, and the findings reported now by Tomasetti *et al.* will continue to do so. The findings point to a clear need for a precise mathematical understanding of cancer. It will take many years to answer in detail the interesting and exciting questions that have been raised.

Cancer is a by-product of the fact that we are made of cells that are individual replicators. Mutations destroy their cooperative program and elicit unwanted replication (defection). Mutants arise whenever cells divide. These normal mutations are due to "bad luck." But because humans don't like to leave it there, we explain the origin of such mutations in more scientific terms such as "thermal fluctuations" or "quantum jumps." Indeed, even Albert Einstein famously said, "God does not play dice with the universe." ■

Mathematical modeling

Mathematical modeling helps to explain factors that affect the correlation between lifetime cancer risk and cell division number.

Let $f(t)$ and $g(\tau)$ denote, respectively, the probability density functions that the first cancer cell arises at age t, and that cancer progresses to full disease after some further time τ. The lifetime risk of cancer can be expressed as

$$R \approx \int_0^T dt\, f(t) \int_0^{T-t} g(\tau) d\tau$$

Here, T is the human life span. We have assumed $R \ll 1$. We expect the initiation rate $f(t)$ to increase with t, and $g(\tau)$ to have a maximum somewhere in the range $\tau = 5 \ldots 20$ years and to fall off for larger τ. Hence, if $T = 80$ years (average life span) and $g(\tau)$ is concentrated in the range $0 < \tau < \tau_{max} \ll T$, then the risk can be approximately written as

$$R \approx \int_0^T f(t)\, dt \quad \times \quad \int_0^{\tau_{max}} g(\tau) d\tau = PQ$$

Here, P and Q are the lifetime probabilities of cancer initiation and progression.

We consider a model in which N stem cells are divided into m "compartments" of size n cells ($n = 1 \ldots 100$). Let ρ be the probability that the progeny of a cancer cell replaces all normal cells in that compartment. If k "hits" are necessary to initiate cancer and the probability of each hit is μ, then the initiation probability P is

$$k = 1\text{: } P \approx \rho\mu D \qquad k = 2\text{: } P \approx \frac{\rho\mu^2 D^2}{2N}$$

where we have assumed $P \ll 1$. Identical results are obtained for the model with just one compartment and large N. The lifetime cancer risk is

$$k = 1\text{: } R \approx \rho\mu DQ \qquad k = 2\text{: } R \approx \frac{\rho\mu^2 Q}{2} \frac{D^2}{N}$$

Risk increases linearly with D for the one-hit model, or linearly with D^2/N for the two-hit model, if all other parameters are constant.

REFERENCES

1. C. Tomasetti, B. Vogelstein, *Science* **347**, 78 (2015).
2. S. Wu, S. Powers, W. Zhu, Y. Hannun, *Nature* **529**, 43 (2016).
3. D. Sornette, M. Favre, *EPJ Nonlin. Biomed. Phys.* **3**, 10 (2015).
4. L. Nunney, B. Muir, *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20150161 (2015).
5. R. Noble, O. Kaltz, M. E. Hochberg, *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20150104 (2015).
6. C. R. Weinberg, D. Zaykin, *J. Natl. Cancer Inst.* **107**, djv125 (2015).
7. C. Tomasetti, L. Li, B. Vogelstein, *Science* **355**, 1330 (2017).
8. M. A. Nowak, F. Michor, Y. Iwasa, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14966 (2003).
9. E. G. Luebeck, S. H. Moolgavkar, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 15095 (2002).
10. D. Wodarz, N. L. Komarova, *Dynamics of Cancer: Mathematical Foundations of Oncology* (World Scientific, 2014).
11. P. M. Altrock, L. L. Liu, F. Michor, *Nat. Rev. Cancer* **15**, 730 (2015).
12. S. A. Frank, Y. Iwasa, M. A. Nowak, *Genetics* **163**, 1527 (2003).
13. E. Lieberman, C. Hauert, M. A. Nowak, *Nature* **433**, 312 (2005).
14. M. A. Nowak, F. Michor, N. L. Komarova, Y. Iwasa, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 10635 (2004).
15. R. J. Noble, O. Kaltz, L. Nunney, M. E. Hochberg, *Cancer Prev. Res.* **9**, 773 (2016).

10.1126/science.aam9746