OXFORD

JNCI J Natl Cancer Inst (2015) 107(7): djv125

doi:10.1093/jnci/djv125
First published online May 8, 2015
Commentary

COMMENTARY

Is Bad Luck the Main Cause of Cancer?

C. R. Weinberg, D. Zaykin

Affiliations of authors: Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC (CRW, DZ).

Correspondence to: C. R. Weinberg, PhD, Biostatistics and Computational Biology Branch, MD A3-03, National Institute of Environmental Health Sciences, P. O. Box 12233, 111 T.W. Alexander Drive, Research Triangle Park, NC 27709 (e-mail: weinber2@niehs.nih.gov).

Abstract

A recent study reports that the log lifetime incidence rate across a selection of 31 cancer types is highly correlated with the log of the estimated tissue-specific lifetime number of stem cell divisions. This observation, which underscores the importance of errors in DNA replication, has been viewed as implying that most cancers arise through unavoidable bad luck, leading to the suggestion that research efforts should focus on early detection, rather than etiology or prevention. We argue that three statistical issues can, if ignored, lead analysts to incorrect conclusions. Statistics for traffic fatalities across the United States provide an example to demonstrate those inferential pitfalls. While the contribution of random cellular events to disease is often underappreciated, the role of chance is necessarily difficult to quantify. The conclusion that most cases of cancer are fundamentally unpreventable because they are the result of chance is unwarranted.

A recent report (1) by Tomasetti and Vogelstein investigated the relationship between the estimated lifetime number of stem cell divisions in selected tissue types and the lifetime incidence rates for 31 corresponding site-specific types of cancer. The logarithm of incidence was strongly related to the logarithm of the estimated number of stem cell divisions (their Figure 1), with an R^2 of 0.65. The idea is that each stem cell division represents another ticket in the cancer lottery: Accordingly, although the number of required mutations may differ across tissues, tissues with more stem cell divisions tend to experience correspondingly higher rates of cancer (2).

The authors conclude, "These results suggest that only a third of the variation in cancer risk among tissues is attributable to environmental factors or inherited predispositions. The majority is due to "bad luck"." An accompanying commentary (2) interprets these findings to suggest that two-thirds of cancers arise stochastically through accrued somatic mutations. Broader implications are proposed, including "that cancer often cannot be prevented and more resources should be funneled into catching it in its infancy." Should we conclude that our cancer dollars would be better spent on screening rather than on etiology and prevention? One scientist quoted in the commentary offered that, "the average cancer patient . . . is just unlucky." This report captured the imagination of the media (http://www.forbes.com/sites/geoffreykabat/2015/01/04/most-cancers-may-simply-be-due-to-bad-luck/), but evoked consternation in the cancer research community. In a subsequent press release (http://www.hopkinsmedicine.org/news/media/releases/bad_luck_of_random_mutations_plays_predominant_role_in_cancer_study_shows), the authors addressed some of the issues raised by online comments. While noting that, "Some have misunderstood our research to say that two-thirds of cancer cases are due to bad luck," they also stated, "we calculate that two thirds of the variation (in cancer) is attributable to the random mutations that occur in stem cell divisions throughout a person's lifetime, while the remaining risk (presumably 1/3) is associated with environmental factors and inherited gene mutations."

In our view, the findings of Tomasetti and Vogelstein (1) can tell us little about the relative importance of luck vs inherited genetic variants or environmental factors. We make three points. First, we argue that a high correlation with lifetime stem cell divisions (even if the R^2 were 1.0) has little bearing on how much cancer could be due to inherited genetic variants or preventable environmental factors. Second, by using tissue-specific characteristics to predict cancer type–specific rates, Tomasetti and Vogelstein (1) present an analysis that addresses variation across tissue types in risk that has been aggregated among individuals, but neglects sources of variation in risk across individuals within particular types of cancer. Finally, the very notion that we should be able to partition causes into bad luck vs other factors, whose contributing fractions sum to 1.0, is false.

Received: February 19, 2015; **Revised:** March 17, 2015; **Accepted:** April 8, 2015

Published by Oxford University Press 2015. This work is written by (a) US Government employee(s) and is in the public domain in the US.

A later press release provided by Tomasetti and Vogelstein (4) used an apt analogy: The lifetime number of stem cell divisions in a particular tissue type is analogous to the number of miles driven in a car trip—the longer the trip, the greater the likelihood of a fatal crash. We will further explore that instructive analogy. Figure 1 shows the log of the rate of death due to automobile accidents for US states and Washington DC in 2012, plotted against the log of the mean number of miles traveled per capita. The R^2 for this regression is 0.7, suggesting that over two-thirds of the variation in log fatality rates across states can be statistically explained by the log of the number of miles traveled per capita. Does this imply that only one-third of traffic deaths could be due to modifiable factors such as inadequate emergency response systems, failure to wear seat belts, automobile design flaws, texting, and drunk driving? Perhaps then, following the reasoning applied earlier, improved safety measures could at best prevent only about one-third of traffic fatalities. We return to this example later. Three issues need to be considered before drawing inferences from this kind of data.

1. R^2 does not explain much

A conceptual problem arises through a misunderstanding of the word "explain" as used in statistical regression analyses. A large R^2 (for example the 0.65 that was calculated by Tomasetti and Vogelstein) simply reflects that the variation in the outcome, Y, is much larger than the variation in the residuals (Y minus the modeled value of Y) around the regression model predictions. By definition, 1 - R^2 is the ratio of the variation among the residuals to the variation in Y. We say that the predictor has "explained" a proportion R^2 of the variation in Y. This jargon is unfortunate, because it does not mean that this proportion of Y has been explained in a causal sense. The finding that the log of the lifetime number of stem cell divisions statistically "explains" two-thirds of the variation among the 31 selected site-specific log cancer incidences does not mean that random errors associated with stem cell divisions explain two-thirds of cancer. Similarly, even though distinct tissues always have the same inherited genomes and consequently the R^2 would be 0 for the role of inherited genetic variants, that does not imply that inherited genetic variants play no role in cancer.

Consider the following hypothetical thought experiment. Suppose an evil agent exposes the entire US population to a powerful new carcinogen that doubles the incidence of all 31 cancers. One might conclude that the fraction explained by this exposure must be one-half, because half the cases would not have occurred had they not been exposed; one might then reason that the fraction explained by stochastic errors in stem cell division would now be correspondingly smaller. But even with the new two-fold higher incidence numbers, the correlation would not change at all, because the points (all being on a log scale) would rigidly shift upward, each by log (2). The fraction of the variability in log incidence that is "explained" (in the statistical sense) by the number of stem cell divisions would remain at two-thirds. Similarly, if the population were administered an anticancer vaccine that could prevent the occurrence of half of cancers, regardless of type, the correlation would still be two-thirds. Clearly one cannot infer from the Tomasetti and Vogelstein data that two-thirds of cancer is unpreventably due to bad luck.

Another way to see this point is to note that the data in their Figure 1 (1) are consistent with a hypothetical (though unlikely) scenario in which cancer is entirely due to environmental mutagens that cause a mutation with a certain probability at each cell division. This far-fetched scenario provides an example to demonstrate that a high correlation with the number of stem cell divisions does not in itself imply an important role for bad luck. The data remain consistent with the possibility that the relationship observed is largely secondary to preventable factors that interfere with faithful replication of DNA, with DNA repair, or that cause dysfunction of protective mechanisms that clear abnormal cells. While bad luck may play an important role in carcinogenesis, the data do not compel that conclusion.

2. Variation in aggregated risk is not the same as variation in risk among individuals

Though important for hypothesis generation, analyses based on aggregates are often misleading. For example, the observation that rates of lung cancer in US counties are negatively

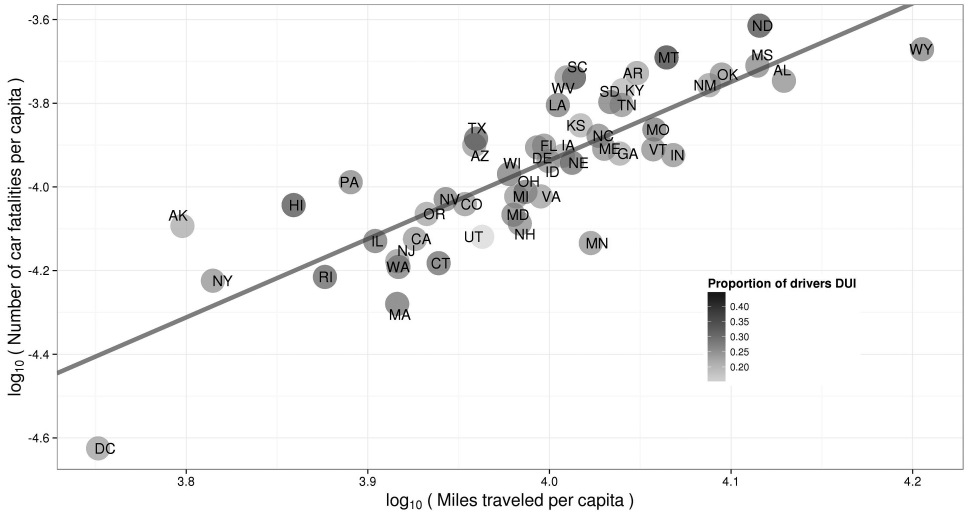

Figure 1. Automobile fatalities, plotted against miles traveled. A high proportion (70%) of the variation in log risk (per year) of a fatal crash can be explained by the log of the average number of miles driven per year per person. (Data from National Highway Traffic Safety Administration, http://www-fars.nhtsa.dot.gov.) The shading indicates the fraction of fatalities where one or both drivers were found to be under the influence of alcohol.

COMMENTARY

correlated with levels of radon (3) was used to promote the questionable notion that background radiation is good for us. Likewise, an observation that, worldwide, countries with higher average body mass index have longer life expectancy might tempt the unwary to infer that obesity is beneficial, forgetting that the data include countries with endemic malnutrition and infectious disease. The analysis provided by Tomasetti and Vogelstein (1) is similarly "ecologic," because it concerns variation in average risk across cancer types but does not address variability in risk across individuals in relation to any single cancer type, a point made in a subsequent commentary (4).

Characterization of cancer sites by their lifetime risk obscures the roles of external factors and inherited genetic variants in risks experienced by individuals. For example, Tomasetti and Vogelstein (1) classify thyroid medullary carcinoma as one of the "replicative" (bad luck) tumors for which "the contribution of external environment and heredity is minimal." While this cancer has a very low lifetime risk, up to 30% of cases are due to completely penetrant hereditary mutations in the RET proto-oncogene (5). Distinguishing the subcategory of thyroid medullary carcinoma arising from that cause would generate a dot with a lifetime risk of 1.0, producing an outlier that would markedly reduce the R^2. In short, the R^2 from an analysis of cancer types based on aggregated risk for each type obscures the contributions of individual risk factors to each cancer type.

3. One cannot partition causes into fractions that add up to 1

A related conceptual issue is even more fundamental: Expanding on a point made recently in an online letter (http://www.sciencemag.org/content/early/2015/02/04/science.aaa6094), environmental exposures, germ-line genetic variants and random events like replicative errors typically act in concert; the effects cannot be treated as separable. It is a mistake to assume that one can partition etiologic factors into contributions that sum to 1.0, as in the notion that two-thirds of cancers are due to bad luck and therefore at most one-third could be due to environmental and inherited genetic factors.

Because of joint effects, contributing causes often have attributable fractions that add to more than 1.0. The intellectual disability syndrome secondary to phenylketonuria is a well-known example where the fraction attributable to genetics is 1.0, while the fraction attributable to environment is also 1.0, because the outcome requires both a dysfunctional metabolic gene and an environmental exposure (dietary phenylalanine). As another example, the fact that 100% of prostate cancer is due to a stochastic event (the random inheritance of a Y chromosome) does not relieve us of the need to search out other causes, some of which may be preventable.

Comment

In the 20th century, physics came to recognize that the behavior of matter at its most fundamental subatomic level is stochastic. Knowledge of a physical system's initial conditions is not enough to specify its future, except statistically. If randomness rules at a subatomic level, perhaps it should not surprise us if the cellular and disease processes that are the object of much of our research prove to be largely stochastic. By contrast, the epidemiologic world-view often presumes that we are subject to a kind of health predestination, such that each person would get a disease if an array of necessary causal factors were in place. The same person would not get the disease in an alternate "*counterfactual*" world where one of those necessary factors was removed from each such array (6).

In contrast to that deterministic perspective, the rest of biology has come to acknowledge an important role for random events at the cellular level. For example, there are rule-bound but random errors in DNA replication, which are normally corrected by the cell's repair mechanisms (7). Some of those random errors accrue over a lifetime of cell divisions, and it is believed that certain sets of acquired "driver" mutations can result in neoplastic transformation. The transformed cells and their progeny may die out through cellular evolution and competition with other lineages, through programmed cell death, or through the police work of our immune system; but if they fail to die, a tumor arises and cancer is diagnosed. Random events are fundamental to these processes. However, we have argued that measuring how much of cancer might be preventable raises important issues that are at once conceptual and statistical.

Let us return to the analogy with traffic fatalities (our Figure 1). The use of state-aggregated data can obscure the causes of individual traffic fatalities (point #2 above), such as drunk driving. As shown by the shading of the individual dots, fatal accidents disproportionately involve drunk drivers (with fractions varying from 16% in Utah to 44% in Montana). Nevertheless, those fractions do not statistically explain much of the variability across states: the R^2 for the log of the fatality rate vs the log of the fraction involving a drunk driver is 0.02. If we were to commit the ecologic error described above as point #2 and improperly draw our inference from the aggregated data, we would conclude that drunk driving plays only a negligible role. In fact, the US National Highway Traffic Safety Administration, after studying thousands of individual accidents, has concluded that close to a third of traffic fatalities in the United States are related to drunk driving (8).

While 70% of the variation across states can be statistically explained by the mean number of miles driven per capita, that R^2 does not imply that 70% of traffic deaths can be attributed to miles driven (point #1 above) and it would be incorrect to conclude (point #3) that at most 30% could be due to other causes that might be preventable. In fact, attempts to improve driving safety have succeeded in reducing traffic fatalities by more than 50% in the past few decades. That impressive reduction was not achieved by persuading people to drive less, as people are driving more now than they did 30 years ago.

Similarly, while tissues with a high number of stem cell divisions do provide more opportunities for errors in DNA replication, causative factors can contribute to rates of cell division, to inducing replicative errors, to the failure to repair those errors, and to the failure to clear abnormal cells. We have little control over the total number of times our stem cells divide, but that fact need not impose a limit on the preventability of cancer. Several letter writers to *Science* noted that age-adjusted cancer rates vary considerably among countries and across time, and change with migration (9,10), suggesting that modifiable environmental and lifestyle factors play a major role in carcinogenesis.

In summary, there is little doubt that bad luck plays an important role in the etiology of disease. The observation that variation in site-specific cancer incidence is related to the lifetime number of stem cell divisions across tissues underscores the importance of replication errors in cancer, but those findings have been overinterpreted (http://www.hopkinsmedicine.org/news/media/releases/bad_luck_of_random_mutations_plays_predominant_

role_in_cancer_study_shows). We need to recognize that the proportion of the variance in incidence rates across cancer sites that is statistically explained by replication of stem cells can tell us little about what proportion of individual cases are caused by bad luck. While bad luck, almost by definition, cannot be prevented, the hypothetical preventability of most cancer remains an open question, and the relative importance of random events in cancer causation will continue to defy meaningful quantification.

Notes

The authors have no conflicts of interest to declare.

Funding

This research was supported by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences, under intramural project ES040006.

References

1. Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*. 2015;347(6217):78–81.
2. Couzin-Frankel J. Biomedicine. The bad luck of cancer. *Science*. 2015;347(6217):12.
3. Cohen BL. How dangerous is low level radiation? *Risk Anal*. 1995;15(6):645–653.
4. Wodarz D, Zauber AG. Cancer: Risk factors and random chances. *Nature*. 2015;517(7536):563–564.
5. Kouvaraki MA, Shapiro SE, Perrier ND, et al. RET proto-oncogene: a review and update of genotype-phenotype correlations in hereditary medullary thyroid cancer and associated endocrine tumors. *Thyroid*. 2005;15(6):531–544.
6. Rothman K, Greenland S, Lash T. *Modern Epidemiology*. Third ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
7. Singh NP, McCoy MT, Tice RR, Schneider EL. A simple technique for quantitation of low levels of DNA damage in individual cells. *Exp Cell Res*. 1988;175(1):184–191.
8. Traffic Safety Facts 2012: Alcohol-Impaired Driving. In: *National Highway Traffic Safety Administration Reports*; 2014.
9. Wild C, Brennan P, Plummer M, Bray F, Straif K, Zavadil J. Cancer risk: role of chance overstated. *Science*. 2015;347(6223):728.
10. Gotay C, Dummer T, Spinelli J. Cancer risk: prevention is crucial. *Science*. 2015;347(6223):728.