Progress in Biophysics and Molecular Biology 124 (2017) 41–48

Contents lists available at ScienceDirect

Progress in Biophysics and Molecular Biology

journal homepage: www.elsevier.com/locate/pbiomolbio

Characteristic patterns of cancer incidence: Epidemiological data, biological theories, and multistage models

Josh Hiller*, Celeste Vallejo, Leo Betthauser, James Keesling

Department of Mathematics, University of Florida, USA

ARTICLE INFO

Article history:
Received 11 August 2016
Accepted 5 November 2016
Available online 10 November 2016

ABSTRACT

We investigate and classify several patterns in cancer incidence and relative risk data which persist across different countries and multiple published studies. We then explore biological hypotheses as well as many mathematical models in the literature that attempt to explain these patterns. A general modeling framework is presented which is general enough to model most of observed behaviors. It is our belief that this model has sufficient flexibility to be adapted to new information as it is discovered. As one application of this framework, we give a model for the effect of aging on the process of carcinogenesis.

Contents

1. Introduction

The epidemiology of cancer and the process of carcinogenesis are intimately linked. Because of this, a classification of general patterns of cancer incidence at the international and global levels,

* Corresponding author.
E-mail address: jphiller1@ufl.edu (J. Hiller).

http://dx.doi.org/10.1016/j.pbiomolbio.2016.11.002
0079-6107/

along with a cohesive biological explanation at the individual level is needed. As Baker said in (Baker, 2012) "A major obstacle to winning the war on cancer is a lack of understanding of how cancer develops."

In particular, three epidemiological phenomena have been prominently documented over the last 60 years. These are: incidence which follows a power law at young age, a decrease in risk at extreme old age, and bimodal incidence with no power-law. Surprisingly, no study has examined these three patterns together as part of a larger theme. Perhaps because of this, none of the available mathematical models can satisfactorily account for all three behaviors. Of course much has changed over the last six decades. One can ask to what extent do these patterns persist in the current epidemiological data, and if they do persist, are they local flukes of some particular region or are they truly characteristics of cancer epidemiology on a global scale?

In this review, we describe these patterns in detail, verify that all three are relevant in a global context by examining papers which draw on cancer incidence data from multiple registries from three continents; we provide biological hypotheses to explain each pattern, and we present and analyze relevant mathematical models from the literature. We also propose a unified mathematical framework which generalizes many existing models and which satisfactorily explains all the described behaviors.

The next section provides some basic background in modeling of incidence and relative risk functions. Having explored the necessary mathematical foundation, we describe our methodology. Our results and discussion section is subdivided into four subsections: the first three of which each describe one of the observed behaviors along with pertinent biological hypotheses and mathematical or statistical models from the literature. We emphasize models based on the multistage model of carcinogenesis because of their widespread use in statistics. The fourth subsection discusses relative risk and carcinogens within a general mathematical and epidemiological context. The fifth section provides a flexible mathematical modeling framework, which combines many of the models discussed and is sufficiently robust as to explain all the patterns presented. We then give a specific application of this general framework by constructing a new variation of the multistage model. Finally, the last section of this paper concludes with possible future directions of inquiry.

2. Mathematical background

In this section we present some basic results on incidence and hazard functions in the context of cancer epidemiology. The reader interested in a deeper understanding is directed to (Mdzinarishvili and Sherman, 2013) and (Wienke, 2003) for theoretical derivations and justifications for several of these functions.

The incidence function for a single population is a function of time defined as

$$I(t) = \frac{C(t)}{S(t)},$$

in this equation, $C(t)$ is the number of diagnosed individuals at time t and $S(t)$ is the survivor function at time t. Often, when modeling this function, we operate under the assumption that non-cancer related death affects all segments of the population equally. With this assumption, incidence can be modeled by the simplified hazard function:

$$h(t) = \frac{f(t)}{1 - F(t)},$$

where $F(t)$ is the probability of developing cancer up to time t (commonly referred to as the *cumulative distribution function*), and $f(t) = F'(t)$ (or alternatively $F(t) = \int_0^t f(\tau)\, d\tau$).

We see then that for sufficiently small values of t, $F(t) \approx 0$ and so

$$I(t) \approx h(t) \approx f(t).$$

This is called the *probability density function approximation of incidence* or the *pdf approximation*. One nice property of the hazard function is that it is easy to observe, and with it, one can calculate F by noticing that the definition of h is a separable differential equation, and so $F(t) = 1 - e^{-\int_0^t h(\tau)d\tau}$.

If we have multiple (but finitely many) classes of people at risk, we can compute the sub-population hazard function for sub-population i by taking a base function $h(t)$ and modifying the hazard function using a random variable which acts multiplicatively on the population hazard function $\theta(t)_i$. This variable has the dual effect of adjusting $h(t)$ for both the change of risk of class i and the proportion of individuals in class i. Using this we have that the hazard function for the sub-population is $h_i(t) = \theta(t)_i h(t)$. From here we see that for any given t, that. $h(t) = \sum_{i=1}^{n} h_i(t)$.

If we would like to compare the risk of getting cancer for two sub-populations (say, for example, sub-population A smoked and sub-population B did not), we would employ the *relative risk function*. Which is defined as

$$RR = \frac{Pr(A)}{Pr(B)},$$

here $Pr(C)$ is the probability that an individual from sub-population $C \in \{A, B\}$ develops cancer.

Finally, two distributions which figure prominently in the cancer modeling literature are the compound Poisson distribution and the Weibull distribution. The cumulative distribution function of the former is

$$F_p(t) = \Pi_{i=1}^{n} \left(1 - e^{-\lambda_i t}\right),$$

where the average waiting time for the ith process to occur is $\frac{1}{\lambda_i}$. The Weibull cumulative distribution function takes the form

$$F_w(t) = 1 - e^{-\left(\frac{t}{\lambda}\right)^n}.$$

3. Patterns of cancer incidence

We will now explore three patterns that each appear many times throughout the literature, regardless of the source of the data or the methodology of collection:

- a power law at early age,
- bimodal incidence (or early peak incidence) with no power law,
- a leveling off or decrease at old age.

3.1. Power law at an early age

Among the earliest general observations on cancer incidence functions is that for many cancers, and across many populations, cancer mortality (and incidence) tends to follow a power law for most of an individual's average lifespan (Armitage and Doll, 1954; Nording, 1951) (in fact this phenomenon is so common that some authors have taken to calling a cancer which exhibits this behavior a "log-log cancer" (Pike et al., 1983)). Biological attempts to explain