CUMULATIVE INCIDENCE RATES OF CANCER

C. T. Lenard,∗ La Trobe University
T. M. Mills,** Loddon Mallee Integrated Cancer Service
Ruth F. G. Williams,*** La Trobe University

Abstract

There is widespread interest in the risk of being diagnosed with cancer. Internationally, especially in developed countries, governments collect and use data on the incidence of cancer for strategic planning to ensure that the nation has the resources that will be required to deal with the disease. Incidence data can also be used to assess the effectiveness of public health campaigns. However, there are several measures for quantifying the incidence of cancer. This paper examines one of them, namely the cumulative incidence rate. We present a review of the method for estimating the cumulative incidence rate of cancer in a population, and for comparing these rates in two populations. We explore the connection between the cumulative incidence rate and the cumulative risk of being diagnosed with cancer by a certain age, with details of the mathematical ideas that underpin these concepts. This expository paper has been written to be useful to researchers and policy makers who have an interest in the incidence of cancer, at a local or regional level, and who wish to understand the details associated with cumulative incidence rate and cumulative risk.

Keywords: Risk; probability; cancer; incidence

2010 Mathematics Subject Classification: Primary 92C60
Secondary 62P10; 60-01; 91B82

1. Introduction

‘Cancer is a leading cause of death worldwide’ (World Health Organisation [10]). One can expect that any new national report on cancer will generate headlines in the mass media on the increasing incidence of cancer in the community because there is widespread interest in cancer statistics among ordinary citizens as well as healthcare professionals and policy makers.

In digesting such statistical reports, it is important, particularly for a reader with a professional interest, to understand the relevant concepts. For example, a recent report on cancer in Australia [1] contained the statement: ‘By the age of 85 years, 1 in 2 males and 1 in 3 females will have been diagnosed with cancer at some stage in their life’. Our aim is to explain the meaning of this statement and the underlying mathematics.

The authors of this report [1] use the term cancer to refer broadly to all cancers combined except basal and squamous cell carcinomas of the skin; we do likewise.
The incidence of cancer is defined to be the number of new cases of cancer diagnosed in a given period (usually one year) and in a given geographic region; thus, incidence is an integer. There are several related measures such as incidence rate, age-specific incidence rate, and age-standardised incidence rate. Another measure that has gained popularity in cancer statistics is the cumulative incidence rate. Henceforth, we will refer to ‘cumulative incidence rate’ simply as ‘cumulative rate’.

In this paper we present an exposition of the definition of cumulative rate, the underlying mathematical theory, and an outline of how this measure is applied in practice. More specifically, we will address the following questions.

1. What is the meaning of cumulative rate?
2. How does one interpret cumulative rate?
3. What is the mathematical basis that underpins this rate?
4. What is the connection between the cumulative rate and the risk of being diagnosed with cancer?
5. How does one estimate the cumulative rate for a given population?
6. How would one compare the cumulative rate of cancer for one population with that of another?

En route, we will review the literature surrounding this measure as some of the references may not be readily available to the reader.

We hope that our exposition will prove useful to those with a special interest in cancer statistics especially as they apply to a particular geographic region.

2. Estimating the rate for a single population

In this section we introduce the cumulative incidence rate for cancer in a single population. The cumulative rate is used to approximate the cumulative risk. Hence, we first introduce the concept of the cumulative risk of being diagnosed with cancer.

2.1. Cumulative risk

In this subsection we present the mathematical approach to defining the cumulative risk of developing cancer by some given age. The notation, terminology, and discussion are taken from [8]; we have included this material to make the presentation as self-contained as possible.

Suppose that we are interested in a particular population. Define two random variables as follows. For a person chosen at random from the population:

\[ T := \text{age of the person either when the person is diagnosed with cancer, or when an intervening event that prevents followup occurs (for example, death from some other cause without ever being diagnosed with cancer)}, \]

\[ \delta := \begin{cases} 1 & \text{if the person is diagnosed with cancer}, \\ 0 & \text{otherwise}. \end{cases} \]

These random variables give rise to the following functions (where \( \mathbb{P}(E) \) denotes the probability of an event \( E \) occurring):

\[ R(t) := \mathbb{P}(T < t), \]

\[ S(t) := \mathbb{P}(T \geq t) = 1 - R(t). \]
Cumulative incidence rates of cancer

\[ p_1 := \mathbb{P}(\delta = 1), \quad (3) \]
\[ R_1(t) := \mathbb{P}(T < t \mid \delta = 1), \quad (4) \]
\[ \pi(t) := \mathbb{P}(T < t \text{ and } \delta = 1) = p_1 R_1(t). \quad (5) \]

The function \( \pi(t) \) is the probability of being diagnosed with cancer by age \( t \).

We now define the function \( \lambda(t) \) which is the instantaneous incidence rate at age \( t \):

\[
\lambda(t) := \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}(t \leq T < t + \Delta t, \delta = 1 \mid T \geq t) \\
= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\mathbb{P}(t \leq T < t + \Delta t, \delta = 1, T \geq t)}{\mathbb{P}(T \geq t)} \\
= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\mathbb{P}(\delta = 1) \mathbb{P}(t \leq T < t + \Delta t \mid \delta = 1)}{\mathbb{P}(T \geq t)} \\
= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{p_1 (R_1(t + \Delta t) - R_1(t))}{1 - R(t)} \\
= \frac{p_1 R_1'(t)}{1 - R(t)} \\
= \frac{p_1 R_1'(t)}{S(t)}.
\]

Therefore,

\[
\lambda(t)S(t) = p_1 R_1'(t). \quad (6)
\]

Using (3), (4), (5), and (6) we find that the probability of being diagnosed with cancer by age \( t \) is as follows:

\[
\pi(t) = \mathbb{P}(T < t \text{ and } \delta = 1) \\
= p_1 R_1(t) \\
= \int_0^t p_1 R_1'(u) \, du \quad \text{(because } R_1(0) = 0) \\
= \int_0^t \lambda(u) S(u) \, du.
\]

The cumulative risk of being diagnosed with cancer was defined by Day [5] as ‘the risk an individual would have of developing the disease in question during a certain age period if no other cause of death were in operation’. We will discuss this surprising assumption in Section 5. (Note that Day used the phrase ‘developing the disease’ whereas we would say ‘being diagnosed with the disease’. While these phrases have different meanings, there is no difference in the modelling so long as one is consistent.) From a mathematical point of view, this assumption means that

\[
p_1 = \mathbb{P}(\delta = 1) = 1; \quad (7)
\]

therefore,

\[
R(t) = \mathbb{P}(T < t) = \mathbb{P}(T < t \mid \delta = 1) \mathbb{P}(\delta = 1) = \mathbb{P}(T < t \mid \delta = 1) = R_1(t).
\]
Hence, we can calculate the cumulative risk, using (2), (6), and (7), as follows:

\[ p_1 R'_1(t) = \lambda(t) S(t), \]

\[ R'_1(t) = \lambda(t) (1 - R(t)), \]

\[ \frac{R'(t)}{1 - R(t)} = \lambda(t) \quad (\text{since } R_1(t) = R(t)), \]

\[- \ln(1 - R(t)) = \int_0^t \lambda(u) \, du \quad (\text{since } R(0) = 0), \]

\[ \ln(1 - R(t)) = - \int_0^t \lambda(u) \, du, \]

\[ 1 - R(t) = \exp\left( - \int_0^t \lambda(u) \, du \right). \]

So we obtain

\[ R(t) = 1 - \exp\left( - \int_0^t \lambda(u) \, du \right). \quad (8) \]

Thus, (8) gives the formula for \( R(t) \), the cumulative risk of being diagnosed with cancer by age \( t \). In this case, where \( p_1 = 1 \), we have

\[ R(t) = \pi(t). \]

Equations (1) and (8) imply that the cumulative risk by age \( t \) is a probability and lies in the interval [0, 1]. One could express the cumulative risk as a percentage.

2.2. Cumulative rate

The cumulative risk by age \( t \) as expressed by (8) involves an integral. The cumulative rate at age \( t \) is an approximation to the cumulative risk, where the integral in (8) is approximated by Riemann sums. We will see that the cumulative rate can be calculated easily from basic demographic and incidence data: herein lies the appeal of this measure.

We illustrate with an example. Suppose that we wish to estimate \( R(75) \), the cumulative risk of being diagnosed with cancer by age 75. Assume that we have demographic and incidence data for one year, aggregated into five-year age groups, as in Table 1.

We approximate \( \int_0^{75} \lambda(u) \, du \) as follows:

\[ \int_0^{75} \lambda(u) \, du \approx 5 \sum_{i=1}^{15} \frac{x_i}{n_i}. \]

<table>
<thead>
<tr>
<th>Age group</th>
<th>Population</th>
<th>Incidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–4</td>
<td>( n_1 )</td>
<td>( x_1 )</td>
</tr>
<tr>
<td>5–9</td>
<td>( n_2 )</td>
<td>( x_2 )</td>
</tr>
<tr>
<td>10–14</td>
<td>( n_3 )</td>
<td>( x_3 )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>70–74</td>
<td>( n_{15} )</td>
<td>( x_{15} )</td>
</tr>
</tbody>
</table>

Table 1: Data in five-year age groups.
Hence, the cumulative rate of the incidence of cancer by age 75, which we will denote by $a(75)$, is given by

$$a(75) := 5 \sum_{i=1}^{15} \frac{x_i}{n_i}.$$ 

Although there is no mathematical reason for this rate to lie in the interval $[0, 1]$, in practice $x_i$ is usually much smaller than $n_i$ and the cumulative rate lies in the interval $[0, 1]$. Often the cumulative rate is expressed as a percentage by multiplying by 100. However, this could mislead a reader into believing that the cumulative rate is a probability. Furthermore, as Bray [4] pointed out: 'The cumulative rate is not in fact a rate, but a dimensionless quantity. In other words, it is not expressed in units “per annum” but simply as a number’.

The cumulative rate is used to approximate the cumulative risk of being diagnosed with cancer by a given age:

$$R(75) \approx 1 - \exp(-a(75)).$$

This is the basis of the calculation that underpins the statement from [1] quoted in Section 1: 'By the age of 85 years, 1 in 2 males and 1 in 3 females will have been diagnosed with cancer at some stage in their life'.

Since $1 - \exp(-x) \approx x$ when $x$ is small and positive, if $a(75)$ is small then $R(75) \approx a(75)$. In this sense, the cumulative rate may be regarded as an approximation to the cumulative risk, which is a probability.

For further discussion of the cumulative rate, see [2] and [3].

### 2.3. A confidence interval for the cumulative rate

In this subsection we consider the question of finding a confidence interval to estimate the cumulative rate. This requires us to interpret the data in Table 1 in a different context; we will adapt some ideas from Dobson et al. [6].

Let us assume that, for each age group $i (i = 1, 2, \ldots, k)$, the incidence of cancer is a random variable $X_i$ which has a Poisson distribution with mean $\theta_i$. This is a reasonable assumption if the number of persons in the age group is large and the probability of being diagnosed with cancer is small. Then we have

$$\mathbb{E}(X_i) = \text{var}(X_i) = \theta_i, \quad i = 1, 2, \ldots, k,$$

where $\mathbb{E}(X)$ denotes the expected value of $X$ and $\text{var}(X)$ denotes the variance of $X$. It is reasonable to assume that $X_1, X_2, \ldots, X_k$ are independent random variables. Then, in Table 1, $x_i$ is the observed value of the random variable $X_i$. Hence,

$$x_i = \text{est} \mathbb{E}(X_i) = \text{est} \text{var}(X_i),$$

where ‘est’ indicates the estimated value. The $n_i$ in Table 1 are known constants.

Now, define the random variable

$$A(75) := 5 \sum_{i=1}^{15} \frac{X_i}{n_i}.$$ 

Then $a(75)$, the cumulative rate at age 75, is the observed value of $A(75)$. Furthermore,

$$\mathbb{E}(A(75)) = 5 \sum_{i=1}^{15} \frac{\theta_i}{n_i}.$$
and, using the independence assumption above about the \( X_i \),

\[
\text{var}(A(75)) = 25 \sum_{i=1}^{15} \frac{\theta_i}{n_i^2}.
\]

Hence, we can estimate the standard deviation of \( A(75) \) by

\[
\text{est SD } A(75) = s(75) = 5 \sqrt{\sum_{i=1}^{15} \frac{x_i}{n_i^2}}.
\]

Finally, we can calculate an approximate 95% confidence interval for \( \mathbb{E}(A(75)) \) as follows:

\[
a(75) \pm 1.96s(75).
\]

Since \( R(75) = 1 - \exp(-a(75)) \), and \( 1 - \exp(-x) \) is a monotonically increasing function of \( x \), the confidence interval corresponding to the cumulative rate \( a(75) \) can be transformed into a confidence interval corresponding to the cumulative risk \( R(75) \).

3. Comparing rates in two populations

Day [5] was the first researcher to promote the measure of cumulative rate in the cancer literature. He introduced it as an alternative to the traditional age-standardised incidence rate. He pointed out that this measure removes ‘the arbitrariness in choosing a standard population’ [5].

Thus, the real value of the cumulative rate arises in comparing the incidence of cancer in two different populations, or in the same population at different times.

Suppose that we have two independent populations which we denote by subscripts ‘1’ and ‘2’. Assume that we have demographic and incidence data for each population as in Table 1. We test the null hypothesis \( H_0 : \mathbb{E}(A_1(75)) = \mathbb{E}(A_2(75)) \) against the alternative hypothesis \( H_1 : \mathbb{E}(A_1(75)) \neq \mathbb{E}(A_2(75)) \) with the test statistic

\[
z = \frac{a_1(75) - a_2(75)}{\sqrt{s_1(75)^2 + s_2(75)^2}},
\]

which has, approximately, a standard normal distribution when \( H_0 \) is true.

4. Variations

In the above calculations, we have considered five-year age groups and an age limit of 75 because these are common assumptions that are used in practice.

However, there are situations in which not all the widths of the age groups are equal to five. For example, because bowel cancer is rare among younger people, one might be presented with incidence data with age groups 0–29, 30–34, 35–39, . . . , 70–74. The above procedures for calculating the cumulative rate and risk can be easily adapted for such situations.

Also, we can readily replace the age 75 by other age limits as we might do in considering childhood cancers [5]. Indeed, the general methods can be adapted to calculate the cumulative rate for specific cancers.

If we note that \( A(75) \) is a weighted sum of Poisson random variables, then we can use the methods in [6] to find other approximate confidence intervals for \( \mathbb{E}(A(75)) \). The methods in [6] have an aesthetic appeal because they take into account the fact that the Poisson distribution is asymmetric: our method does not take this into account, as is evident from (9).
5. Conclusion

Our aim in writing this expository paper was to present a discussion of the cumulative rate on the incidence of cancer, and the associated cumulative risk of developing cancer by a given age.

A surprising assumption that underpins these calculations is that all causes of death other than cancer are excluded from consideration. One might regard this assumption as unrealistic, which it is. However, the main use of the cumulative rate (and risk) is to find some common ground for comparing the incidence of cancer in two populations. Calculations involved in age-standardisation rates involve assumptions about the age-distribution of the population that are also unrealistic.

Risk perception is associated with many issues and is not well understood [7]. For this reason, the assumption about excluding all causes of death other than cancer from consideration should be stated when presenting the results of calculations of cumulative rate or risk [9]. For an excellent discussion of the perceptions of risks associated with cancer, see [7].

Also, we prefer the phrase ‘cumulative risk by age 75’ to the less precise term ‘lifetime risk’. There is a big difference between the impacts of the statements ‘One in three people will get cancer during their lifetime’ and ‘If we assume that everyone will be diagnosed with cancer sometime, then one in three people will be diagnosed with cancer by age 75’.

The cumulative rate is suitable for comparing the incidence of cancer in two populations, or in one population at two different times. The cumulative rate for a single population and a given time is a measure that is not particularly informative.

We conclude with some sage advice from Boniol and Heanue [2]. ‘Therefore it cannot be stressed enough that neither the age-standardised rate nor the cumulative rate are alternatives to the age-specific incidence rates, which should always be the starting point and foundations for any thorough analysis of the incidence data’.

References