Simulation modelling to validate the flow method for estimating completeness of case ascertainment by cancer registries

Paul B. S. Silcocks, David Robinson

1Trent Cancer Registry, Fulwood House, 5 Old Fulwood Road, Sheffield S10 3TG, UK
2Trent Research and Development Support Unit, University of Nottingham, Nottingham NG7 2UH, UK
3King's College London, Thames Cancer Registry, 1st Floor Capital House, 42 Weston Street, London SE1 3QD, UK

Address correspondence to David Robinson, E-mail: dave.robinson@kcl.ac.uk

ABSTRACT

Background To validate estimates of completeness of cancer ascertainment obtained by the flow method.

Methods We generated a computer simulation of patient-level cancer registration processes, based loosely on the age distribution and survival of colorectal carcinoma patients, and utilizing a mixture of 'cured' and 'killed' subjects with an age-dependent fraction of 'cured' cases. The simulated data were then used in an analysis of completeness using the flow method. Validation of the simulation process was based on similarity of outputs to those obtained using real data, and validation of the flow method on its ability to correctly estimate the known proportion of cases in the simulated data which would never be registered.

Results We successfully generated realistic data and have shown that completeness estimated by the flow method is close to the true value, whereas another method of estimating completeness (Ajiki's) was shown to be strongly biased. We also modelled what happens to completeness estimates when a new registry is set up.

Conclusions When its assumptions are met (steady state for incidence, survival and stable population structure), the flow method works well but is biased for cancers with good survival. Further research is required to assess the robustness of the method when these conditions are not met.

Keywords cancer, epidemiology, statistical methods

Introduction

The flow method introduced by Bullard et al. was a major advance in addressing the issue of completeness of case ascertainment by cancer registries. For the first time, a method of estimating completeness was available which explicitly modelled the time dependence of the registration process, yet could be applied routinely to large data sets and to a large number of different tumour sites, requiring only a general-purpose software package. (Copies of the completeness software are available on request from the authors.)

Briefly, the flow method estimates completeness of registration as a function of three time-dependent probabilities that can be calculated using routine cancer registry data:

- $s(t_i)$ is the probability that a cancer patient is still surviving at time $t_i$ after diagnosis, estimated from a life table analysis on a sample of incident cases;
- $m(t_i)$ is the probability that the death certificate of a patient who dies in the time interval $(t_i, t_{i+1})$ after diagnosis includes a mention of cancer. This is estimated from the same sample as the life table as follows: Deaths with mention/All cancer patient deaths in the interval;
- $n(t_i)$ is the probability that a patient surviving until time $t_i$ after diagnosis is still unregistered. This is based on a survival model with registration as the event and censoring at death.

These three probabilities are combined to give the proportions ($i$) alive but unregistered, i.e. missing ($M$) and (ii) dead and unregistered and cancer not mentioned, i.e. lost ($L$) at time $T$ after diagnosis as:

\[
M(T) = s(t_i) \cdot n(t_i) \quad \text{and} \quad L(T) = \sum_{i=0}^{n} \left[ (s(t_i) - s(t_{i+1})) \cdot (1 - m(t_i)) \cdot n(t_i) \right]
\]
with \( t_e \leq T < t_{e+1} \). Completeness at time \( T \) is then given by
\[
C(T) = 1 - M(T) - L(T).
\]

However, no evidence of the validity of the procedure was given in the original paper—instead, examples were given and the assumptions (a steady state for survival, ascertainment rates, etc) stated. In practice, of course, cancer incidence changes as the population ages, older patients survive less well and may be less accurately ascertained. In addition, secular trends in general population mortality operate and improvements in medical care for the condition of interest further confuse the issue. A program (comlims) developed at the Thames Cancer Registry (TCR), which implemented the flow method using the statistical package Stata,² has been adopted by a number of registries, but the lack of validation still causes some concern.

Two issues need to be distinguished with respect to validation. First, no ‘gold standard’ method exists against which the flow method can be validated, and the flow method could arguably be regarded as the gold standard itself. Secondly, no gold standard data exist. There is no real data set for which the true number of cases is known and which covers a long time period during which cancer incidence rates, age structure and survival have all remained constant.

Recognizing these limitations, this paper attempts to address the problem by generating simulated data that realistically mimic the salient features of registry data. The face validity of the simulation process (i.e. whether the procedure seems reasonable) is demonstrated by the use of parameters based on real data, and the generation of ‘realistic-looking’ incidence and survival data, whereas the criterion validity of the simulation process (i.e. comparison with some objective standard) is shown by the ability of the simulated data to mimic features observed in real registry data. Finally, the estimation validity of the flow method itself is assessed by comparing its estimate of completeness with the true completeness in the simulated data.

### Materials and methods

The simulation proceeds in two phases. Phase I consists of generating incident cases mimicking a specific cancer type (e.g. colorectal) for each year of a user-specified period of many years’ duration. The data items generated are:

- year of incidence;
- ‘cured’ flag;
- date of death;
- birth date;
- ‘mention’ flag (i.e. whether cancer was mentioned on death certificate);
- death certificate initiated (DCI) flag (indicating that registration was based on a death certificate);
- death certificate only (DCO) flag (i.e. a DCI which could not be linked to additional survival data through trace-back in hospital records);
- observed diagnosis date (which allows for incomplete trace-back of DCI cases);
- date of registration;
- ‘lost’ flag—used to give the true proportion of missing cases;
- ID number.

The data for individual years are then combined into a single data set, which Phase II converts to mimic the database that would actually be held by a registry and generates the files that would be used to estimate completeness. For this purpose, two files are extracted from the database: a file of incident cases for a given year, usually 5 years or so earlier than the extraction date, and a file of deaths (typically for a year midway between the extraction and incident years). In a steady state, the choice of these files should have little impact on the estimated completeness. In practice, the most recent years would be chosen, bearing in mind that at least 5 years of follow-up are required for the incident cases and that the deaths should be as complete as possible.

Phase I is the core of the exercise and is therefore described in more detail.

### Phase I

#### Stage 0: set-up

In this stage, the notional start year of registrations is specified, together with the number of years (e.g. 30) for which data are required. The simulation must allow for the fact that when patients with cancer are registered, inevitably some registrations are only initiated after the patient’s death—these are termed DCI registrations. Most of these can be subsequently linked to hospital records from which the original date of diagnosis can be established. However, in a proportion of cases, this trace-back is unsuccessful and the registration remains on the record as a DCO registration. Since the only evidence of the cancer for such cases is the death certificate, the recorded date of diagnosis is the same as the date of death and survival time is zero.

Survival is modelled as a mixture of patients who are ‘cured’ (who experience the age-specific mortality rates of the general population from the time of diagnosis) and ‘killed’ (who die as a result of their disease and who experience much poorer survival). Not all cancers are statistically curable in this sense, but the model can mimic this in principle...
by reducing the survival of cured cases or the proportion that are cured.

In addition, parameters are specified which give:

- the expected number of cases per year;
- the probability that the death certificate mentions cancer (this depends on age at death and survival time);
- the probability that a case will be DCI given that the death certificate mentions cancer;
- the fractional shortening of the true survival time for partially traced-back DCIs;
- the proportion of DCI cases which become DCO registrations.

Unless otherwise stated, parameter values displayed correspond to estimates made for colorectal cancer using Trent Cancer Registry incidence data for 1999 (and deaths in 2003).

**Stage 1: generate the age distribution of cases diagnosed in a given year**

This employs a scaled beta distribution with a range of 100 years, mean of 71.34 years and standard deviation of 11.94 years. Clearly, the mean and variance should be appropriate for the chosen tumour site. Ageing of the population could in principle be modelled by changing these parameters.

**Stage 2: set up true survival times for ‘cured’ and ‘killed’ cases**

In this stage, cases are randomly defined as ‘cured’ or not, based on the following logistic model:

\[
Pr(\text{cure}) = \begin{cases} 
1 & \text{if age} \geq 35 \\
\frac{1 + e^{0.336+0.0137'(age-35)}}{0.416781451} & \text{otherwise}
\end{cases}
\]

These parameters are based on 5-year relative survival figures for colorectal cancer, assuming that 5-year age-specific relative survival is a measure of cure, and fitting a logistic response with age as explanatory variable.

Survival times for ‘killed’ subjects are then generated using an age-dependent Weibull distribution with parameters:

\[
\lambda = e^{(-1.3316+0.014894\times\text{age})} \\
\rho = 0.8306871
\]

and survival function \( S(t) = \exp (-\lambda^{\rho}) \).

The parameter values are those obtained from a mixture model analysis of colorectal cancer data previously carried out within the Trent Cancer Registry.

To mimic lung cancer, which has poor survival, we assume all cases are ‘killed’ with survival modelled by a generalized F distribution, the four parameters of which confer flexibility in shape. Poorer survival with advancing age is modelled for these sites by an age-dependent hazard ratio, the logarithm of which is defined by a polynomial function. For breast cancer, which has relatively good survival but with a low proportion cured—patients experience mortality rates in excess of the general population even after 20 years—a similar approach is adopted, except that we assume that patients aged over 50 at diagnosis are cured if their initially modelled survival time exceeds 15 years. Note that this does not preclude the use of a pre-defined cured fraction if so desired.

The survival of the general population experienced by ‘cured’ subjects is modelled by a Gompertz–Makeham distribution with parameters

\[
A = 0.000351 \\
B = \exp(-10.8029) \\
C = \exp(0.099257) \\
M = \frac{B}{\ln(C)}
\]

with the probability of living from birth to a specified age being given by:

\[
S(\text{age}) = \exp[-A \times \text{age} - M \times (C^{\text{age}} - 1)]
\]

The parameters for the Gompertz–Makeham distribution are those estimated by non-linear maximum-likelihood-based Poisson regression, using England and Wales mortality statistics for 2001. By the rules of conditional probability, the probability of survival of a subject aged \( a \) years for another \( t \) years is given by:

\[
S(a + t) = \frac{S(0, a + t)}{S(0, a)}
\]

where \( S(a + t) \) is the probability of survival for additional \( t \) years from age \( a \) years, \( S(0, a + t) \) the probability of survival from birth to age \( a + t \) years, and \( S(0, a) \) the probability of survival from birth to age \( a \) years.

This has no closed-form solution for \( t \), which is obtained numerically by a Newton–Raphson iteration.

**Stage 3: convert true survival times into diagnosis, death and birth dates**

For each subject, the true diagnosis date is drawn from a uniform distribution of dates in the current year and the date of birth obtained as the diagnosis date minus age. Date of death for each subject is then assigned using date of diagnosis + survival time.
Stage 4: generate a mention of cancer on the death certificate and identify DCI and DCO cases
The probability of a mention of cancer on the death certificate is dependent on age at death and survival in years. Suitable parameter values were obtained from Trent Cancer Registry colorectal cancer deaths data for 2003, the formula used being:

\[
\Pr(\text{mention}) = \frac{1}{1 + e^{-7.716 + 0.0665 \times \text{age at death} + 0.3748 \times \text{survival}}}
\]

This formula is evaluated for each case and compared with a uniform random number to determine whether the case should be recorded as having a ‘mention’ or not. Slightly different parameter values are currently used for the other tumour sites simulated.

A study on survival of DCI cases has suggested that if on average \(~40\%\) of the true survival were untraced for colorectal cases, this would account for all the worse survival of DCI cases apart from selection bias. A figure of \(20\%\) for the proportion of the true survival that is untraced on average was chosen for the simulation so that \(k\), the proportion untraced, has a uniform distribution in the range \(0–0.4\); the apparent survival (years) being given as:

\[
\text{Apparent survival} = (1 - k) \times \text{True survival}
\]

(and rounded to three decimal places). This process mimics incomplete trace-back as when a case is recorded on the basis of a death certificate, but the record in the treating hospital only records admission for treatment of metastases (the record of the original admission for the primary tumour having been several years earlier, maybe in a different hospital, and now unavailable). This value of \(k\) is also used for cancers with poor survival such as lung, but for cancers with better survival such as breast, \(k\) is taken to be \(0.1\).

Next, a DCI flag is generated conditional on having a ‘mention’ (and which is therefore indirectly dependent on age at death and survival time):

\[
\Pr(\text{DCI}|\text{mention}) = 0.339252
\]

The value shown is for colorectal cancer, with slightly different values for simulated lung and breast cancers. Again, for each case, the setting of the flag was determined by comparison of this probability with a uniform random number.

DCOs are generated at this point as a random sample of DCIs, the fraction being estimated from registry-based data for the tumour site being considered.

Stage 5: define registration dates
An age-dependent registration date is created for cases registered (hence ‘known’) while alive, the time from diagnosis to registration being a random exponential function with hazard \(1/14\) on top of a minimum delay of 3 days.

For DCI cases, an age-independent interval from date of death to registration was generated—related to Office for National Statistics practice—by another exponential function with hazard \(1/14\) on top of a minimum delay of 3 days.

Stage 6: create ‘lost case’ flag
This is a simple step to flag ‘lost’ cases, defined as follows. Cases registered while alive must have registration date prior to the date of death. Technically, ‘missing’ cases are still alive and, though not yet registered, are potentially still registrable while alive. ‘Lost’ cases are those not registered while alive, who have died without mention of cancer on the death certificate, and it is these cases which are denoted by the ‘lost’ flag. Current parameter settings for simulated colorectal cancer result in \(~2.5\%\) ‘lost’ cases, this true proportion being obtained from the whole simulated data set. A key part of the complims output is an estimate of this proportion.

Stages 7 and 8: drop surplus variables, save to file and proceed to the next year
Finally, only the variables that would be seen in a registry database are saved and the data from individual years are pooled into a single ‘full data’ file representing ‘the truth’.

Phase II
Stage 9: create data ‘as observed’ at a registry and create deaths and diagnoses files
After pooling the data, the first step is an extraction process to mimic the data that would be available to the registry at a given time point. Thus, for simulated data covering the period 1975–2015, with an extraction of date 31 December 2005, deaths and diagnoses files must be created for use by complims—for example, a file of cases diagnosed in 1999 and a file of cases who died in 2003.

For the diagnoses file, this involves:

- dropping cases with diagnosis or registration dates later than the extraction date;
- dropping cases that are ‘lost’;
- setting the death date, DCO flag and DCI flag to missing if death is later than the extraction date;
- keeping only cases with diagnosis year equal to that specified.
For the deaths file:

- keeping cases if the year of death is that specified for the deaths file;
- dropping cases that are ‘lost’.

At each run of the simulation, the parameters can be altered to assess the effect, for example, of a trend in survival, or a different age distribution or the effect of setting up a new registry.

**Results**

**Face validity**

A graph of the life table function (used to generate survival in ‘cured’ cases) is displayed in Fig. 1, corresponding to an expectation of life of 79 years. The survival of ‘killed’ subjects is shown in Fig. 2, demonstrating the variation with age. The overall survival (which may be compared with real data for colorectal cancer from Trent) is shown in Fig. 3, and the proportion of DCO and DCI cases and the age distribution of real and simulated data are shown in Table 1. In general, the patterns are similar—exact agreement is not expected here.

**Criterion validity**

Although the function \( s(t) \) estimated by complims is a standard Kaplan–Meier survival curve, the other functions are specific to complims. The simulated data should generate results with features qualitatively similar to those seen in the analysis of real data. Figure 4a–c shows how the output of complims is similar for both real (Trent Cancer Registry 1999 registrations and year 2003 deaths) and simulated data. For the purpose of smoothing, when estimating the probability of a ‘mention’, complims uses pooled 30-day mean values, and these are plotted in Fig. 4b.

**Estimation validity**

This is demonstrated in Table 2, which shows how the percent completeness estimated by complims approaches the true value with time.

**The effect of a new registry**

Simulated data were generated for cases diagnosed between 1975 and 2014. An extract ‘as of’ 31 December 1980 was then performed to provide a snapshot of the registry database 5 years after its initial set-up, and complims was run using a file of cases diagnosed in 1975 and a file of cases who died in 1979, with follow-up until the end of 1979. This gave a 5-year estimate of completeness of 96.5%. Repeating the process using an extract ‘as of’ 31 December 2005 (i.e. when the registry was well established), and running complims with diagnoses from 2000, deaths from 2004 and follow-up to the end of 2004 gave a 5-year completeness estimate of 96.7%. As the true percentage of ‘lost’ cases was 2.5%, applying the flow method too soon after the set-up of the registry gave a falsely (but marginally) low estimate of completeness.

These findings reflect the situation observed when applying the flow method to real data from the National Cancer Registry of Ireland, which was set up in 1994. Applying the method in 2001, using 1994 diagnoses and 1997 deaths, followed-up to the end of 1998 gave a 5-year completeness estimate for colorectal cancer of 97.0%. Repeating the process in 2005, using 1998 diagnoses and 2001 deaths, followed-up to the end of 2002 gave a corresponding estimate of 98.5%.

**Comparison with Ajiki’s method**

Ajiki et al. described a method for estimating completeness using the proportion of DCI registrations and the ratio of
deaths to registrations. The formula is:

\[
\text{% completeness} = 100 \times \frac{(100 - \% \text{DCI} \times (\text{Incidence/Deaths}))}{(100 - \% \text{DCI})}
\]

and it can be shown to be equivalent to estimating the living but unregistered cases by capture–recapture (where the two sources are taken to be deaths and registrations during life).

A key assumption of the method is that the case fatality is the same among cases notified during life as in those missed. Other assumptions are that the pattern of survival and incidence rates remain constant over time and that the cause of death is accurately specified on death certificates.

Ajiki’s method was applied to simulated data for extraction years 1975–2014. The estimated percentage complete rose rapidly the later the extraction year (i.e. the closer the data were to a steady state). Nevertheless, even based on the final year the Ajiki method grossly underestimated completeness, giving a value of 86.57% as opposed to the true value of 97.49%.

**Effects of good/poor survival**

We have also investigated simulated and real data for lung and breast cancer. Once again salient features were well mimicked.

<table>
<thead>
<tr>
<th>Age at diagnosis</th>
<th>Trent</th>
<th>Simulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (years)</td>
<td>71.3</td>
<td>71.4</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>11.9</td>
<td>12.0</td>
</tr>
<tr>
<td>% DCI</td>
<td>25.5%</td>
<td>23.9%</td>
</tr>
<tr>
<td>% DCO</td>
<td>5.5%</td>
<td>5.3%</td>
</tr>
</tbody>
</table>

In terms of estimation validity, for lung the true completeness was 98.63%, whereas the estimates from the flow method (based on nominal years of incidence and deaths of 1999 and 2007, respectively) rose from 91.26% at 1 year through 98.88% at 5 years to 99.09% at 15 years after diagnosis. For breast cancer, the true completeness was 98.05%, with estimates of 65.39% at 1 year through 93.02% at 5 years and 95.05% at 15 years after diagnosis. However, in contrast, Ajiki’s method gave completeness estimates of 77.65% and 80.72% for lung and breast, respectively (Table 2).

**Discussion**

**Main findings of this study**

We have been able to create simulated cancer registry data in a steady state and for which the completeness is known, for direct comparison with the estimate generated by the flow method using these simulated data. The simulated data are free of secular trends in incidence, survival, diagnostic fashion and population migration and can be created for any length of time—longer indeed than any real cancer registry has existed. The simulated data display emergent properties (that is, properties not directly programmed) which resemble those of comparable real data. We have also been able to mimic observations seen when a new registry is set up.

For the lung cancer model, the long-term estimated proportion complete was actually slightly higher than the true value, although it is unlikely that in practice completeness estimates would be based on a 15-year follow-up and both values were very high. For breast cancer, the estimated completeness approached the true value but underestimated it even at 15-years follow-up. The fact that this gives a conservative estimate may reassure those who are concerned about possible under-ascertainment of cases. The results for breast cancer are intuitively understandable because the long
survival allows cases to die of other causes. Ajiki’s method grossly underestimates the completeness for all three cancer models.

What is already known on this topic

Previously, methods for estimating completeness have been validated against some other method (e.g. death certificates, intensive case-finding or multiple data sources). None of these is satisfactory for the flow method as they ignore the dimension of time. Moreover, none of these validation methods can provide the ‘true’ value of the quantity being estimated.

What this study adds

This is the first example of validation of a method for estimating completeness of ascertainment by simulation rather than against some other method. We have also shown that a standard method for estimating completeness based on the percentage of DCI cases, incidence and mortality seriously overestimates the proportion of missing cases even when the necessary assumptions are met.

Our conclusion is that we can confirm that, under ideal circumstances and if its assumptions are met, the flow method does indeed accurately estimate completeness. However, this is an asymptotic result that may not be
achieved in a reasonable length of time for cancers with good survival. Completeness is slightly overestimated for cancers with poor survival.

Limitations of this study
We have yet to explore the effects of secular trends in survival and incidence, and we plan a future publication which will address these issues.

Competing interests
David Robinson was a co-author of the original paper on the flow method, and both authors have promoted the use of this method in the UK, Irish and European cancer registries. There are no commercial interests.

References