DCRC:
Tabulation outline

Diabetes and Cancer Research Consortium
2
June 2010

www.biostat.ku.dk/~bxc/DCRC

(Compiled Tuesday 29th June, 2010 at 17:27)

Bendix Carstensen Steno Diabetes Center, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk
http://www.biostat.ku.dk/~bxc/

Edwin Gale University of Bristol, UK
Edwin.Gale@bristol.ac.uk

Jeff Johnson ACHORD, University of Alberta, Edmonton, Cananda
jeff.johnson@ualberta.ca

Helen Colhoun University of Dundee
h.colhoun@chs.dundee.ac.uk

Sara Wild University of Edinburgh
sarah.wild@ed.ac.uk

Henrik Møller King's College & Thames Cancer Registy, London
henrik.moller@kcl.ac.uk

Reijo Sund National Institute for Health and Welfare, Helsinki
reijo.sund@thl.fi

Junmei Jonasson University of Göteborg
junmei.jonasson@gu.se

Mail to all

Contents

1 Introduction

This document is meant as a guideline for a general tabulation of follow-up data for the contributing members for the Diabetes and Cancer Research Consortium. The purpose is to establish summary datasets (tables) that allows joint analysis of cancer incidence across centers, with the specific aim of evaluating the effects of exposure to certain drugs used in diabetes treatment, that be causal or assignment effects.

2 Exposures (covariates)

2.1 Drugs

The following pharmaceutical exposures are of interest:

- Metformin (A10BA)
- SU (A10BB))
- TZDs (A10BG)
- Insulin (A10A)

This means that follow-up should be classified indicator variables of whether these drugs are actually being taken in any given follow-up interval.

2.2 Timescales

The following timescales are of interest:

- Time since AD 0 (current calendar time)
- Time since birth (current age)
- Time since DM diagnosis (current disease duration)
- Time since first Metformin dispensation
- Time since first SU dispensation
- Time since first TZD dispensation
- Time since first Insulin dispensation

For the last 5 timescales, persons who are not (yet) diagnosed with diabetes should be coded 0; persons diagnosed with diabetes and/or on any of the drugs but where duration is unknown should be coded NA (missing, Not Available).

The latter will enable analyses investigating duration effects excluding persons without duration information, as well as analyses including all persons ignoring the duration variable.

Also note that these duration variables are defined as time since first dispensation, that is they keep increasing, even if a given drug is stopped. Moreover, we will not have any handle on time off a drug, but we will know by the interaction between (time since first dispensation > 0) and the indicator of the drug being taken whether persons off the drug have a higher or lower risk that those who have never been on it.

We require that age and calendar time at follow-up as well as date of birth be tabulated in 1-year intervals.

If we require that the 5 duration timescales be tabulated in 6-month intervals, we may very well produce a very large dataset indeed, most likely substantially exceeding 10,000,000 records (see appendix).

2.3 Timescale practicalities

Allocation of events and follow-up time to intervals on the many timescales requires that the dataset be split on these.

Technically this can be done in Stata using the command `stsplit`. In SAS by using the macro `%Lexis`, available as `http://staff.pubhealth.ku.dk/~bxc/Lexis/Lexis.sas`, which contains guidance to the use in the file itself. In R is a machinery called `splitLexis`, but since R keep all data in memory this may be prohibitive by the sheer data size.

2.4 Covariates

The indicators and the timescales define 11 variables, so including sex and date of birth we will have 13 explanatory variables.

3 Outcomes

We propose to include follow-up only until first primary cancer, and censor persons at this event. This will simplify analyses, since the follow-up time will be the same for all cancer incidence outcomes considered.

In the cancer epidemiology literature there are varying practices on this point; some prefer to follow persons till the occurrence of the primary cancer of interest, disregarding earlier occurrence of other primary cancers. By the same token, only persons with the particular cancer diagnosed prior to start of follow up should be excluded.

The following outcomes should be tabulated for all:

- Follow-up time (person-years) before death.
- Death.
- Follow-up time (person-years) before first primary cancer of any kind.
- Any primary cancer.
- Any primary cancer except non-melanoma skin cancer.
- (all the other cancers — specify; ICD10 codes.)

These outcomes thus define $5 + \{\text{number of cancer sites}\}$ variables in the dataset.

4 Non-DM follow-up

Not all studies have access to a full database of the entire population, but only to demographic data from the statistical bureau (population size and hence derived population follow-up time), and to cancers for the entire population, typically by sex, age, calendar time and date of birth.

In this case, the follow-up time and cancer cases in the DM population must be subtracted from that of the total population to give the cases and follow-up time in the non-DM population.

5 Specifics

A number of specific features of the different studies must be taken into account:

Scotland The diabetes classification is incomplete prior to 2003(????) and hence the follow-up among those coded as non-diabetics contains a fraction of DM patients. Thus analyses that compares rates between non-DM and DM are not valid for the period prior to this date. But comparisons *internally* in the group of DM patients are.

Canada (BC) The data is not a complete enumeration of the follow-up in the population, but only among diabetes patients and a sample of non-DM persons matched to the DM-persons at date of diagnosis. Hence, the DM patients can only be followed from the date of DM, and the matched non-DM persons only from the matching date.

THIN/GPRD The cancer diagnoses are based on extracts from GP databases, and are therefore less reliable, and particularly some persons with a previous cancer may be included.

6 Appendix: The tabulation squeeze

In pharmacoepidemological studies of diabetes we are interested in many timescales beyond the fundamental three, current age, current calender time and disease duration. Each drug of interest will typically require two timescales, namely time since initiation of the drug and time since the cessation of it.

The point of this note is not to discuss the finer points of the definition of these variables, here we shall just assume that algorithms are available to define all relevant time scale variables at any desired point of follow up for all persons in the study.

The purpose of this note is to discuss practical data processing problems with many timescales.

6.1 Data requirement for follow-up data

If multiple timescales are to be accommodated, it is required that the follow up time is subdivided by each of these. Splitting of follow-up time by age and calendar time as well as by a number of time scales will result in a very large number of units from each patient, and potentially also a very large number of cells in the required cross-classification of timescales.

6.2 The tabulation squeeze

If four drugs and diabetes are to be classified by duration in say 6-month intervals, then we will with 15 years of follow up have 30 intervals on each time scale, that is potentially $30^5 = 24.3$mio. intervals, which additionally must be classified by age, calendar time and diabetes duration. A substantial fraction of these potential combinations will of course be empty, but with the additional tabulation by age and period, we can easily run into hundreds of millions of combinations, which currently is not feasible as analysis unit.

6.3 A practical solution

It should be noted that it is only the diabetes patients' follow-up that need subdivision by drug-exposures. And so far we have only a few hundred thousand patients in each data base. So if the follow-up of diabetes patients is only split by time since diagnosis (duration of diabetes), in say 6-month intervals, we will have up to 30 intervals per person. In the Danish diabetes and cancer study there is about 1,000,000 person-years, so this tabulation would result in some 2,000,000 intervals, which is in the range of analytical possibilities.

The advantage of this is that we can define all of the required timescales for any of these intervals, so this approach is robust to inclusion of any number of time scales, as the number of units will stay the same regardless of further time-scales being added.

The follow-up of the non-diabetic population is still classified and tabulated by current age, calendar time and date of birth. In order to make the follow-up of the diabetes patients comparable to this, the age and calendar time assigned to each interval should be the age and calendar time 3 months (*i.e.* half the tabulation length) after the left endpoint of the interval (which for most intervals will the midpoint).

The duration variables however, should correspond to the left endpoint of the intervals.

This way we will be able to produce a dataset which in the case of Danish data will have some 2 million records, and which can accommodate any number of timescales for analysis. Hence, it will only be the the number of events that limits the complexity of the models, not the tabulation possibilities.

6.4 Confidentiality

The resulting dataset will be a dataset which has very large numbers of person-years for each age, period, cohort class for the non-diabetic population and very small amounts of follow-up for each combination of the many timescales for the diabetes population. Some (presumably most) contributions from the diabetes population will only contain follow-up data from one person.

It will however be totally uninformative about the the persons identity, because it will only concern a small piece of the follow-up from the person, and there will be no way to link this piece of information to the rest of the information from the same person. Hence, there will be no way to link any of the follow-up tabulated this way back to the individuals. Unless, of course, all the information contained in the record is known from some other source, in which case the confidentiality issue would be somewhere else.