ADVANCED ANALYSIS OF LINKED HEALTH DATA

Topics and Technologies

A Short-Course with

Professor David B. Preen

SHORT COURSE WORKBOOK

Version 2.7 December 2022

COURSE INFORMATION

Course Description

Advanced Analysis of Linked Health Data is an intensive course of five days in duration, designed to instruct participants in the theory and practice of analysis of large sets of linked health data at an intermediate to advanced level. Advanced principles of health care epidemiology are combined with hands-on practical exercises in the implementation of computing solutions.

The modular structure of the unit provides students with a theoretical grounding in the classroom on each topic, followed by a training session on the corresponding technological computing solutions. Students use fictitious but realistic linked data files in the hands-on exercises. One-on-one coaching and instruction is available in the computing laboratory session each afternoon on how to problem-solve complex research scenarios covered by the hands-on exercises.

Course Content Overview

MODULE 1:	***PRINCIPLES OF ADVANCED LINKED DATA ANALYSIS***
Topic Session1:	Transition from Introductory to Advanced Analytical Concepts
Tech Session 1:	Preparing a Complex File Group for Analysis
Review 1:	Review of 'Principles of Advanced Linked Data Analysis'
MODULE 2:	***CLINICAL POPULATION DYNAMICS USING LINKED DATA***
Topic Session 2:	Advanced Topics in Clinical Population Dynamics
Tech Session 2:	Health Occurrences in Partitioned At Risk Populations
Review 2:	Review of 'Clinical Population Dynamics Using Linked Data'
MODULE 3:	***COMPLEX LONGITUDINAL DESIGNS USING LINKED DATA***
Topic Session 3:	Advanced Topics in Longitudinal Research Design
Tech Session 3:	Implementing Complex Longitudinal Designs
Review 3:	Review of 'Complex Longitudinal Designs Using Linked Data'
MODULE 4:	***CASE DISTRIBUTION DESIGNS USING LINKED DATA***
Topic Session 4:	Advanced Topics in Case Distribution Studies
Tech Session 4:	Implementing Case Distributions Designs
Review 4:	Review of 'Case Distribution Designs Using Linked Data'
MODULE 5:	***RISK MODIFICATION AND ADJUSTMENT USING LINKED DATA***
Topic Session 5:	Advanced Topics in External and Internal Validity of Effect Measures
Tech Session 5:	Optimising Effect Estimation
Review 5:	Review of 'Risk Modification and Adjustment Using Linked Data'

Learning Objectives

Advanced Analysis of Linked Health Data provides health researchers with the opportunity to build on their pre-existing theoretical knowledge and skills in the analysis of linked health data by exploring a number of advanced topics. Upon completion the participant will:

- 🕮 have consolidated their grasp of foundation concepts of health and social epidemiology and data analysis;
- 🕮 possess an advanced understanding of methods for the conceptualisation and construction of valid measures and effect measures of health care utilisation and outcomes based on complex, multi-sourced linked data sets;
- 🕮 understand complex longitudinal research designs and how to implement them using multi-sourced linked data sets;
- 🕮 understand case distribution designs and how to implement them using multi-sourced linked data sets;
- 🕮 have skills in the analysis of linked mortality, institutional, pharmaceutical and primary care health data; and
- 🕮 be able to write computing syntax to prepare complex linked data files for analysis, derive exposure and outcome variables, relate numerators and denominators and produce results from advanced statistical procedures.

Course Prerequisites

The course assumes that students have completed *Introductory Analysis of Linked Health Data* or have *equivalent* knowledge. The computing component of the unit assumes a facile competence in the preparation of computing syntax for SPSS, SAS, Stata or R and familiarity with the statistical analysis of linked data files at an introductory to intermediate level.

Helpful Syntax Writing Resources

The advanced course is offered in SPSS, SAS, Stata and R with syntax code and solution files provided for each of these software packages to allow you to select the statistical package that best suits you. Computer lab demonstrators, proficient in each package, will also be on-hand to assist you with syntax writing and trouble-shooting in each afternoon computer exercise session.

As outlined above, the advanced course assumes students already possess competence in the preparation of computing syntax for SPSS, SAS, Stata and/or R. However, should you wish to brush up on your syntax writing skills, some resources are available through the weblinks below which provide information on the use, rules, structure, main commands and implementation of statistical syntax.

SPSS
https://stats.idre.ucla.edu/spss/

https://libguides.library.kent.edu/SPSS/Syntax

SAS
https://stats.idre.ucla.edu/sas/

https://libguides.library.kent.edu/SAS/syntax

Stata
https://stats.idre.ucla.edu/stata/

https://libguides.library.kent.edu/statconsulting/stata

R
https://stats.idre.ucla.edu/r/

https://libguides.library.kent.edu/statconsulting/r

Class Participation and Adult Learning Philosophy

Advanced Analysis of Linked Health Data is an introductory to intermediate-level short course that adopts adult learning principles in the pursuit of its stated learning objectives. The quality of the learning experience gained by each participant is dependent not only on the design and organisation of the course and the teaching performance of the instructor, but also on the contributions to class information, discussion and training sessions made by every member of class and the individual participant's commitment to the discovery of knowledge manifest especially through completion of the training exercises.

Expression of different points of view by participants, including views different to those of the instructor, are welcomed and encouraged, as are contributions that add value to class material by sharing of knowledge and expertise held by participants. Every person in the classroom is both a student and a teacher.

Text

There is no prescribed text for the course although detailed course manuals and notes are provided. These are the subject of copyright and should not be reproduced without permission from the course instructor. Bibliographies are also provided for those who wish to read further.

About the Course Instructor

Professor David Preen is the Chair in Public Health at the School of Population and Global Health, The University of Western Australia (UWA). He was also the Director of the UWA Centre for Health Services Research (CHSR) from 2006-2016. The CHSR was established in 1995 to work with linked administrative data provided through the WA Data Linkage System and has been involved with developing novel techniques for applying epidemiological method to the analysis of complex linked datasets for over two decades. Professor Preen also holds a honorary professorial position at Swansea University (UK).

Professor Preen's current research activities include conducting public health and health services research in areas such as: i) cancer service delivery, ii) pharmaco-epidemiology, iii) chronic disease management, iv) social determinants of health, and v) methodological advances using data linkage.

Professor Preen has been awarded over 75 competitive research grants as chief investigator (totalling >$57 million) from State, national and international funding agencies such as the NHMRC, Australian Research Council (ARC), Australian Research Data Commons (ARDC), the National Prescribing Service, Cancer Australia, the National Breast Cancer Foundation, Cancer Council WA, the Commonwealth Department of Health and Ageing, the Australian Institute of Health and Welfare, the WA Department of Health and the UK Medical Research Council.

He has over 300 publications in international and national peer-reviewed journals, most of which have used whole-population linked administrative health data, and has authored 18 commissioned large-scale published research reports for agencies such as the MundiPhama International, Heart Foundation, WA Department of Health, the Australian Council on Smoking, Health and the Public Health Advocacy Institute, and Medibank Health Solutions. In addition, he has been awarded prizes at national and international level for his work, including the Peter Reizenstein Prize from the International Journal of Quality in Health Care, Editor's Choice Award from the Medical Journal of Australia and research prizes from multiple national and international conferences.

Professor Preen has been involved with the teaching of epidemiology and linked data analysis (at both the introductory and advanced level) for the last 10 years, being responsible for coordination of the UWA linked data training curriculum since 2012. He has presented linked data analysis courses in the UK, Austria, Singapore and New Zealand as well as multiple Australian States.

Contact Details

Professor David Preen
School of Population and Global Health, The University of Western Australia,
35 Stirling Hwy, Crawley, Western Australia 6009.
Ph: +61 8 6488 1251; Email: david.preen@uwa.edu.au

The instructors' administrative secretary is Ms Jennifer Di Risio
Ph: +61 8 6488 1318; Email: jenifer.dirisio@uwa.edu.au

About the Course Founder

Emeritus Professor D'Arcy Holman
MBBS LLB(*Hons*) MPH(*Harv.*) PhD GAICD FACE FPHAA FAFPHM FAIM AM
School of Population Health
The University of Western Australia

Course history and evolution

Based on his experience since 1995 in leading the development and exposition of the Western Australian Data Linkage System, Professor D'Arcy Holman, in 2001, wrote and delivered the world's first training course devoted to the theory and practical skills needed to analyse complex linked health data sets. The course was available to health researchers, clinicians and managers and served not only to convey the necessary theory and skills, but also an appreciation of the utility of linked data to answer questions related to national health priorities.

Demand for this unique course grew quickly and by 2007, the offering was enhanced with the first-ever advanced-level course, thus creating a training curriculum in two articulating parts: *Introductory Analysis and Linked Health Data: Principles and Hands-On Applications*; and *Advanced Analysis of Linked Health Data: Topics and Technologies*. Both now exist as self-contained, intensive, five-day training courses delivered face-to-face using classroom facilities.

Professor David Preen, whilst working with D'Arcy Holman on a number of research projects, became involved in developing and delivering the courses from 2007 onwards, and with D'Arcy's encouragement took over the leadership of the curriculum commencing from 2012. Professor Holman remains an advisor on the curriculum and often makes guest appearances when possible.

1. TRANSITION FROM INTRODUCTORY TO ADVANCED ANALYTICAL CONCEPTS

1.1 Review of Introductory Analytical Concepts

The prerequisite short-course, *Introductory Analysis of Linked Health Data*, aimed to familiarise participants with data linkage methods and systems; to teach the foundation principles of epidemiology; and to teach the skills needed to implement epidemiological concepts in computing syntax capable of analysing linked health data sets.

Foundation Principles of Epidemiology

The foundation principles of epidemiology include the following topics:

- Conceptualisation and calculation of basic epidemiologic measures.
- Relationships between basic measures.
- Rationale for and method of actuarial survival analysis.
- Conceptualisation and calculation of effect measures.
- Estimating effect measures from 2x2 tables.
- Sources of random and systematic error.
- Hypothesis testing and effect measurement.
- Theory of the cohort study.
- Theory of the case-control study, including nested case-control study.
- Confounding including confounding by indication.
- Effect modification.
- Methods of risk adjustment used in study design.
- Methods of risk adjustment used in study analysis.
- Multivariate modelling.

Within these topics you should be familiar with correct definitions of the following terms:

- Health events and states.
- Cohort and dynamic populations.
- Cross-section.
- Ratio, proportion, rate, odds and risk.
- Prevalence, cumulative incidence, survival and incidence rate.
- Random error and systematic error.
- Precision and internal validity.
- p-value and confidence interval.
- Time zero.
- Censoring date.
- Post-censoring event.
- Restriction.
- Stratification.

Be aware of the central place in epidemiological thinking of the ***population at risk*** of a particular health event or state. Sometimes populations at risk are defined *ab initio*, being conceptualised accurately prior to an exact specification of the health events or states to be enumerated; eg, the cohort of WA residents first admitted to a WA hospital with mention of diabetes mellitus in 1980-2005. This is what the theoretical epidemiologist, Olli Miettinen, calls a ***primary study base***. A different approach occurs when the initial focus is on a health event or state that is the object of research; eg, first-time renal dialysis in WA for diabetic renal failure in 1980-2005. A specification of the population at risk then follows from a careful conception of the people who have at least the potential to incur this outcome, taking into account the nature of the research question to be posed. This is what Miettinen calls a ***secondary study base***. A person cannot be a member of the population at risk if there is no possibility of them ever incurring the outcome event or state. Thus, in the latter example, a diabetic who is already on a dialysis program is not a member of the population at risk because it is no longer possible for that patient to incur their first-time renal dialysis.

Foundation Computing Syntax Concepts

The computing syntax concepts covered in the introductory course are essential foundations for this course and include the following technologies:

- Read data files into SPSS, SAS, Stata or R.
- Merge files.
- Tag records mentioning target conditions or services.
- Create sequence variables.
- File processing order and information backflow.

Within these topics you should be familiar with correct definitions of the following terms:

- End of record loading.
- Syntax functions including date functions and lag functions.
- Syntax commands including loops and conditionals.
- Scratch variables.
- Index record.
- Upside-down file, vector file and (in Stata) explicit subscripting.

1.2 Making the Transition to Advanced Analytical Concepts

Need for More Detailed Health System Knowledge

The content and philosophy of the advanced course builds on the foundations laid down in the introductory course. One motivation behind the advanced course is to prepare participants for research projects that entail the use of more than one (and often three or more) large files of health data that contain multiple and variable numbers of records per individual. For example, such a circumstance is typical of ***cross-jurisdictional linkage projects*** using data sourced from multiple statutory and claims data collections:

Typical statutory data collections (mostly 'State' in Australia)	**Typical claims data collections** (mostly 'Commonwealth' in Australia)
Birth registrations Midwives' notifications Cancer notifications Mental health system encounters Hospital separations Death registrations Others; eg mental illness	Medical benefits claims Pharmaceutical benefit claims Residential aged care subsidies Others

In Australia, a common 'shape' to cross-jurisdictional linkage projects is that the Commonwealth data collections provide information on service exposures (eg, occasions of primary medical care, medical procedures, laboratory tests and pharmaceutical prescriptions) and the State data collections provide information on health outcomes (eg, perinatal outcomes, cancers, diseases and injuries causing hospitalisation or death). However, this is not invariable. The informational role of the two jurisdictions may be reversed, as in a study of the extent to which people with mental illness (ascertained using State data: the 'exposure') make use of primary medical care services (ascertained from Commonwealth data: the 'outcome'). In some instances, data on exposures or outcomes may be drawn from both jurisdictions, as with exposure to a procedure, such as vasectomy, which may be performed as either a public inpatient (State data only), an outpatient in private rooms (Commonwealth data only) or a private inpatient (ascertainable from both State and Commonwealth data). Some disease outcomes may also be ascertainable to a partial extent in both jurisdictions. For example, a disease outcome of epilepsy could be identified as a diagnosis on a hospital separation, mental health record or death registration (all State data collections) as well as identified in MBS and PBS data using item numbers for laboratory tests and medications that are pathognomic of epilepsy (from Commonwealth data).

Due to complexities of the type illustrated above, the competent analyst of cross-jurisdictional linked data must have a sound grasp of the health care processes relevant to their research topic; for example, the respective roles of different jurisdictions. Only then can the different data sets, State and Commonwealth, be used to maximum advantage to support the researcher's objectives.

While much of the necessary knowledge can only accrue with experience, this short-course aims to launch participants on the required path. In the longer term, working with clinical collaborators may be the most effective means of ensuring that the requisite insights are available into how the Australian health system, with its split funding arrangements, works in practice.

Technical Challenges of Complex File Groups

Particularly at issue in terms of technological challenge is that medical service and pharmaceutical data are characterised by very large numbers of records per individual, albeit that the clinical content of these records is sparse by comparison with hospital separation data, for example. A range of new skills are needed to work effectively with these very large data sets, especially when research questions may require a high degree of interaction between the data files. The technical challenges encountered in this working environment may include:

- The lack of a census of all cases of a target condition or service in any one file, necessitating a search across multiple files to identify all eligible individuals.
- The bringing together and characterisation of index records that may be located in different files.
- Combining information derived from multiple records in two or more files.
- Loading information derived from multiple records onto a record in another file.
- The lack of explicit information on diagnoses in the medical service and pharmaceutical data.

The advanced course is designed to familiarise participants with effective solutions to these problems.

Advanced Topics in Epidemiology

Richer data resources lend themselves to a number of newer and exciting methods of epidemiological analysis. Unlike the introductory course, which endeavoured to lay a broadly based foundation as the framework for thinking about epidemiological problems, this advanced course adopts a topics-based approach to theory rather than a comprehensive approach. Each topic will be dealt with in relative isolation and sometimes in considerable depth. The aim is to equip participants with a set of specialised tools that may prove useful in future research. Topics covered in this course include the following:

- Backcasting method for correction of the prevalent pool effect.
- Capture-recapture methods for correction of case under-ascertainment.
- Sentinel event, reference point and active disease models for characterising clinical populations and patterns of care.
- Complex longitudinal designs using domain and time restrictions.
- Nested case-control studies.
- Case-crossover, case-time-control and crossover longitudinal designs.
- Propensity score analysis.
- Instrumental variable analysis.
- Assessment of effect modification.

1.3 Working with Medical Claims Data

Medical Claims Data Coverage

Eligible claimants: Medical claims data will cover services delivered to those eligible for a benefit. For private health insurance schemes, this will be those with paid-up policies of sufficient duration to exceed any waiting period. But even universal government schemes will have eligibility criteria. In Australia, for example, Medicare benefits are available only to Australian residents and foreign visitors from countries with a reciprocal health care agreement with Australia. In some jurisdictions, war veterans or military personnel are covered by a separate scheme.

Eligible services: Services eligible for a benefit may vary considerably across schemes and within the same scheme over time. Some areas where policy settings often vary include the following:

- Consultation or procedural services from general practitioners and medical specialists according to whether provided in private rooms, in public hospital outpatient or emergency department facilities, or provided to hospital inpatients who are public or private patients.
- Diagnostic tests and examinations to manage illness as distinct from screening services and health assessments; eg, for life insurance purposes. Also cosmetic versus reconstructive surgery.
- Services from non-physicians such as dentists, optometrists, podiatrists and physiotherapists, according to whether provided in private rooms or a hospital setting; eg, inpatient surgical procedures by dentists; eye test by optometrists.

Processing of claims: For data to be created, the patient receiving the service must register a claim.

Medical Claims Data Elements

The available data items will vary across systems. A typical medical claims data set might include:

- Patient's identifier (usually encrypted).
- Patient's age, gender and postcode.
- Service item number.
- Date of service.
- Provider's identifier (usually encrypted) and business postcode.

A common issue in medical claims data is the extent of available diagnostic information.

Medical Claims Data Coding

Researchers should be aware that, typically, there are periodic changes to the way services are classified and there is also potential for claim inaccuracies. Areas prone to medical advances and policy changes may have incomplete longitudinal series. A large part of a medical claims data set may contain only standard consulting codes (eg, 'short consultation', 'long consultation').

1.4 Working with Pharmaceutical Claims Data

Pharmaceutical Claims Data Coverage

Eligible claimants: As with medical claims, pharmaceutical claims data will also include only pharmaceuticals dispensed to those eligible for a benefit. War veterans or military personnel might be covered by a separate scheme. In some systems, the claimant need not be the patient; eg, if the total out-of-pocket expense of pharmaceuticals in a calendar year exceeds a certain threshold, in some jurisdictions this may trigger safety net entitlements, whereby benefits are paid to the nominated head of the family. If the patient is not recorded separately, this may cause identity problems.

Eligible pharmaceuticals: Usually only prescription pharmaceuticals and not over-the-counter products are included. Eligible pharmaceuticals may vary considerably across schemes and within the same scheme over time. Some areas where policy settings often vary include the following:

- Not all prescription pharmaceuticals may be scheduled for a benefit and, even then, the treatment circumstances may be restricted. For example, some chemotherapy agents may be scheduled only for treatment of certain cancers, despite emerging therapeutic effectiveness in other cancers. The lag between drug availability and scheduling can be prolonged.
- The prescription setting may be relevant; for example, whether pharmaceuticals dispensed in public hospitals are eligible for a benefit.
- The level of front-end deductable may affect which pharmaceuticals are eligible. Low cost drugs (eg, some antibiotics) may have a price below the front-end deductable so the patient pays the full amount and no claim is possible. In Australia, the maximum level of patient contribution is higher for those with general entitlements compared with pensioners and others with concessional entitlements. Where co-payment levels vary, pharmaceuticals may become effectively eligible in the hands of some claimants, but not others. If there is a maximum annual copayment, some pharmaceuticals in the hands of some claimants may become eligible for a benefit towards the latter part of each year.

Processing of claims: For data to be created, the patient must fill the prescription and register a claim. Not all prescriptions will be filled. In some systems, claims may be lodged directly by pharmacists; others may require action by the patient after the prescription is filled.

Pharmaceutical Claims Data Elements

The available data items will vary across systems. A typical data set might include:

- Patient's or claimant's identifier (usually encrypted).
- Patient's age, gender and postcode.
- Pharmaceutical item number.
- Date of supply and type of prescription (original, repeat, special authority, etc).
- Benefit payment category (general, concessional, safety net, etc).
- Pharmacist identifier (usually encrypted) and business postcode.

Pharmaceutical claims data rarely include diagnostic information.

Pharmaceutical Claims Data Coding

Researchers using pharmaceutical claims data should be aware that pharmaceutical item numbers are regularly added and deleted from benefit schedules and that published schedules may change more than once per year.

Advanced Theory Topic A:

1.4 Backcasting Method

Linked disease-related event data can be used to estimate the incidence of serious disease. However, incidence rates calculated from first-time incident events tend to be overestimated as a result of erroneous inclusion of prevalent cases that had their true first-time event prior to the available period of observation. A simple solution to this problem of the ***prevalent pool effect*** is to enforce a ***clearance period***, whereby data from the initial years of observation are discarded based on a visual inspection of where the prevalent pool effect appears to end. Although easy to implement, use of a clearance period has the disadvantage of loss of a number of years of information. There is also the concern that a true downward trend in incidence rate in the early years of observation may be mistaken for a substantial prevalence pool artefact.

The Brameld-Holman ***backcasting method*** was developed as a solution to the prevalent pool, which does not necessitate discarding data from early years. It implements an actuarial retrograde survival model to derive correction factors for the levels of over-ascertainment of first-time events according to the number of years of linked data on which estimates are based. The backcasting method exists in two forms: (i) correction factors for consecutive annual periods of observation; and (ii) correction weights for individual patients. In this course we cover the latter of the two methods, because it is considered to be the more accurate, albeit still with limitations.

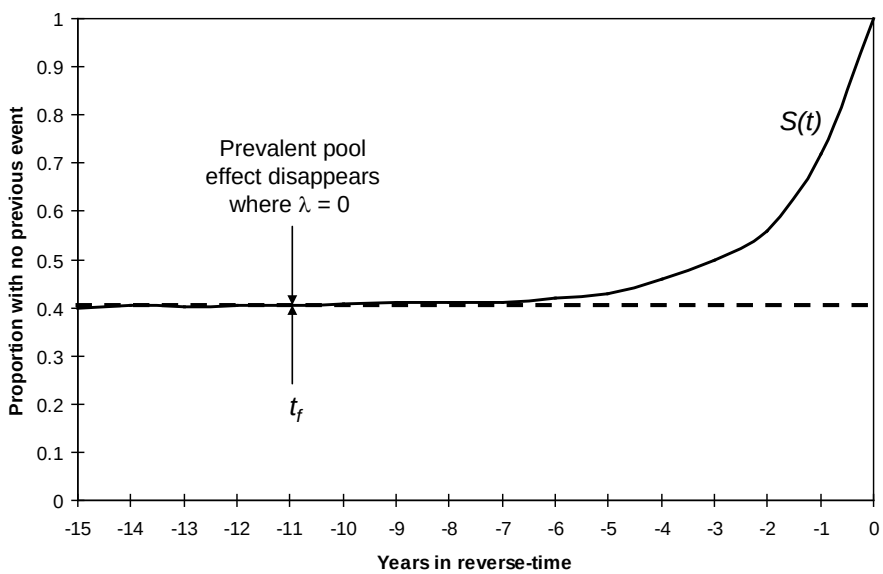
General Theory of the Backcasting Method

Taking time-zero to be the date of any relevant event observed in a patient, whether apparently a first-time event or not, retrograde follow-up time proceeds backwards in ***reverse time*** (t = 0, -1, -2, etc) and ends when the patient experiences an earlier event. For patients with no previous event during the study observation period, retrograde follow-up time is censored on the last date on which event information is available. A ***retrograde survival curve*** is produced from the data such that:

$$S(t) = \prod_{l=0}^{t+1}\left(1 - \frac{d_l}{n_l}\right)$$

where *S(t)* is the survival by reverse-time *t* from a previous event, d_l is the number of patients with previous events in the retrograde follow-up interval (*l*, *l*-1), and n_l is the number of patients at risk at the start of each interval. Thus d_l/n_l is the cumulative incidence of a previous event in (*l*, *l*+1). The underlying hazard rate of a previous event at time point *t*, going backwards in time, is given by:

$$\lambda(t) = \underset{\delta t \to 0}{Lim} \frac{d(t)}{n(t)\delta t}$$



where $d(t)$ is the number of events in the interval $(t, t - \delta t)$ and $n(t)\delta t$ is the person-time at risk in that interval. For a chronic disease for which the retrograde hazard is monotonic and decreasing with reverse-time and where the survival function is decreasing with reverse-time, it is assumed that a point exists where all previous events have been accounted for and there is no further hazard. We denote this point in reverse-time where $\lambda = 0$, and the patient is risk free, as t_f.

To estimate t_f, we differentiate the survival curve or fit a regression curve to the estimates of $\lambda(t)$ at t = 0, -1, -2, … , and solve $\lambda(t) = 0$ for t or the earliest reasonable approximation to zero if the fitted curve is asymptotic in form. Thus $S(t_f)$ is the proportion of patients who have truly never experienced a previous event. A correction factor for over-ascertainment of incident events due to the prevalent pool effect is now given by the conditional probability of an event being the true first event, conditional on no previous event having been observed since the commencement of the study period (see below for formula).

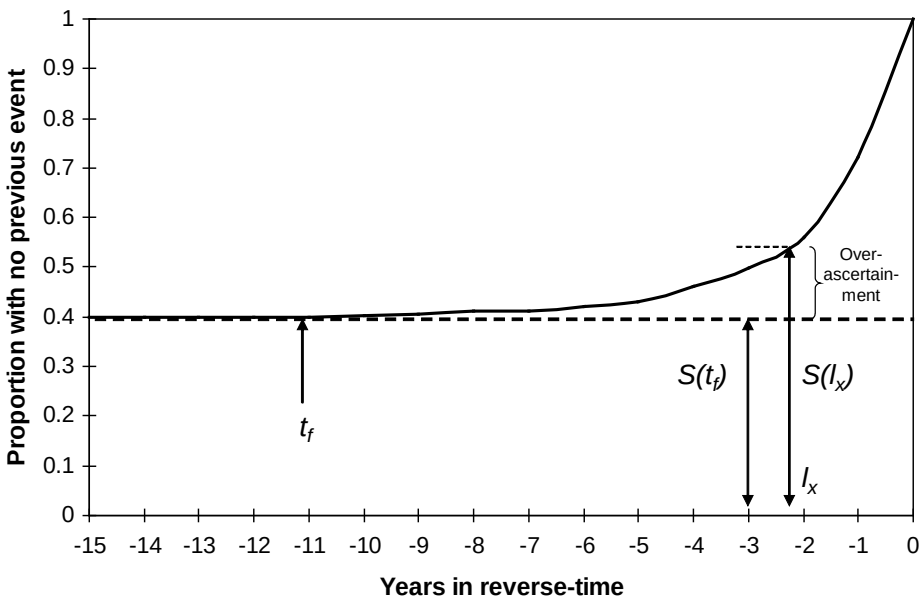
Backcasting Correction Weights for Individual Patients

The correction weight, C_x, is the probability that the individual apparent first-time case x, is truly incident, as distinct from a member of the prevalent pool of previously incident cases. Taking a file initially of all cases (not merely those that are apparently first-time events), let a denote the date of an event, b the date of the most recent previous event, and c the first date of the study observation period. The retrograde follow-up time is $l_x = b - a$ for cases with previous events or $l_x = c - a$ for censored cases.

Retrograde hazard and survival curve are generated and t_f and $S(t_f)$ are thereby estimated by fitting a regression curve as explained above. The C_x for apparent first-time cases are estimated as:

$$C_x = \frac{S(t_f)}{S(l_x)} \text{ for } l_x > t_f \quad \text{or} \quad C_x = 1 \quad \text{for} \quad l_x \le t_f$$

Note that l_x and t_f are both ≤ 0, thus $l_x > t_f$ is the same as $[l_x] < [t_f]$.



Corrected estimates of incident first-time events in the study are then obtained by weighing each apparent first-time event in the file by its appropriate C_x given the length of available retrograde clearance time, l_x, back to the start of the study observation period.

Limitation of the Backcasting Method

The backcasting method outlined above is not without limitations. For a start its validity depends on the following requirements and assumptions:

- It requires retrospective decay in the likelihood of an earlier event.
- It assume no age-related or calendar related restrictions.
- It requires sufficient data to estimate t_f.
- It assumes that the propensity for a repeat event, conditional on the first, does not change greatly over time.

It thus works best for chronic conditions and starts to fall down if incidence rates are subject to extreme variation over time; eg, in the case of surveillance of acute epidemic diseases that visit the population in distinctly separate 'waves' of incidence. A further complication may occur where the outcome event was impossible prior to a defined calendar date (the second point above), such as where a new treatment was previously unavailable. However, under the latter circumstances there is often little or no need for prevalent pool correction.

Advanced Theory Topic B:

1.5 Capture-Recapture Methods

Capture-recapture methods were originally used in estimating animal populations, where they were based on marking and releasing a batch of captured animals, to see how many were recaptured in the next batch of animals caught.

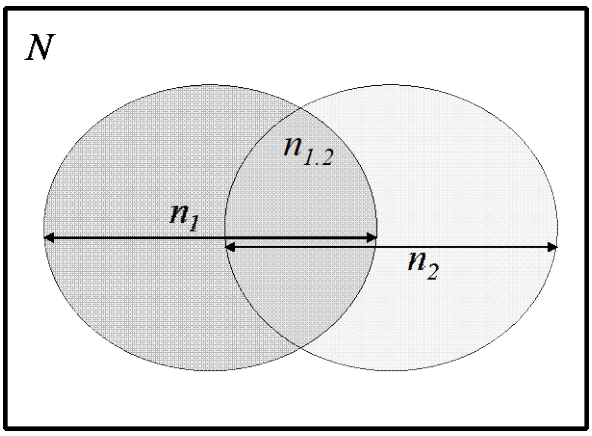
Case Finding Based on Two Sources - the Chapman Estimator

If two independent case-finding procedures identify n_1 and n_2 cases respectively, with $n_{1.2}$ cases being common to both sources (see diagram below), the Chapman estimator of the size of the total underlying population, N, is defined by (note that the reason for the +1's in the formula is the indivisibility of small numbers):

$$N = \frac{(n_1+1)(n_2+1)}{(n_{1.2}+1)} - 1$$

and Seber's approximate unbiased estimate of the variance of N is given by:

$$Var(N) = \frac{(n_1+1)(n_2+1)(n_1-n_{1.2})(n_2-n_{1.2})}{(n_{1.2}+1)^2(n_{1.2}+2)}$$



Case Finding Based on Three or More Sources – Log-Linear Modelling

The two-sample capture-recapture method is based on assumptions that are seldom met in practice, because case finding procedures are rarely independent. If cases ascertained in one data collection are especially likely to be identified in the other data collection, the calculated total using Chapman's estimator will be an underestimate. The opposite can also occur if cases ascertained from one source are likely to be missed by the other; then the estimated total will be an overestimate.

This bias can be largely controlled by log-linear modelling where there are three or more case-finding sources (also known as the ***multiple recapture census***). However, even this method will underestimate the total where there is a certain type of case who is 'uncatchable'; eg, highly disadvantaged populations who make no contact whatsoever with the health system, thus generating no administrative records.

To apply log-linear modelling the data are arranged in multi-level contingency tables. For three sources of case ascertainment, the scheme would appear as follows where values of 'no' and 'yes' denote if a case is identified in each respective data source.

Source A	Source B	Source C	Number of cases
No	No	No	F_{000}
		Yes	F_{001}
	Yes	No	F_{010}
		Yes	F_{011}
Yes	No	No	F_{100}
		Yes	F_{101}
	Yes	No	F_{110}
		Yes	F_{111}

If x is the number of sources, the contingency table will contain 2^x cells, including one missing cell, corresponding to the absence in all sources. Thus the table has $(2^x - 1)$ observations and any non-saturated log-linear model may contain at most $(2^x - 2)$ constant, main effect terms and interaction effect terms drawn from the saturated model, which for three data collection sources will be selected from:

$$ln(F_{ijk}) = \alpha + \beta_A + \beta_B + \beta_C + \beta_{AB} + \beta_{AC} + \beta_{BC} + \beta_{ABC}$$

where $ln(F_{ijk})$ is the natural log of the expected cell frequency of the cases in cell ijk in the contingency table and the β's are terms representing the main and interaction effects. Thus for a four-way multi-level contingency table (ie, four sources of case ascertainment) there will be $2^4 = 16$ cells, 15 sets of observations and a model consisting of a constant, four main effect terms and nine of the ten possible interaction terms. The log-linear model is specified to have a ***cell structure*** that recognises that the zero observations in the missing cell is not free to vary empirically (a ***structural zero***). The best fitting model is determined using goodness of fit statistics. The frequency for the missing cell is then estimated as the antilog of a linear combination of applicable parameter coefficients from the chosen model.

References

Borgmeier I, Holman CDJ. **Does vasectomy reversal protect against prostate cancer?** *Ann Epidemiol* 2004; 14: 748-749.

Brameld KJ, Holman CDJ, Lawrence DM, *et al*. **Improved methods for estimating incidence from linked hospital morbidity data.** *Int J Epidemiol* 2003; 32: 617-624.

Brameld KJ, Holman CDJ, Thomas MAB, *et al*. **Use of a state data bank to measure incidence and prevalence of a chronic disease: end-stage renal failure.** *Am J Kidney Dis* 1999; 34: 1033-1039.

Bruno G, LaPorte RE, Merletti F, *et al*. **National diabetes programs. Application of capture-recapture to count diabetes?** *Diabetes Care* 1994; 17: 548-556.

Carrao G, Bagnardi V, Vittadini G, *et al*. **Capture-recapture method to size alcohol related problems in a population.** *J Epidemiol Community Hth* 2000; 54 603-610.

Chapman DG. **Some properties of the hyper-geometric distribution with applications to zoological sample census.** *Berkeley, University of California Publications in Statistics* 1951; 1: 131-160.

Einarsdottir K, Preen DB, Emery J, *et al*. **Regular primary care decreases the likelihood of mortality and hospitalisation in older epilepsy patients.** *Med Care* 2010; 48: 472-476.

Einarsdottir K, Preen DB, Emery JD, *et al*. **Regular primary medical care lowers hospitalisation and mortality in seniors with chronic respiratory disease.** *J Gen Int Med* 2010; 25: 766-773.

Einarsdottir K, Preen DB, Sanfilippo FM, *et al*. **Mortality in Western Australian seniors with chronic respiratory diseases: a cohort study.** *BMC Public Health* 2010; 10: 385, doi:10.1186/1471-2458-10-385.

Goldacre M, Shiwach R, Yeates D. **Estimating incidence and prevalence of treated psychiatric disorders from routine statistics: the example of schizophrenia in Oxforshire.** *J Epidemiol Community Health* 1994; 48: 318-322.

Holman CDJ, Preen DB, Baynham NJ, *et al*. **A multipurpose Australian comorbidity scoring system performed better than the Charlson index.** *J Clin Epidemiol* 2005; 58: 1006-1014.

Huff L, Bogdan G, Burke K, *et al*. **Using hospital discharge data for disease surveillance.** *Public Health Rep* 1996; 111: 78-81.

Jeansonne A. **Loglinear models**. Available at:
http://online.sfsu.edu/~efc/classes/biol710/loglinear/Log%20Linear%20Models.pdf

Kelman CW, Kortt MA, Becker NG, *et al*. **Deep vein thrombosis and air travel: record linkage study.** *BMJ* 2003; 327: 1072-1076.

Mclean M, Duclos P, Jacobs P, *et al*. **Incidence of Guillain-Barre syndrome in Ontario and Quebec, 1983-1989, using hospital service databases.** *Epidemiology* 1994; 5: 443-448.

Miettinen OS. **Theoretical Epidemiology. Principles of Occurrence Research in Medicine.** New York: John Wiley & Sons, 1985.

Moorin RE, Holman CDJ. **Patient-initiated switching between private and public inpatient hospitalisation in Western Australia 1980-2001: an analysis using linked data.** *ANZ J Health Policy* 2005; 2: 12.

Qun Mai, Holman CDJ, Sanfilippo FM, *et al*. **Do Users of mental health services lack access to general practitioner services?** *Med J Aust* 2010; 192: 501-506.

Primatesta P, Goldacre MJ. **Crohn's disease and ulcerative colitis in England and the Oxford record linkage study area: a profile of hospitalized morbidity.** *Int J Epidemiol* 1995; 24: 922-928.

Seber GAF. **The effect of trap response on tag recapture estimates.** *Biometrics* 1970; 26: 13-22.

The Nova Scotia – Saskatchewan Cardiovascular Disease Epidemiology Group. **Trends in incidence and mortality from acute myocardial infarction in Nova Scotia and Saskatchewan 1974 to 1985.** *Can J Cardiology* 1992; 8: 253-58.

Tonilolo P, Psisani P, Vigano C, *et al*. **Estimating incidence of cancer from a hospital discharge reporting system.** *Rev Epidemiol Sante Publique* 1986; 34: 23-30.

TRAINING SESSION 1:
PREPARING A COMPLEX FILE GROUP FOR ANALYSIS

IMPORTANT NOTICE

The practice data sets provided for training purposes in this course are fictitious. All results based on analysis of these fictitious data are scientifically useless. The course organisers and instructor take no responsibility for any misuse of the spurious results arising from the analysis of these fictitious data sets.

General Notes About the Training Sessions

The aim of the training sessions is to give course participants sufficient briefing and explanation to complete the hands-on exercise set for the day. Several of the exercises are based on real studies using linked health data from Western Australia.

There are several points to note about the training sessions and practical exercises:

1. Although the practice data sets are fictitious and scientifically useless, the pattern of relationships between variables have been modelled to mimic qualitatively, but not quantitatively, the type of results observed in the studies using real data.
2. Many different techniques of analysis of linked data files are possible using different software packages. The approach adopted in this course, using ***rectangular file structures***, represents only one possible approach. It is not necessarily the best approach; it is merely an approach that can be understood by a competent SPSS, SAS, Stata or R user. In general, a highly skilled user would probably adopt a more sophisticated approach, including, for example, use of ***nested file structures*** or even the use of ***code-generating syntax*** in SAS.
3. The course uses a general philosophy of analysis developed within the UWA Centre for Health Services Research known as ***end of record (EOR) loading***. Again, it is emphasised that EOR loading is not the only possible technique, nor necessarily the optimal technique. It does have the advantage that it is easy to understand. An alternative approach preferred by some linked file analysts using SPSS is the use of the ***aggregate command*** to aggregate selected and derived information from blocks of records pertaining to each individual.
4. Following the training session, course participants should endeavour to implement the objectives set out in the practice exercise during the afternoon computer laboratory.
5. Seek assistance from Professor Preen (or another demonstrator) if you have persistent difficulties completing the exercise or if there is anything of importance in the theory class that you did not understand.
6. In the first session each morning of days 2-5, a review will be conducted of the practical exercise and its potential solutions. Others in the class are welcome to come forward with alternative or better solutions that they have devised, so that the class (including the instructor!) may gain the benefit of being exposed to different ideas and approaches.

Background to Training Session 1

For your first advanced training session, you are going to 'jump in the deep end' by implementing an analysis platform suitable for performing research on a group of four internally and externally linked data sets, covering deaths, hospital morbidity, medical benefit scheme ('MBS') claims and pharmaceutical benefit scheme ('PBS') claims in a population of patients with diabetes mellitus. Although fictitious, the data have qualitative similarities with real data used to study the patterns of health care and subsequent health outcomes in diabetic patients.

In the course, *Introductory Analysis of Linked Health Data*, participants were introduced to three different types of data files:

Type 1 data files: These should contain only one valid record per individual (eg, a file of death registrations).

Type 2 data files: These contain mostly only one record per individual, but some individuals have a limited number of multiple records (eg, a file of cancer registrations).

Type 3 data files: These contain a multiple and variable number of records per individual as the norm (eg, a file of hospital morbidity data).

A type 2 data file has the characteristic that it is usually practicable to convert it into a type 1 data file. The boundary used to limit the demands placed upon participants in the introductory course contained two conditions: (i) the practical exercises were limited to situations where there was no more than one type 3 data file; and (ii) when present, the type 3 data file provided a census of all of the individuals to be included in analysis. When these conditions are observed an analyst is said to be working with a ***simple file group***. In the present advanced exercise you will meet the situations where one or both of these limiting conditions is violated. You will therefore work with a ***complex file group***. The combination of deaths, hospital separations, MBS and PBS claims provide an ideal training resource with which to demonstrate how a data linkage analyst may cope with the additional complexities that arise when working with multiple type 3 data files, none of which provides a census of individuals to be included in the analysis.

Training Session Resources

The Course includes the following files, available on LMS, for use in this training session:

If you are using this statistical package …	*… you should use this save file.*
SPSS	Dthdata.sav HMDSdata.sav MBSdata.sav PBSdata.sav
SAS	Dthdata.sas7bdat HMDSdata.sas7bdat MBSdata.sas7bdat PBSdata.sas7bdat
Stata	Dthdata.dta HMDSdata.dta MBSdata.dta PBSdata.dta
R	Dthdata.rda HMDSdata.rda MBSdata.rda PBSdata.rda

Note that in all of your practice data sets, the unique identifier that pertains to records on the same individual is the variable known as the rootlpno. This name stands for *root linkage project number*, but please note that it is an arbitrary name that could be different in another system. A form of encrypted rootlpno is assigned by the WA Data Linkage Unit and results from the links created between records, both within data files and between data files from different sources.

Note that the following codes are taken to be pathognomic of diabetes for the purpose of the exercise. It should be noted that this is a simplified list for training purposes. There are a number of additional codes that may be considered and there are circumstances, such as certain uncommon diagnoses, where some of the PBS and other codes may not be pathognomic of diabetes.

Data Set	Coding System	Codes
MBS	*Medicare Benefits Schedule*	Glycolated Hb test and fructosamine Hb test:. 66551, 66554, 66557, 73815, 73840, 66319, 66322
PBS	*Pharmaceutical Benefits Schedule*	Oral hypoglycaemics: 1202, 1801, 2178, 2430, 2440, 2449, 2607, 2720, 2939, 2940, 8188, 8189, 8391, 8392, 8450, 8451, 8452, 8533, 8535, 8607, 8687, 8688, 8689, 8690, 8691, 8692, 8693, 8694, 8695, 8696, 8810, 8811 Insulin: 1425, 1426, 1429, 1430, 1431, 1461, 1462, 1531, 1532, 1533, 1534, 1535, 1537, 1591, 1592, 1710, 1711, 1713, 1715, 1716, 1718, 1721, 1722, 1761, 1762, 1763, 2061, 2062, 8006, 8084, 8085, 8212, 8390, 8435, 8571, 8609
HMDS	*ICD-9* and *ICD-9-CM* *ICD-10-AM (1st ed)*	Diabetes mellitas: 250.XX, V77.1 E10.XX thru E14.XX
Deaths	*ICD-9* *ICD-10-AM (1st ed)*	Diabetes mellitus: 250X, V77.1 E10.X thru E14.X

Note that '*X*' is a *wild* character, which may represent 0-9. To assist you in syntax preparation, the text of this table is contained in a word document file called AALHD.doc in LMS.

WARNING: You should not rely on these codes in any analysis separate from the course as they may contain errors or be incomplete.

In addition, simply using all the codes indicated above to identify a clinical population of people with diabetes would yield an incorrect ascertainment of the clinical population of interest. In practice, the above codes would be used to create an algorithm for identifying people with diabetes based on how these health services are accessed and provided in real-world clinical practice. Such decisions should be made in consultation with practising clinicians, pharmacists and other health professionals working in relevant areas. As such, it is important for an analyst to appreciate that ***understanding linked health data is synonymous with understanding current clinical practice***.

For example, a simple approach using the above data sources for identifying people with diabetes may be to select any person with any record indicating: i) insulin or oral hypoglycaemic drug used (PBS data), ii) any record of HbA1c testing (MBS data), iii) any diabetic hospital record (HMDS data) or iv) any diabetic death record (Death data). However, this would not accurately reflect real-world clinical practice and a more appropriate algorithm may be as shown below.

Table 2: Two-stage algorithm for identifying people with diabetes from administrative data that was developed as part of the Cross-jurisdictional Diabetes Pilot Project.

Diabetes identification algorithm	Data Set Used
Stage 1—Person identified as diabetic if:	
Prescribed insulin or oral hypoglycaemic agent for > 3 months	PBS
Regular annual HbA_1C testing performed	MBS
Regular annual cycle of care for diabetes-specific MBS items	MBS
Primary diagnosis of diabetes in hospital records	HMDS
Primary cause of death due to diabetes	Death Registry
Stage 2—Person, identified from Stage 1, excluded if:	
Prescribed metformin for polycystic ovarian syndrome	All data sets
Prescribed oral hypoglycaemics concurrently with high dose corticosteroids	All data sets
Prescribed sulphonylurea drugs for diabetes insipidus	All data sets
Death due to acute non-diabetes related hyperglycaemic event	All data sets

Analysis Platforms for Complex File Groups

In each training session of the Introductory Course, a hospital morbidity data file was used as the working file, or ***platform***, onto which information from other files such as death registrations or cancer notifications was appended in the ***end of record loading area***. A hospital morbidity data file, or other type 3 data file, that provides a census of the research subjects is the logical choice for an analysis platform when working with a simple file group. The reasons are that the analyst may be confident of having at least one record per research subject in that file and the information from all of the other data sets may be reduced to type 1 files, which are then used readily as keyed look-up tables to supply additionally needed data for each individual.

The analysis of a complex file group will usually differ in one or both of the following ways from the analysis of a simple file group:

1. If the complex file group has multiple type 3 data files, it will usually necessitate ***preliminary processing*** of at least one type 3 file prior to definitive analysis. This may include internal or external loading of new information onto the EOR loading area(s). The aim of preliminary processing is to distil information in a type 3 file down to a single record of information for inclusion of each individual in the analysis, which is loaded onto each individual's index record.

2. If the complex file group has at least one data file that contains a census of all individuals to be included in the analysis, it is possible to use that file as a ***primary platform*** in a manner similar to what occurs with a simple file group. However, if the complex file group lacks any data file containing a census, it may be necessary to create a ***secondary platform*** that provides such a census. Even in the situation where one of the source files does contain a census, it may still be preferred to work with a secondary platform for the sake of clarity in project management and, in some instances, to reduce computer processing time.

The needs for preliminary processing and the construction of a secondary platform will demand a great deal more forethought from the analyst than is the case with a simple file group. It is wise, therefore, to set out the desired end result and the steps necessary to achieve it in a documented plan, which might include conceptual diagrams and the steps to be taken in building the platform. The plan should bear a relationship to the scientific objectives and study methods of the research.

Commonly, the platform file has a corresponding conceptualisation in epidemiology, being a census of those individuals who make up the denominator or, less often, the numerator of one more epidemiologic measures.

Construction of a Secondary Platform

We call the separate linked data files making up the complex file group the ***source files***. Each source file is assumed to have potential to identify individuals who possess characteristics, such as target conditions or health services, that are the object of research investigation. It is also assumed in some instances that each source file may not be the only one to contain relevant information about a particular individual. The overall aim of constructing a secondary platform is to draw together all of the information in the source files needed for the research project into one *de novo* file. A general formulation of the steps used to construct a secondary platform is as follows:

Step 1: The records in each source file that mention the ***target condition(s) or service(s)*** are tagged. This procedure was covered in the *Introductory Course*.

Step 2: A single ***index record*** is identified for each individual with the target condition/service in each source file. For source files containing multiple records per person this will be defined as the first record to mention the target condition/service when the records are sorted in preferred sequential order. Creation of a ***condition/service sequence variable*** enables identification of index records (sequence value = 1) as well as other records that mention the target condition/service (sequence value = 2+). It is desirable that each index record includes

a ***condition/service date***, which for the index record is the ***index date*** being the earliest recorded date on which knowledge of the target condition or service is expressed or implied.

The following are appropriate index dates for different types of administrative records:

Type of administrative health record	Appropriate index date
Death registrations	Date of death
Conditions or procedures in hospital data	Date of admission
Cancer notifications	Date of pathology diagnosis
Midwives' notification and birth registrations	Date of birth
Medicare benefit claims	Date of service
Pharmaceutical benefit claims	Date of supply/dispensing

Step 3: A copy (not the original) of each source file is substantially cut down to contain only the records mentioning the target condition/service and only the following variables [**SPSS users:** note the use of the DROP and KEEP subcommands]:

(i) rootlpno;

(ii) the condition/service date, date;

(iii) basic demographic variables such as sex and age at index date or date of birth;

(iv) a sequence variable that always has a value ≥ 1, denoting the sequence number of the tagged record in its source file; and

(v) a new variable, source, that denotes the identity of the source file, which has the same value for all of the records in the same source file (eg, 1 = MBS; 2 = PBS; 3 = hospital morbidity; and 4 = death).

Note that the same variables must be included in each of the cut down source files. Thus, if age is one of the basic demographic variables, but it is missing from one source file, a new blank variable for age will need to be created in that source file. It is essential that all of the cut down source files have an identical structure with identical variable specifications.

Step 4: The cut down source files are merged vertically into one long and 'skinny' file using the correct procedure depending on your statistical software package:

Package:	SPSS	SAS	Stata	R
Merge files with the same structures so as to add cases from an external file to a working file.	ADD FILES (*data* | *merge files* | *add case* in pull-down menu mode)	data all files; set file1 file2 file3; run;	append using "filename"	append using rbind() from base R or rbind.fill() from package [plyr]

Once all the merging is complete, the file is sorted in ascending order of rootlpno, date and source, such that all of an individual's records are blocked together and the records within a block are in chronological order of condition/service date. There will also be a predictable order when two index records for the same individual from different source files bear the same date. The resultant file is the ***antepenultimate secondary platform*** and should be saved, because there are situations where this file is useful in its own right. An example is the use of the antepenultimate file in incidence estimate, including application of the Brameld-Holman backcasting method for correction of the prevalent pool effect.

Step 5: A copy (not the original) of the antepenultimate secondary analysis platform file is cut down further by selecting the index records from all sources (sequence = 1) and deleting the other records. The variable sequence may also then be deleted. The resultant file is the ***penultimate secondary platform*** and should be saved, because it also is useful in it own right. An example is where the penultimate file is used to undertake a capture-recapture analysis of completeness of case ascertainment. Another example is where this precursor file is used for

an inter-'observer' reliability analysis of the demographic data, such as sex and date of birth, recorded on the index records of different source files.

Step 6: The final step is to select the first record in each block of records for an individual and to delete all other records. This produces the ultimate secondary platform (or simply '*platform file*' or '*platform*'), which bears the following characteristics:

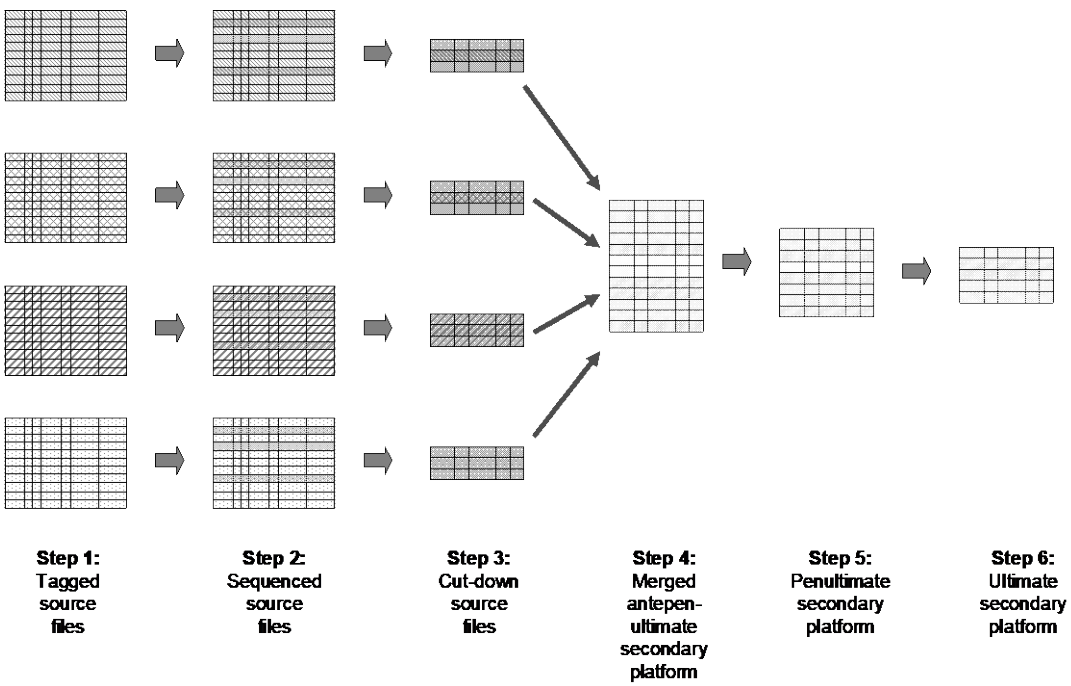
(i) One ***inaugural index record*** per individual identified by its rootlpno (the '***aug***');

(ii) a variable indicating the source of the inaugural index record;

(iii) an index date, date, that denotes the first time that the target condition/service was mentioned in any source file; and

(iv) a set of basic demographic data such as sex and age.

This becomes the ultimate working file for statistical analyses on data from two or more source files. The ***data streams*** from the source files, +/- preliminary processing, are fed into the EOR loading area of the platform file. The diagram below illustrates the steps:



It should be emphasised that the above steps are but one possible formulation of how to create a platform file. In particular, the creation of an antepenultimate platform file is optional, depending on whether the research aims of the study include analysis of incidence trends. There are many other possible approaches, and the methods taught in this course should be viewed as providing a solution that is neither unique nor necessarily optimal.

Creating a Condition/Service Sequence Variable

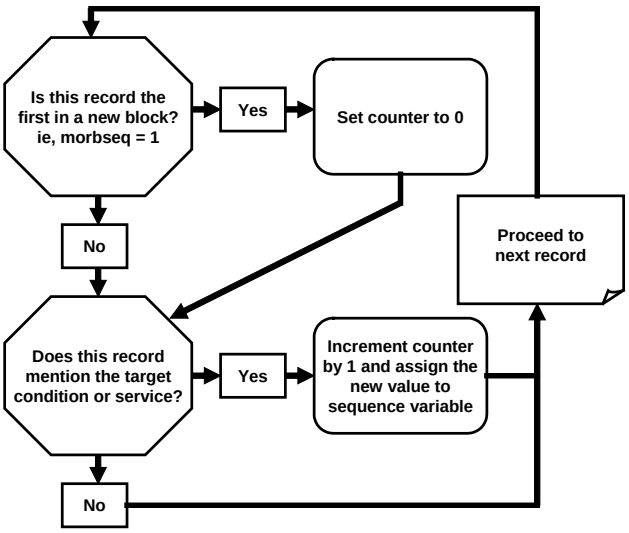
In Step 2 above you will need to create a condition/service sequence variable. In the introductory course you learnt how to create an index sequence variable, which identified the ordinal sequence of records in a file pertaining to the same individual, but commencing with the value of 1 for the index record. A condition sequence variable (suggested name condseq) or service sequence variable (suggested name servseq) is similar, but only takes on an incremented value for each record in a block mentioning the target condition/service. For other records the condition/service sequence variable is left blank.

The method of creating a condition/service sequence variable is quite straightforward.

SPSS – Users of SPSS may simply use a #counter scratch variable that is initialised to zero at the start of a new block of records and increments by 1 each time a record with a target condition/service is found. The sequence variable is simply assigned the value of the #counter for each such record (see flow chart opposite).

SAS – As for SPSS, set up a counter using the retain statement.

Stata – Users of Stata can use the egen command to create a cumulative sum of records that mention the condition or service within each block of records.

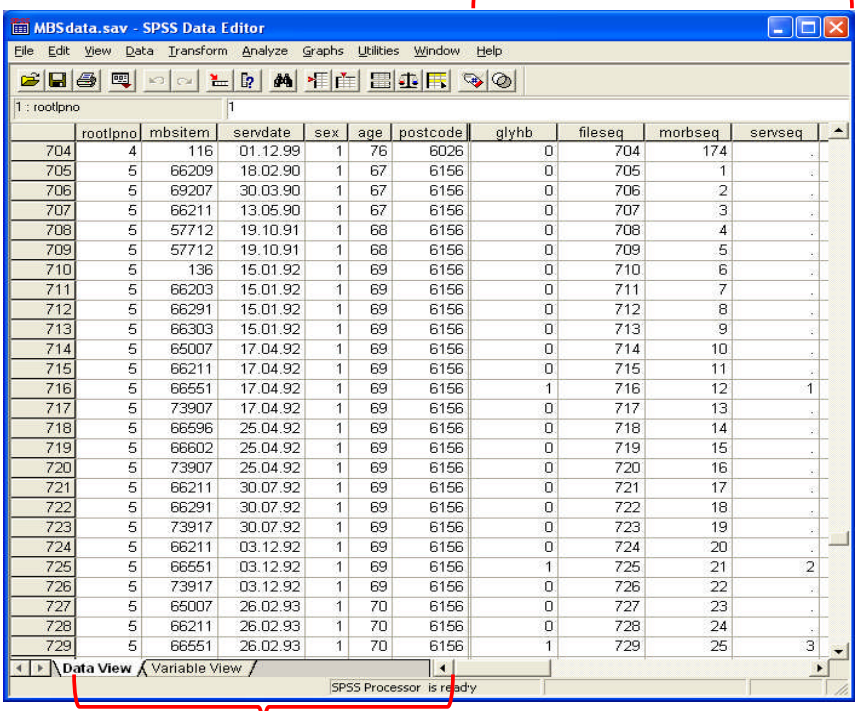


R – One way is to use the group_by() and mutate() functions from the package [dplyr], in conjunction with the function seq_along() from base R.

End-of-record (EOR) loading area.

In the exercises you will encounter the term, ***end-of-record (EOR) loading area***. This is simply the right hand side of the file where new variables appear. New variables can be added to the data set through either ***internal loading*** of indicator information for the original data or alternatively from ***external loading*** through file merges with other data sets. These concepts were presented in the introductory analysis of linked data course.

Newly created indicator variables in *end-of-record (EOR) loading area*.



	rootlpno	mbsitem	servdate	sex	age	postcode	glyhb	fileseq	morbseq	servseq
704	4	116	01.12.99	1	76	6026	0	704	174	.
705	5	66209	18.02.90	1	67	6156	0	705	1	.
706	5	69207	30.03.90	1	67	6156	0	706	2	.
707	5	66211	13.05.90	1	67	6156	0	707	3	.
708	5	57712	19.10.91	1	68	6156	0	708	4	.
709	5	57712	19.10.91	1	68	6156	0	709	5	.
710	5	136	15.01.92	1	69	6156	0	710	6	.
711	5	66203	15.01.92	1	69	6156	0	711	7	.
712	5	66291	15.01.92	1	69	6156	0	712	8	.
713	5	66303	15.01.92	1	69	6156	0	713	9	.
714	5	65007	17.04.92	1	69	6156	0	714	10	.
715	5	66211	17.04.92	1	69	6156	0	715	11	.
716	5	66551	17.04.92	1	69	6156	1	716	12	1
717	5	73907	17.04.92	1	69	6156	0	717	13	.
718	5	66596	25.04.92	1	69	6156	0	718	14	.
719	5	66602	25.04.92	1	69	6156	0	719	15	.
720	5	73907	25.04.92	1	69	6156	0	720	16	.
721	5	66211	30.07.92	1	69	6156	0	721	17	.
722	5	66291	30.07.92	1	69	6156	0	722	18	.
723	5	73917	30.07.92	1	69	6156	0	723	19	.
724	5	66211	03.12.92	1	69	6156	0	724	20	.
725	5	66551	03.12.92	1	69	6156	1	725	21	2
726	5	73917	03.12.92	1	69	6156	0	726	22	.
727	5	65007	26.02.93	1	70	6156	0	727	23	.
728	5	66211	26.02.93	1	70	6156	0	728	24	.
729	5	66551	26.02.93	1	70	6156	1	729	25	3

Original data

Overview of Exercise 1

Steps 1-4. Tag, sequence and cut down MBSdata file.

Steps 5-8. Tag, sequence and cut down PBSdata file.

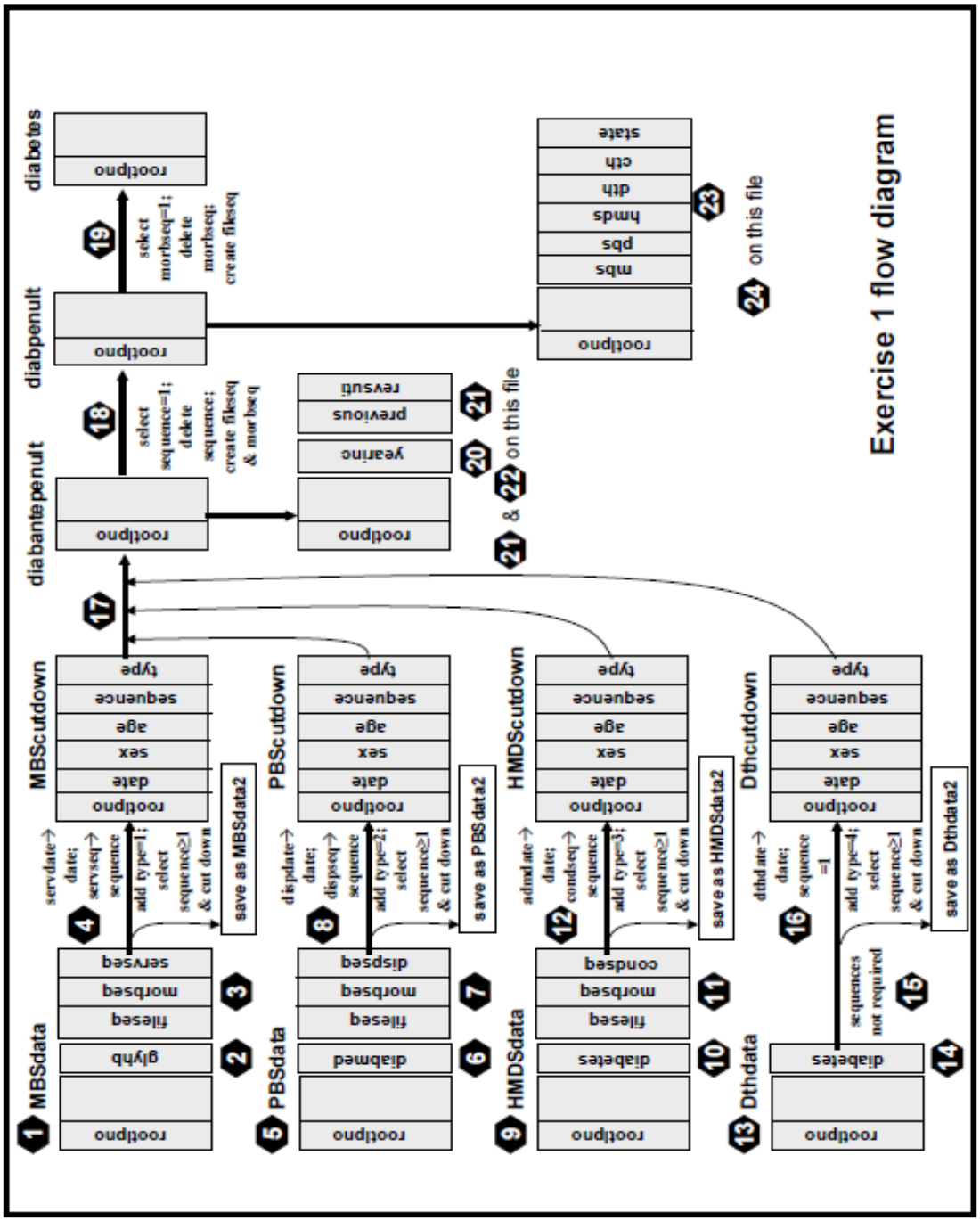
Steps 9-12. Tag, sequence and cut down HMDSdata file.

Steps 12-16. Tag, sequence and cut down Dthdata file.

Steps 17-19. Create antepenultimate → penultimate → ultimate platform file.

Steps 20-22. Backcasting correction for incidence estimation.

Step 23. Capture-recapture correction for prevalence estimation

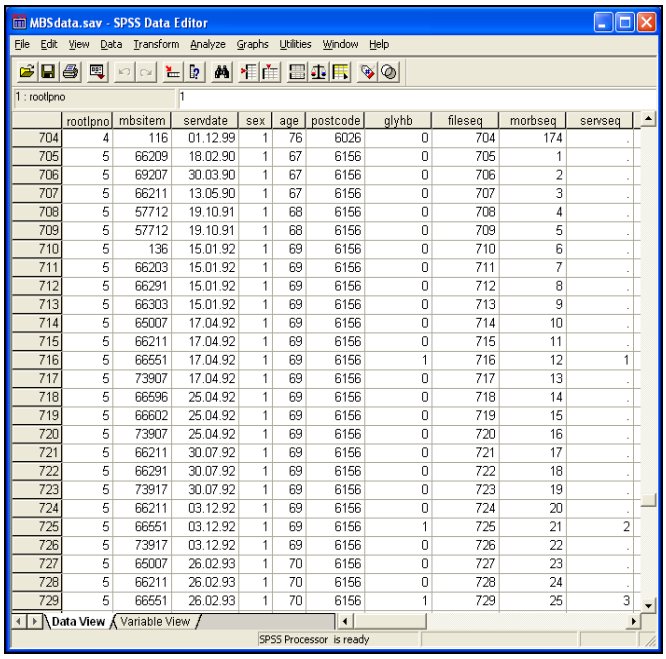


Exercise Instructions

Preparation of the MBS Data

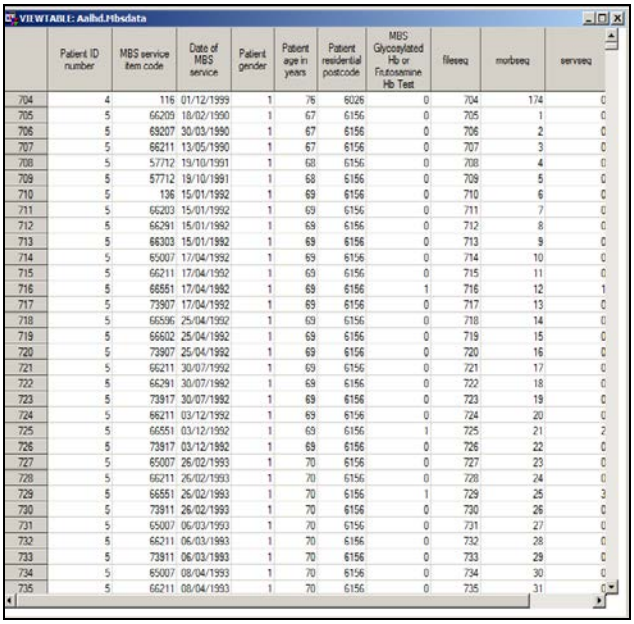
1. Open the correct MBSdata file for your statistical software package and become familiar with its structure. The 1,963,465 records are sorted by ascending servdate within ascending rootlpno.
2. Tag all records that mention a diabetes-related MBS item, being either a glycosylated Hb test or a fructosamine Hb test, by creating a new variable, glyhb, in the EOR loading area. Set glyhb to '1' for a tagged record and '0' otherwise.
3. Create the following sequence variables in the EOR loading area: (i) a fileseq variable that numbers all records in ascending order; (ii) a morbseq variable that numbers the records within each block that relate to an individual; and (iii) a servseq variable that numbers only the tagged diabetes-related records within each block. Your file should look something like the screen shown opposite.

SPSS



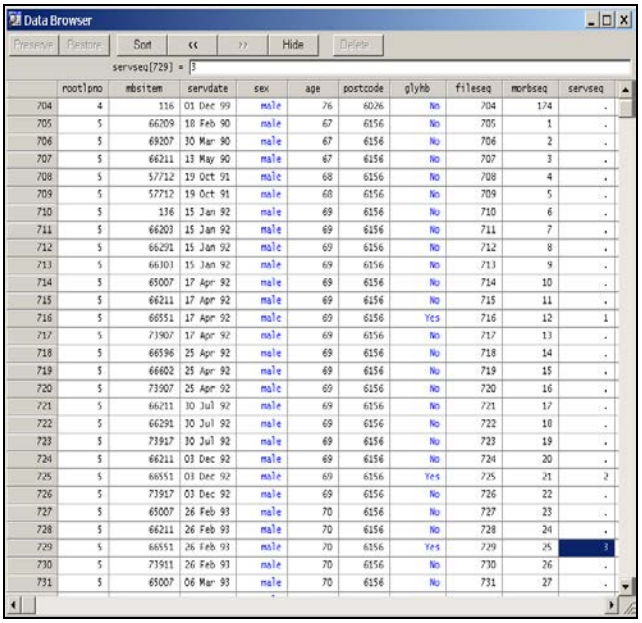
	rootlpno	mbsitem	servdate	sex	age	postcode	glyhb	fileseq	morbseq	servseq
704	4	116	01.12.99	1	76	6026	0	704	174	.
705	5	66209	18.02.90	1	67	6156	0	705	1	.
706	5	69207	30.03.90	1	67	6156	0	706	2	.
707	5	66211	13.05.90	1	67	6156	0	707	3	.
708	5	57712	19.10.91	1	68	6156	0	708	4	.
709	5	57712	19.10.91	1	68	6156	0	709	5	.
710	5	136	15.01.92	1	69	6156	0	710	6	.
711	5	66203	15.01.92	1	69	6156	0	711	7	.
712	5	66291	15.01.92	1	69	6156	0	712	8	.
713	5	66303	15.01.92	1	69	6156	0	713	9	.
714	5	65007	17.04.92	1	69	6156	0	714	10	.
715	5	66211	17.04.92	1	69	6156	0	715	11	.
716	5	66551	17.04.92	1	69	6156	1	716	12	1
717	5	73907	17.04.92	1	69	6156	0	717	13	.
718	5	66596	25.04.92	1	69	6156	0	718	14	.
719	5	66602	25.04.92	1	69	6156	0	719	15	.
720	5	73907	25.04.92	1	69	6156	0	720	16	.
721	5	66211	30.07.92	1	69	6156	0	721	17	.
722	5	66291	30.07.92	1	69	6156	0	722	18	.
723	5	73917	30.07.92	1	69	6156	0	723	19	.
724	5	66211	03.12.92	1	69	6156	0	724	20	.
725	5	66551	03.12.92	1	69	6156	1	725	21	2
726	5	73917	03.12.92	1	69	6156	0	726	22	.
727	5	65007	26.02.93	1	70	6156	0	727	23	.
728	5	66211	26.02.93	1	70	6156	0	728	24	.
729	5	66551	26.02.93	1	70	6156	1	729	25	3

SAS



	Patient ID number	MBS service item code	Date of MBS service	Patient gender	Patient age in years	Patient residential postcode	MBS Glycosylated Hb or Fructosamine Hb Test	fileseq	morbseq	servseq
704	4	116	01/12/1999	1	76	6026	0	704	174	0
705	5	66209	18/02/1990	1	67	6156	0	705	1	0
706	5	69207	30/03/1990	1	67	6156	0	706	2	0
707	5	66211	13/05/1990	1	67	6156	0	707	3	0
708	5	57712	19/10/1991	1	68	6156	0	708	4	0
709	5	57712	19/10/1991	1	68	6156	0	709	5	0
710	5	136	15/01/1992	1	69	6156	0	710	6	0
711	5	66203	15/01/1992	1	69	6156	0	711	7	0
712	5	66291	15/01/1992	1	69	6156	0	712	8	0
713	5	66303	15/01/1992	1	69	6156	0	713	9	0
714	5	65007	17/04/1992	1	69	6156	0	714	10	0
715	5	66211	17/04/1992	1	69	6156	0	715	11	0
716	5	66551	17/04/1992	1	69	6156	1	716	12	1
717	5	73907	17/04/1992	1	69	6156	0	717	13	0
718	5	66596	25/04/1992	1	69	6156	0	718	14	0
719	5	66602	25/04/1992	1	69	6156	0	719	15	0
720	5	73907	25/04/1992	1	69	6156	0	720	16	0
721	5	66211	30/07/1992	1	69	6156	0	721	17	0
722	5	66291	30/07/1992	1	69	6156	0	722	18	0
723	5	73917	30/07/1992	1	69	6156	0	723	19	0
724	5	66211	03/12/1992	1	69	6156	0	724	20	0
725	5	66551	03/12/1992	1	69	6156	1	725	21	2
726	5	73917	03/12/1992	1	69	6156	0	726	22	0
727	5	65007	26/02/1993	1	70	6156	0	727	23	0
728	5	66211	26/02/1993	1	70	6156	0	728	24	0
729	5	66551	26/02/1993	1	70	6156	1	729	25	3
730	5	73911	26/02/1993	1	70	6156	0	730	26	0
731	5	65007	06/03/1993	1	70	6156	0	731	27	0
732	5	66211	06/03/1993	1	70	6156	0	732	28	0
733	5	73911	06/03/1993	1	70	6156	0	733	29	0
734	5	65007	08/04/1993	1	70	6156	0	734	30	0
735	5	66211	08/04/1993	1	70	6156	0	735	31	0

Stata



	rootlpno	mbsitem	servdate	sex	age	postcode	glyhb	fileseq	morbseq	servseq
704	4	116	01 Dec 99	male	76	6026	No	704	174	.
705	5	66209	18 Feb 90	male	67	6156	No	705	1	.
706	5	69207	30 Mar 90	male	67	6156	No	706	2	.
707	5	66211	13 May 90	male	67	6156	No	707	3	.
708	5	57712	19 Oct 91	male	68	6156	No	708	4	.
709	5	57712	19 Oct 91	male	68	6156	No	709	5	.
710	5	136	15 Jan 92	male	69	6156	No	710	6	.
711	5	66203	15 Jan 92	male	69	6156	No	711	7	.
712	5	66291	15 Jan 92	male	69	6156	No	712	8	.
713	5	66303	15 Jan 92	male	69	6156	No	713	9	.
714	5	65007	17 Apr 92	male	69	6156	No	714	10	.
715	5	66211	17 Apr 92	male	69	6156	No	715	11	.
716	5	66551	17 Apr 92	male	69	6156	Yes	716	12	1
717	5	73907	17 Apr 92	male	69	6156	No	717	13	.
718	5	66596	25 Apr 92	male	69	6156	No	718	14	.
719	5	66602	25 Apr 92	male	69	6156	No	719	15	.
720	5	73907	25 Apr 92	male	69	6156	No	720	16	.
721	5	66211	30 Jul 92	male	69	6156	No	721	17	.
722	5	66291	30 Jul 92	male	69	6156	No	722	18	.
723	5	73917	30 Jul 92	male	69	6156	No	723	19	.
724	5	66211	03 Dec 92	male	69	6156	No	724	20	.
725	5	66551	03 Dec 92	male	69	6156	Yes	725	21	2
726	5	73917	03 Dec 92	male	69	6156	No	726	22	.
727	5	65007	26 Feb 93	male	70	6156	No	727	23	.
728	5	66211	26 Feb 93	male	70	6156	No	728	24	.
729	5	66551	26 Feb 93	male	70	6156	Yes	729	25	3
730	5	73911	26 Feb 93	male	70	6156	No	730	26	.
731	5	65007	06 Mar 93	male	70	6156	No	731	27	.

4. Save your MBS source file with the changes made above as MBSdata2. Now rename the variable servdate to just date, rename servseq to sequence and add a type variable with its value is set to '1', meaning 'MBS', for every record in the file. Cut the file down vertically so that it contains only the tagged records (sequence ≥ 1; **Stata users:** Also require 'not missing' because 'missing' has a maximum value), and horizontally so it only contains rootlpno, date, sex, age, sequence and type. Check that you are on track by confirming that your cut down file contains 29,571 records, including 7,239 records where sequence = 1. If all is well, save the cut down MBS source file with the filename MBScutdown.

Preparation of the PBS Data

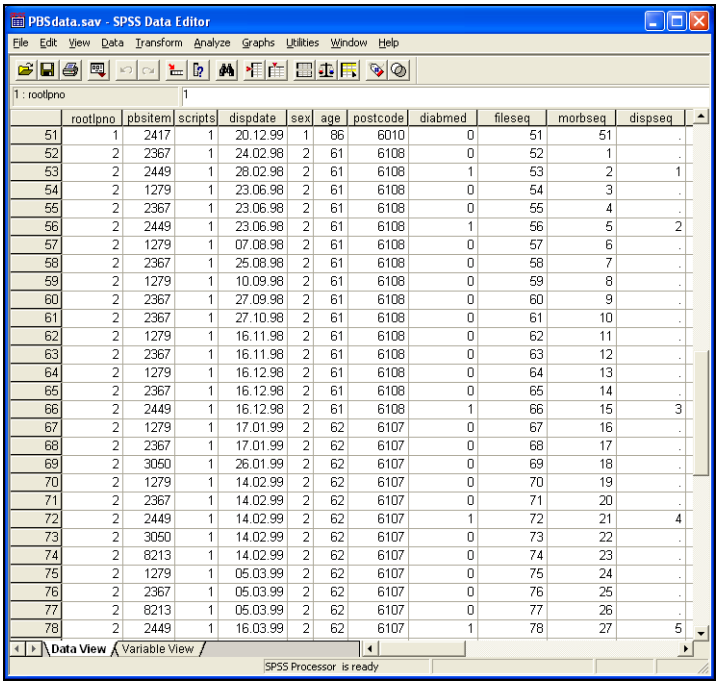
5. Open the correct PBSdata file for your statistical software package and become familiar with its structure. The 1,546,335 records are sorted by ascending dispdate within ascending rootlpno.

6. Tag all records that mention a diabetes-related PBS item, using a new variable, diabmed. Relevant scripts will either be for an oral hypoglycaemic agent (diabmed '1') or an insulin compound (diabmed '2'). Otherwise set diabmed to '0'.

7. Create the following sequence variables in the EOR loading area: (i) a fileseq variable that numbers all records in ascending order; (ii) a morbseq variable that numbers the records within each block that relate to an individual; and (iii) a dispseq variable that numbers only the tagged diabetes-related records within each block. Your file should look something like the screen shown below on the right.

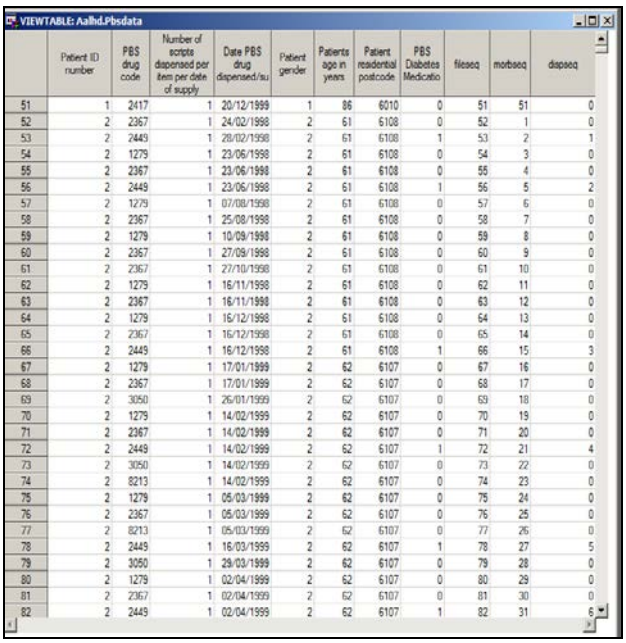
8. Save your PBS source file with the changes made above as PBSdata2. Now rename the variable dispdate to just date, rename dispseq to sequence and add a type variable with its value is set to '2', meaning 'PBS', for every record in the file. Cut the file down vertically so that it contains only the tagged records (sequence ≥ 1), and horizontally so it only contains rootlpno, date, sex, age, sequence and type. Check that you are on track by confirming that your cut down file contains 174,131 records, including 6,820 records where sequence = 1. If all is well, save the cut down PBS source file with the filename PBScutdown.

SPSS



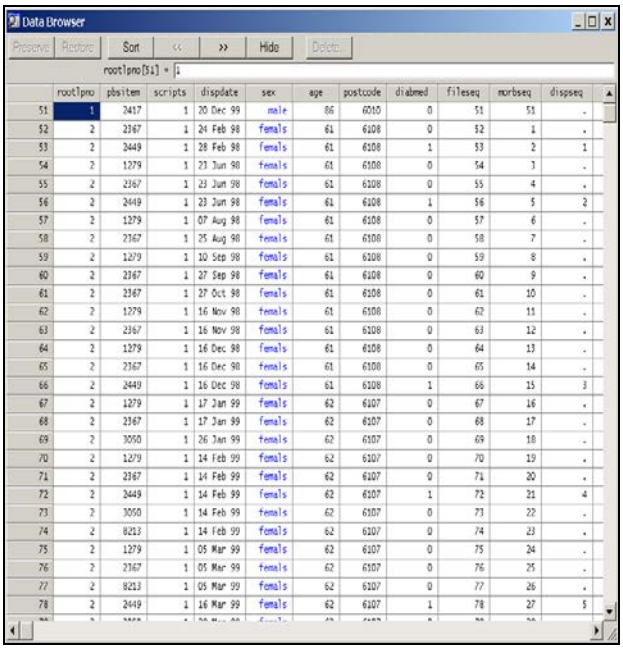
	rootlpno	pbsitem	scripts	dispdate	sex	age	postcode	diabmed	fileseq	morbseq	dispseq
51	1	2417	1	20.12.99	1	86	6010	0	51	51	.
52	2	2367	1	24.02.98	2	61	6108	0	52	1	.
53	2	2449	1	28.02.98	2	61	6108	1	53	2	1
54	2	1279	1	23.06.98	2	61	6108	0	54	3	.
55	2	2367	1	23.06.98	2	61	6108	0	55	4	.
56	2	2449	1	23.06.98	2	61	6108	1	56	5	2
57	2	1279	1	07.08.98	2	61	6108	0	57	6	.
58	2	2367	1	25.08.98	2	61	6108	0	58	7	.
59	2	1279	1	10.09.98	2	61	6108	0	59	8	.
60	2	2367	1	27.09.98	2	61	6108	0	60	9	.
61	2	2367	1	27.10.98	2	61	6108	0	61	10	.
62	2	1279	1	16.11.98	2	61	6108	0	62	11	.
63	2	2367	1	16.11.98	2	61	6108	0	63	12	.
64	2	1279	1	16.12.98	2	61	6108	0	64	13	.
65	2	2367	1	16.12.98	2	61	6108	0	65	14	.
66	2	2449	1	16.12.98	2	61	6108	1	66	15	3
67	2	1279	1	17.01.99	2	62	6107	0	67	16	.
68	2	2367	1	17.01.99	2	62	6107	0	68	17	.
69	2	3050	1	26.01.99	2	62	6107	0	69	18	.
70	2	1279	1	14.02.99	2	62	6107	0	70	19	.
71	2	2367	1	14.02.99	2	62	6107	0	71	20	.
72	2	2449	1	14.02.99	2	62	6107	1	72	21	4
73	2	3050	1	14.02.99	2	62	6107	0	73	22	.
74	2	8213	1	14.02.99	2	62	6107	0	74	23	.
75	2	1279	1	05.03.99	2	62	6107	0	75	24	.
76	2	2367	1	05.03.99	2	62	6107	0	76	25	.
77	2	8213	1	05.03.99	2	62	6107	0	77	26	.
78	2	2449	1	16.03.99	2	62	6107	1	78	27	5

SAS



	Patient ID number	PBS drug code	Number of scripts dispensed per item per date of supply	Date PBS drug dispensed/su	Patient gender	Patients age in years	Patient residential postcode	PBS Diabetes Medicatio	fileseq	morbseq	dispseq
51	1	2417	1	20/12/1999	1	86	6010	0	51	51	0
52	2	2367	1	24/02/1998	2	61	6108	0	52	1	0
53	2	2449	1	28/02/1998	2	61	6108	1	53	2	1
54	2	1279	1	23/06/1998	2	61	6108	0	54	3	0
55	2	2367	1	23/06/1998	2	61	6108	0	55	4	0
56	2	2449	1	23/06/1998	2	61	6108	1	56	5	2
57	2	1279	1	07/08/1998	2	61	6108	0	57	6	0
58	2	2367	1	25/08/1998	2	61	6108	0	58	7	0
59	2	1279	1	10/09/1998	2	61	6108	0	59	8	0
60	2	2367	1	27/09/1998	2	61	6108	0	60	9	0
61	2	2367	1	27/10/1998	2	61	6108	0	61	10	0
62	2	1279	1	16/11/1998	2	61	6108	0	62	11	0
63	2	2367	1	16/11/1998	2	61	6108	0	63	12	0
64	2	1279	1	16/12/1998	2	61	6108	0	64	13	0
65	2	2367	1	16/12/1998	2	61	6108	0	65	14	0
66	2	2449	1	16/12/1998	2	61	6108	1	66	15	3
67	2	1279	1	17/01/1999	2	62	6107	0	67	16	0
68	2	2367	1	17/01/1999	2	62	6107	0	68	17	0
69	2	3050	1	26/01/1999	2	62	6107	0	69	18	0
70	2	1279	1	14/02/1999	2	62	6107	0	70	19	0
71	2	2367	1	14/02/1999	2	62	6107	0	71	20	0
72	2	2449	1	14/02/1999	2	62	6107	1	72	21	4
73	2	3050	1	14/02/1999	2	62	6107	0	73	22	0
74	2	8213	1	14/02/1999	2	62	6107	0	74	23	0
75	2	1279	1	05/03/1999	2	62	6107	0	75	24	0
76	2	2367	1	05/03/1999	2	62	6107	0	76	25	0
77	2	8213	1	05/03/1999	2	62	6107	0	77	26	0
78	2	2449	1	16/03/1999	2	62	6107	1	78	27	5
79	2	3050	1	29/03/1999	2	62	6107	0	79	28	0
80	2	1279	1	02/04/1999	2	62	6107	0	80	29	0
81	2	2367	1	02/04/1999	2	62	6107	0	81	30	0
82	2	2449	1	02/04/1999	2	62	6107	1	82	31	6

Stata



	rootlpno	pbsitem	scripts	dispdate	sex	age	postcode	diabmed	fileseq	morbseq	dispseq
51	1	2417	1	20 Dec 99	male	86	6010	0	51	51	.
52	2	2367	1	24 Feb 98	female	61	6108	0	52	1	.
53	2	2449	1	28 Feb 98	female	61	6108	1	53	2	1
54	2	1279	1	23 Jun 98	female	61	6108	0	54	3	.
55	2	2367	1	23 Jun 98	female	61	6108	0	55	4	.
56	2	2449	1	23 Jun 98	female	61	6108	1	56	5	2
57	2	1279	1	07 Aug 98	female	61	6108	0	57	6	.
58	2	2367	1	25 Aug 98	female	61	6108	0	58	7	.
59	2	1279	1	10 Sep 98	female	61	6108	0	59	8	.
60	2	2367	1	27 Sep 98	female	61	6108	0	60	9	.
61	2	2367	1	27 Oct 98	female	61	6108	0	61	10	.
62	2	1279	1	16 Nov 98	female	61	6108	0	62	11	.
63	2	2367	1	16 Nov 98	female	61	6108	0	63	12	.
64	2	1279	1	16 Dec 98	female	61	6108	0	64	13	.
65	2	2367	1	16 Dec 98	female	61	6108	0	65	14	.
66	2	2449	1	16 Dec 98	female	61	6108	1	66	15	3
67	2	1279	1	17 Jan 99	female	62	6107	0	67	16	.
68	2	2367	1	17 Jan 99	female	62	6107	0	68	17	.
69	2	3050	1	26 Jan 99	female	62	6107	0	69	18	.
70	2	1279	1	14 Feb 99	female	62	6107	0	70	19	.
71	2	2367	1	14 Feb 99	female	62	6107	0	71	20	.
72	2	2449	1	14 Feb 99	female	62	6107	1	72	21	4
73	2	3050	1	14 Feb 99	female	62	6107	0	73	22	.
74	2	8213	1	14 Feb 99	female	62	6107	0	74	23	.
75	2	1279	1	05 Mar 99	female	62	6107	0	75	24	.
76	2	2367	1	05 Mar 99	female	62	6107	0	76	25	.
77	2	8213	1	05 Mar 99	female	62	6107	0	77	26	.
78	2	2449	1	16 Mar 99	female	62	6107	1	78	27	5

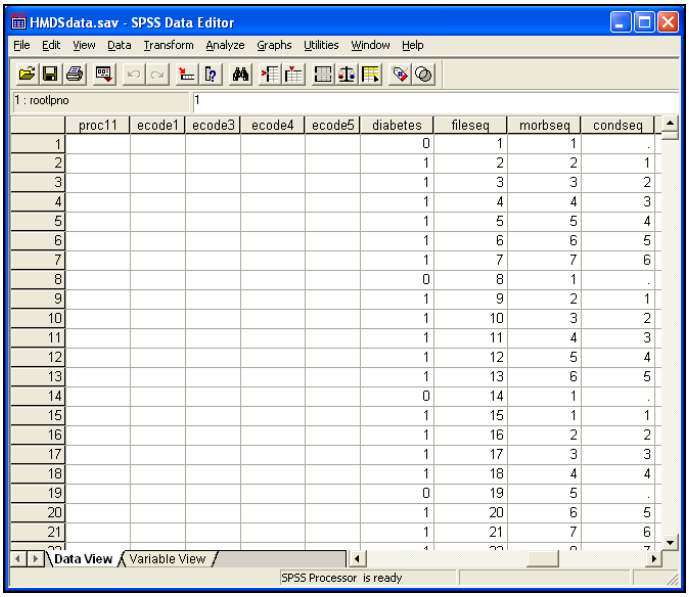
Preparation of the Hospital Morbidity Data

9. Open the correct HMDSdata file for your statistical software package and become familiar with its structure. Note that the 62,356 records are sorted by ascending admdate within ascending rootlpno.

10. Tag all records that mention a diabetes-related diagnosis code in any field, using a new variable, diabetes. Set diabetes to '1' for a tagged record and '0' otherwise.

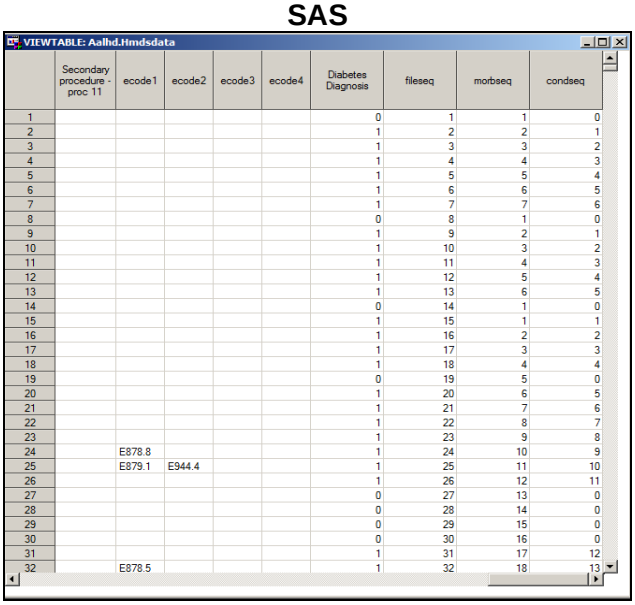
11. Create the following sequence variables in the EOR loading area: (i) a fileseq variable that numbers all records in ascending order; (ii) a morbseq variable that numbers the records within each block that relate to an individual; and (iii) a condseq variable that numbers only the tagged diabetes-related records within each block. Your file should look something like the screen shown opposite.

SPSS



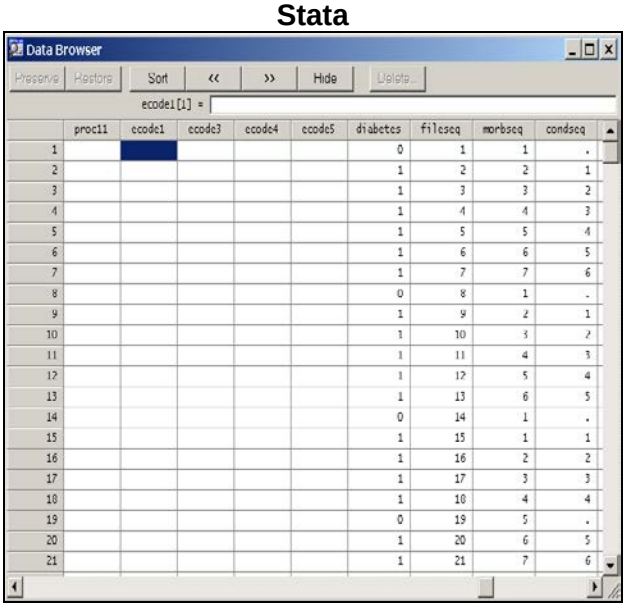
	proc11	ecode1	ecode3	ecode4	ecode5	diabetes	fileseq	morbseq	condseq
1						0	1	1	.
2						1	2	2	1
3						1	3	3	2
4						1	4	4	3
5						1	5	5	4
6						1	6	6	5
7						1	7	7	6
8						0	8	1	.
9						1	9	2	1
10						1	10	3	2
11						1	11	4	3
12						1	12	5	4
13						1	13	6	5
14						0	14	1	.
15						1	15	1	1
16						1	16	2	2
17						1	17	3	3
18						1	18	4	4
19						0	19	5	.
20						1	20	6	5
21						1	21	7	6

SAS



	Secondary procedure - proc 11	ecode1	ecode2	ecode3	ecode4	Diabetes Diagnosis	fileseq	morbseq	condseq
1						0	1	1	0
2						1	2	2	1
3						1	3	3	2
4						1	4	4	3
5						1	5	5	4
6						1	6	6	5
7						1	7	7	6
8						0	8	1	0
9						1	9	2	1
10						1	10	3	2
11						1	11	4	3
12						1	12	5	4
13						1	13	6	5
14						0	14	1	0
15						1	15	1	1
16						1	16	2	2
17						1	17	3	3
18						1	18	4	4
19						0	19	5	0
20						1	20	6	5
21						1	21	7	6
22						1	22	8	7
23						1	23	9	8
24		E878.8				1	24	10	9
25		E879.1	E944.4			1	25	11	10
26						1	26	12	11
27						0	27	13	0
28						0	28	14	0
29						0	29	15	0
30						0	30	16	0
31						1	31	17	12
32		E878.5				1	32	18	13

Stata



	proc11	ecode1	ecode3	ecode4	ecode5	diabetes	fileseq	morbseq	condseq
1						0	1	1	.
2						1	2	2	1
3						1	3	3	2
4						1	4	4	3
5						1	5	5	4
6						1	6	6	5
7						1	7	7	6
8						0	8	1	.
9						1	9	2	1
10						1	10	3	2
11						1	11	4	3
12						1	12	5	4
13						1	13	6	5
14						0	14	1	.
15						1	15	1	1
16						1	16	2	2
17						1	17	3	3
18						1	18	4	4
19						0	19	5	.
20						1	20	6	5
21						1	21	7	6

12. Save your HMDS source file with the changes made above as HMDSdata2. Now rename the variable admdate to just date, rename condseq to sequence and add a type variable with its value is set to '3', meaning 'HMDS', for every record in the file. Cut the file down vertically so that it contains only the tagged records (sequence ≥ 1), and horizontally so it only contains rootlpno, date, sex, age, sequence and type. Note that the date variable is initially in the wrong position (it needs to follow rootlpno) and, therefore, you should move its position (eg, by copying and pasting in this relatively small data set; or using the move command in Stata). Check that you are on track by confirming that your cut down file contains 21,274 records, including 5,328 records where sequence = 1. If all is well, save the cut down HMDS source file with the filename HMDScutdown.

Preparation of the Death Data

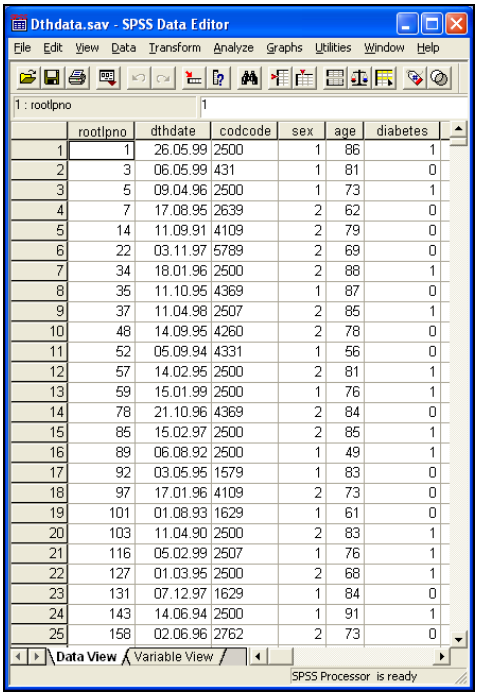
13. Open the correct Dthdata file for your statistical software package and become familiar with its structure. Note that the 2,260 records are sorted by ascending rootlpno.

14. Tag all records with a diabetes-related cause of death code, using a new variable, diabetes. Set diabetes to '1' for a tagged record and '0' otherwise.

15. Sequence variables are not required in this instance as there is only one record per individual. Your file should look something like the screen shown opposite.

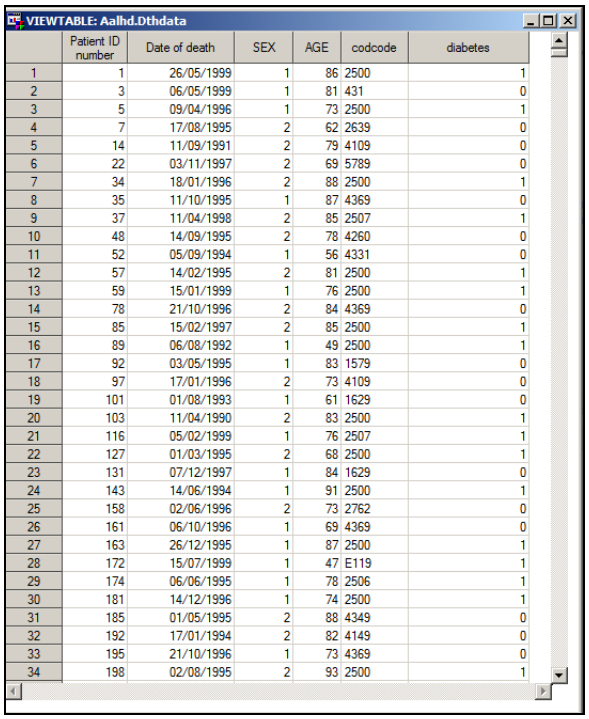
16. Save your death source file with the changes made above as Dthdata2. Now rename the variable dthdate to just date, add a sequence variable set to '1' for all records and a type variable set to '4', meaning 'deaths'. Cut the file down vertically so that it contains only the tagged records (diabetes = 1) and horizontally so it only contains rootlpno, date, sex, age, sequence and type. Check that you are on track by confirming that your cut down file contains 965 records. If all is well, save the cut down death source file with the filename Dthcutdown.

SPSS



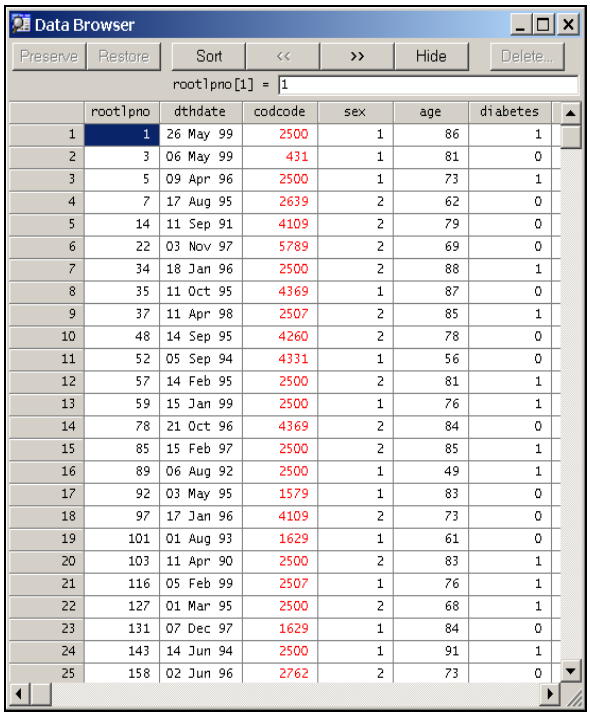
	rootlpno	dthdate	codcode	sex	age	diabetes
1	1	26.05.99	2500	1	86	1
2	3	06.05.99	431	1	81	0
3	5	09.04.96	2500	1	73	1
4	7	17.08.95	2639	2	62	0
5	14	11.09.91	4109	2	79	0
6	22	03.11.97	5789	2	69	0
7	34	18.01.96	2500	2	88	1
8	35	11.10.95	4369	1	87	0
9	37	11.04.98	2507	2	85	1
10	48	14.09.95	4260	2	78	0
11	52	05.09.94	4331	1	56	0
12	57	14.02.95	2500	2	81	1
13	59	15.01.99	2500	1	76	1
14	78	21.10.96	4369	2	84	0
15	85	15.02.97	2500	2	85	1
16	89	06.08.92	2500	1	49	1
17	92	03.05.95	1579	1	83	0
18	97	17.01.96	4109	2	73	0
19	101	01.08.93	1629	1	61	0
20	103	11.04.90	2500	2	83	1
21	116	05.02.99	2507	1	76	1
22	127	01.03.95	2500	2	68	1
23	131	07.12.97	1629	1	84	0
24	143	14.06.94	2500	1	91	1
25	158	02.06.96	2762	2	73	0

SAS



	Patient ID number	Date of death	SEX	AGE	codcode	diabetes
1	1	26/05/1999	1	86	2500	1
2	3	06/05/1999	1	81	431	0
3	5	09/04/1996	1	73	2500	1
4	7	17/08/1995	2	62	2639	0
5	14	11/09/1991	2	79	4109	0
6	22	03/11/1997	2	69	5789	0
7	34	18/01/1996	2	88	2500	1
8	35	11/10/1995	1	87	4369	0
9	37	11/04/1998	2	85	2507	1
10	48	14/09/1995	2	78	4260	0
11	52	05/09/1994	1	56	4331	0
12	57	14/02/1995	2	81	2500	1
13	59	15/01/1999	1	76	2500	1
14	78	21/10/1996	2	84	4369	0
15	85	15/02/1997	2	85	2500	1
16	89	06/08/1992	1	49	2500	1
17	92	03/05/1995	1	83	1579	0
18	97	17/01/1996	2	73	4109	0
19	101	01/08/1993	1	61	1629	0
20	103	11/04/1990	2	83	2500	1
21	116	05/02/1999	1	76	2507	1
22	127	01/03/1995	2	68	2500	1
23	131	07/12/1997	1	84	1629	0
24	143	14/06/1994	1	91	2500	1
25	158	02/06/1996	2	73	2762	0
26	161	06/10/1996	1	69	4369	0
27	163	26/12/1995	1	87	2500	1
28	172	15/07/1999	1	47	E119	1
29	174	06/06/1995	1	78	2506	1
30	181	14/12/1996	1	74	2500	1
31	185	01/05/1995	2	88	4349	0
32	192	17/01/1994	2	82	4149	0
33	195	21/10/1996	1	73	4369	0
34	198	02/08/1995	2	93	2500	1

Stata

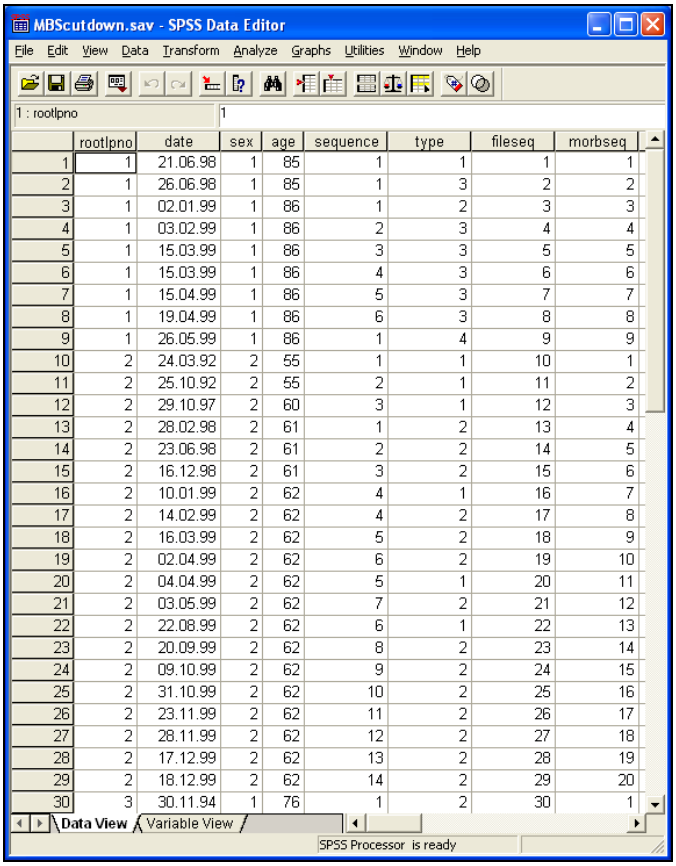


	rootlpno	dthdate	codcode	sex	age	diabetes
1	1	26 May 99	2500	1	86	1
2	3	06 May 99	431	1	81	0
3	5	09 Apr 96	2500	1	73	1
4	7	17 Aug 95	2639	2	62	0
5	14	11 Sep 91	4109	2	79	0
6	22	03 Nov 97	5789	2	69	0
7	34	18 Jan 96	2500	2	88	1
8	35	11 Oct 95	4369	1	87	0
9	37	11 Apr 98	2507	2	85	1
10	48	14 Sep 95	4260	2	78	0
11	52	05 Sep 94	4331	1	56	0
12	57	14 Feb 95	2500	2	81	1
13	59	15 Jan 99	2500	1	76	1
14	78	21 Oct 96	4369	2	84	0
15	85	15 Feb 97	2500	2	85	1
16	89	06 Aug 92	2500	1	49	1
17	92	03 May 95	1579	1	83	0
18	97	17 Jan 96	4109	2	73	0
19	101	01 Aug 93	1629	1	61	0
20	103	11 Apr 90	2500	2	83	1
21	116	05 Feb 99	2507	1	76	1
22	127	01 Mar 95	2500	2	68	1
23	131	07 Dec 97	1629	1	84	0
24	143	14 Jun 94	2500	1	91	1
25	158	02 Jun 96	2762	2	73	0

Creation of the Antepenultimate Platform

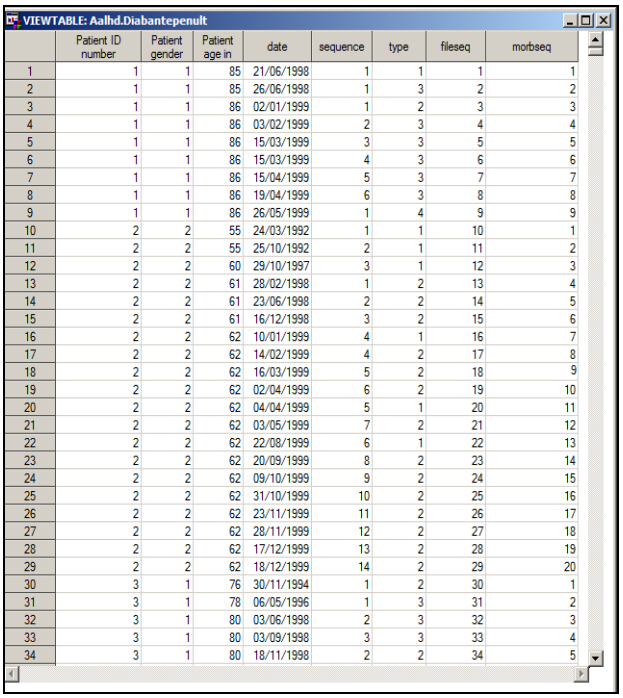
17. Re-open MBScutdown as your working file and using PBScutdown as the external file, merge the latter on to the bottom of the former. Repeat using HMDScutdown as the external file, and then again using Dthcutdown as the external file. Sort the final merged file by date within rootlpno in ascending order. Now create new fileseq and morbseq variables for this file. Check that your file of 225,941 records looks like the one opposite and, if so, save it as diabantepenult.

SPSS



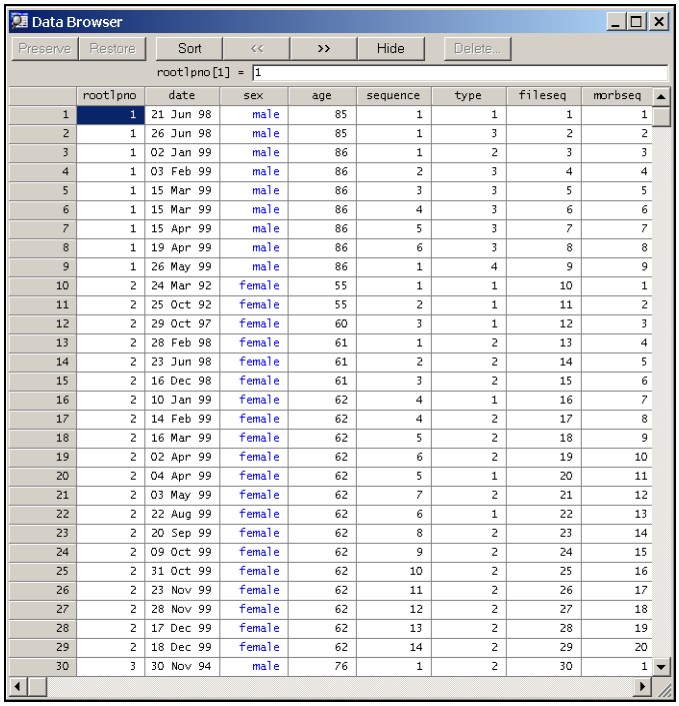
	rootlpno	date	sex	age	sequence	type	fileseq	morbseq
1	1	21.06.98	1	85	1	1	1	1
2	1	26.06.98	1	85	1	3	2	2
3	1	02.01.99	1	86	1	2	3	3
4	1	03.02.99	1	86	2	3	4	4
5	1	15.03.99	1	86	3	3	5	5
6	1	15.03.99	1	86	4	3	6	6
7	1	15.04.99	1	86	5	3	7	7
8	1	19.04.99	1	86	6	3	8	8
9	1	26.05.99	1	86	1	4	9	9
10	2	24.03.92	2	55	1	1	10	1
11	2	25.10.92	2	55	2	1	11	2
12	2	29.10.97	2	60	3	1	12	3
13	2	28.02.98	2	61	1	2	13	4
14	2	23.06.98	2	61	2	2	14	5
15	2	16.12.98	2	61	3	2	15	6
16	2	10.01.99	2	62	4	1	16	7
17	2	14.02.99	2	62	4	2	17	8
18	2	16.03.99	2	62	5	2	18	9
19	2	02.04.99	2	62	6	2	19	10
20	2	04.04.99	2	62	5	1	20	11
21	2	03.05.99	2	62	7	2	21	12
22	2	22.08.99	2	62	6	1	22	13
23	2	20.09.99	2	62	8	2	23	14
24	2	09.10.99	2	62	9	2	24	15
25	2	31.10.99	2	62	10	2	25	16
26	2	23.11.99	2	62	11	2	26	17
27	2	28.11.99	2	62	12	2	27	18
28	2	17.12.99	2	62	13	2	28	19
29	2	18.12.99	2	62	14	2	29	20
30	3	30.11.94	1	76	1	2	30	1

SAS



	Patient ID number	Patient gender	Patient age in	date	sequence	type	fileseq	morbseq
1	1	1	85	21/06/1998	1	1	1	1
2	1	1	85	26/06/1998	1	3	2	2
3	1	1	86	02/01/1999	1	2	3	3
4	1	1	86	03/02/1999	2	3	4	4
5	1	1	86	15/03/1999	3	3	5	5
6	1	1	86	15/03/1999	4	3	6	6
7	1	1	86	15/04/1999	5	3	7	7
8	1	1	86	19/04/1999	6	3	8	8
9	1	1	86	26/05/1999	1	4	9	9
10	2	2	55	24/03/1992	1	1	10	1
11	2	2	55	25/10/1992	2	1	11	2
12	2	2	60	29/10/1997	3	1	12	3
13	2	2	61	28/02/1998	1	2	13	4
14	2	2	61	23/06/1998	2	2	14	5
15	2	2	61	16/12/1998	3	2	15	6
16	2	2	62	10/01/1999	4	1	16	7
17	2	2	62	14/02/1999	4	2	17	8
18	2	2	62	16/03/1999	5	2	18	9
19	2	2	62	02/04/1999	6	2	19	10
20	2	2	62	04/04/1999	5	1	20	11
21	2	2	62	03/05/1999	7	2	21	12
22	2	2	62	22/08/1999	6	1	22	13
23	2	2	62	20/09/1999	8	2	23	14
24	2	2	62	09/10/1999	9	2	24	15
25	2	2	62	31/10/1999	10	2	25	16
26	2	2	62	23/11/1999	11	2	26	17
27	2	2	62	28/11/1999	12	2	27	18
28	2	2	62	17/12/1999	13	2	28	19
29	2	2	62	18/12/1999	14	2	29	20
30	3	1	76	30/11/1994	1	2	30	1
31	3	1	78	06/05/1996	1	3	31	2
32	3	1	80	03/06/1998	2	3	32	3
33	3	1	80	03/09/1998	3	3	33	4
34	3	1	80	18/11/1998	2	2	34	5

Stata



	rootlpno	date	sex	age	sequence	type	fileseq	morbseq
1	1	21 Jun 98	male	85	1	1	1	1
2	1	26 Jun 98	male	85	1	3	2	2
3	1	02 Jan 99	male	86	1	2	3	3
4	1	03 Feb 99	male	86	2	3	4	4
5	1	15 Mar 99	male	86	3	3	5	5
6	1	15 Mar 99	male	86	4	3	6	6
7	1	15 Apr 99	male	86	5	3	7	7
8	1	19 Apr 99	male	86	6	3	8	8
9	1	26 May 99	male	86	1	4	9	9
10	2	24 Mar 92	female	55	1	1	10	1
11	2	25 Oct 92	female	55	2	1	11	2
12	2	29 Oct 97	female	60	3	1	12	3
13	2	28 Feb 98	female	61	1	2	13	4
14	2	23 Jun 98	female	61	2	2	14	5
15	2	16 Dec 98	female	61	3	2	15	6
16	2	10 Jan 99	female	62	4	1	16	7
17	2	14 Feb 99	female	62	4	2	17	8
18	2	16 Mar 99	female	62	5	2	18	9
19	2	02 Apr 99	female	62	6	2	19	10
20	2	04 Apr 99	female	62	5	1	20	11
21	2	03 May 99	female	62	7	2	21	12
22	2	22 Aug 99	female	62	6	1	22	13
23	2	20 Sep 99	female	62	8	2	23	14
24	2	09 Oct 99	female	62	9	2	24	15
25	2	31 Oct 99	female	62	10	2	25	16
26	2	23 Nov 99	female	62	11	2	26	17
27	2	28 Nov 99	female	62	12	2	27	18
28	2	17 Dec 99	female	62	13	2	28	19
29	2	18 Dec 99	female	62	14	2	29	20
30	3	30 Nov 94	male	76	1	2	30	1

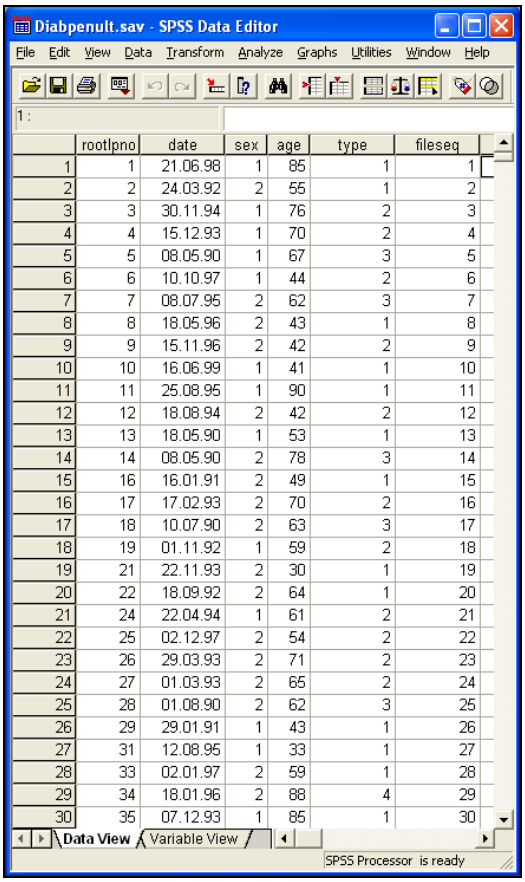
Creation of the Penultimate Platform

18. Using your diabantepenult file, select the index records from each source file, ie, all records where sequence = 1, and then delete the remaining records as well as the sequence variable, which is no longer of use. Create new fileseq and morbseq variables for this file. Check that your file has 20,352 records and, if so, save it as diabpenult.

Creation of the Ultimate Platform

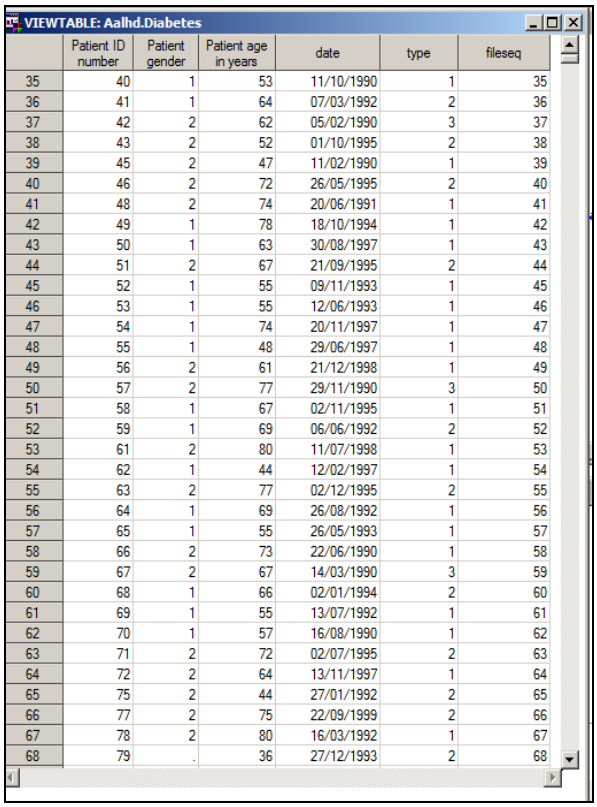
19. Using your diabpenult file, select the first record in each block, ie, all records where morbseq = 1, and then delete the remaining records. Now create a new fileseq variable and delete the morbseq variable, which is no longer of use. Check that your file of 10,675 records looks like the one opposite and, if so, save it as diabetes. Run a frequency distribution on type to confirm that 5,020 of the diabetics were first identified in the MBS data, 3,200 in the PBS data, 2,425 in the hospital morbidity data and 30 in the death data.

SPSS



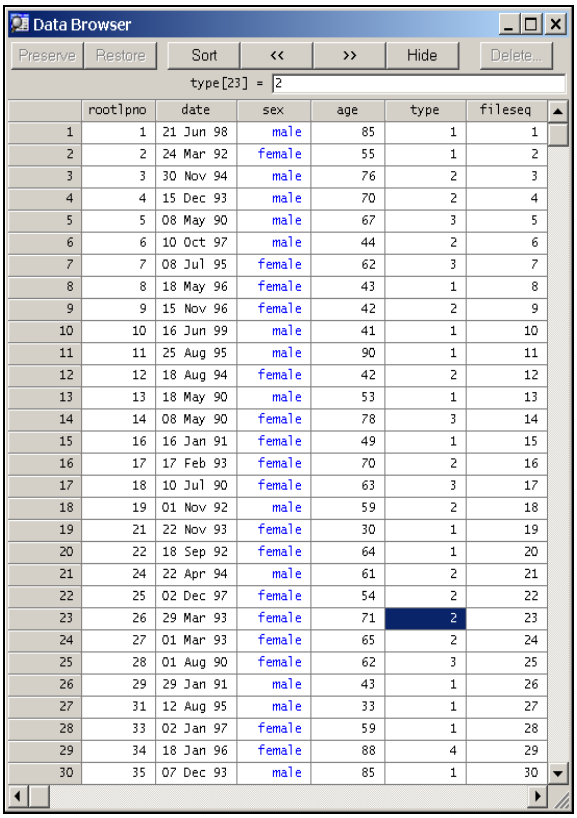
	rootlpno	date	sex	age	type	fileseq
1	1	21.06.98	1	85	1	1
2	2	24.03.92	2	55	1	2
3	3	30.11.94	1	76	2	3
4	4	15.12.93	1	70	2	4
5	5	08.05.90	1	67	3	5
6	6	10.10.97	1	44	2	6
7	7	08.07.95	2	62	3	7
8	8	18.05.96	2	43	1	8
9	9	15.11.96	2	42	2	9
10	10	16.06.99	1	41	1	10
11	11	25.08.95	1	90	1	11
12	12	18.08.94	2	42	2	12
13	13	18.05.90	1	53	1	13
14	14	08.05.90	2	78	3	14
15	16	16.01.91	2	49	1	15
16	17	17.02.93	2	70	2	16
17	18	10.07.90	2	63	3	17
18	19	01.11.92	1	59	2	18
19	21	22.11.93	2	30	1	19
20	22	18.09.92	2	64	1	20
21	24	22.04.94	1	61	2	21
22	25	02.12.97	2	54	2	22
23	26	29.03.93	2	71	2	23
24	27	01.03.93	2	65	2	24
25	28	01.08.90	2	62	3	25
26	29	29.01.91	1	43	1	26
27	31	12.08.95	1	33	1	27
28	33	02.01.97	2	59	1	28
29	34	18.01.96	2	88	4	29
30	35	07.12.93	1	85	1	30

SAS



	Patient ID number	Patient gender	Patient age in years	date	type	fileseq
35	40	1	53	11/10/1990	1	35
36	41	1	64	07/03/1992	2	36
37	42	2	62	05/02/1990	3	37
38	43	2	52	01/10/1995	2	38
39	45	2	47	11/02/1990	1	39
40	46	2	72	26/05/1995	2	40
41	48	2	74	20/06/1991	1	41
42	49	1	78	18/10/1994	1	42
43	50	1	63	30/08/1997	1	43
44	51	2	67	21/09/1995	2	44
45	52	1	55	09/11/1993	1	45
46	53	1	55	12/06/1993	1	46
47	54	1	74	20/11/1997	1	47
48	55	1	48	29/06/1997	1	48
49	56	2	61	21/12/1998	1	49
50	57	2	77	29/11/1990	3	50
51	58	1	67	02/11/1995	1	51
52	59	1	69	06/06/1992	2	52
53	61	2	80	11/07/1998	1	53
54	62	1	44	12/02/1997	1	54
55	63	2	77	02/12/1995	2	55
56	64	1	69	26/08/1992	1	56
57	65	1	55	26/05/1993	1	57
58	66	2	73	22/06/1990	1	58
59	67	2	67	14/03/1990	3	59
60	68	1	66	02/01/1994	2	60
61	69	1	55	13/07/1992	1	61
62	70	1	57	16/08/1990	1	62
63	71	2	72	02/07/1995	2	63
64	72	2	64	13/11/1997	1	64
65	75	2	44	27/01/1992	2	65
66	77	2	75	22/09/1999	2	66
67	78	2	80	16/03/1992	1	67
68	79	.	36	27/12/1993	2	68

Stata



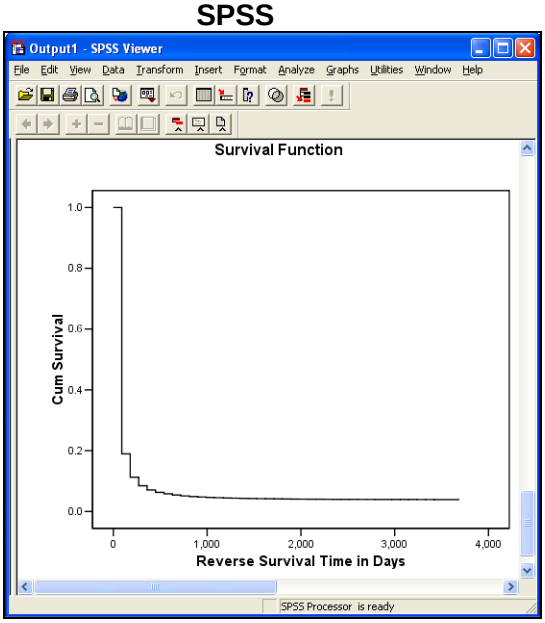
	rootlpno	date	sex	age	type	fileseq
1	1	21 Jun 98	male	85	1	1
2	2	24 Mar 92	female	55	1	2
3	3	30 Nov 94	male	76	2	3
4	4	15 Dec 93	male	70	2	4
5	5	08 May 90	male	67	3	5
6	6	10 Oct 97	male	44	2	6
7	7	08 Jul 95	female	62	3	7
8	8	18 May 96	female	43	1	8
9	9	15 Nov 96	female	42	2	9
10	10	16 Jun 99	male	41	1	10
11	11	25 Aug 95	male	90	1	11
12	12	18 Aug 94	female	42	2	12
13	13	18 May 90	male	53	1	13
14	14	08 May 90	female	78	3	14
15	16	16 Jan 91	female	49	1	15
16	17	17 Feb 93	female	70	2	16
17	18	10 Jul 90	female	63	3	17
18	19	01 Nov 92	male	59	2	18
19	21	22 Nov 93	female	30	1	19
20	22	18 Sep 92	female	64	1	20
21	24	22 Apr 94	male	61	2	21
22	25	02 Dec 97	female	54	2	22
23	26	29 Mar 93	female	71	2	23
24	27	01 Mar 93	female	65	2	24
25	28	01 Aug 90	female	62	3	25
26	29	29 Jan 91	male	43	1	26
27	31	12 Aug 95	male	33	1	27
28	33	02 Jan 97	female	59	1	28
29	34	18 Jan 96	female	88	4	29
30	35	07 Dec 93	male	85	1	30

Backcasting Correction for Incidence Estimation.

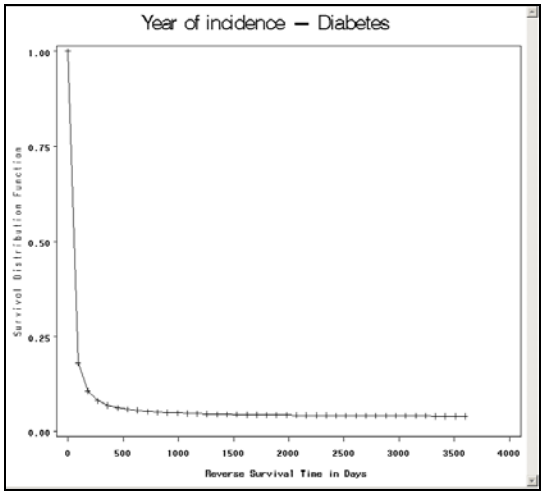
20. Reopen your Diabantepenult file and create a variable, yearinc, which draws the calendar year out of the date variable using the appropriate date function. Select the first record for each individual (ie, morbseq = 1) and obtain a frequency distribution on yearinc to observe the relatively larger numbers within the early years of 1990-92 (eg, 1,450 in 1990 compared with 973 in 1999). Some of this excess of apparent incidence of first mention of diabetes on an administrative record in the early years is likely to be due to a prevalent pool effect.

21. Reselect all records in the file. The next step is to flag every record in each block that is preceded by an earlier record on the same individual (ie, morbseq > 1), using a 0-1 binary indicator, previous. Calculate a reverse survival time, revsurti, for each record. For the first record in a block (ie, previous = 0) revsurti is the number of days elapsed from 31.12.1989 to date. For subsequent records in a block (ie, previous = 1) revsurti is the number of days from the date on the record immediately beforehand to the date on the current record. Use previous as the binary outcome and revsurti as the time variable in an actuarial survival analysis, where you should specify analysis intervals of 90 days up to 3690 days and request a survival curve plot. The plot should appear like the one opposite.

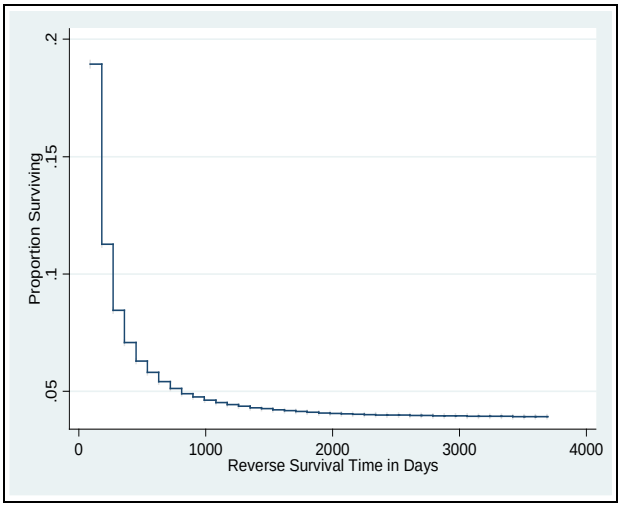
SPSS



SAS



Stata



Examine the life table and note where the hazard rate (slope of the curve) hits zero, which is at 2,160 days (make sure your hazard rates are displayed to 4-5 decimal places, otherwise a zero hazard rate may appear to occur earlier). At this point the survival is only 0.040 (4%). This means that in the absence of any documented history of whether or not an earlier event occurred, there is only a 4% chance that an apparent first-time event is genuine.

22. To assign an individual correction factor to every inaugural index record in the file, being the probability that each such record is a genuine first-time event, one must fit an equation to the relationship between revsurti and the points on the survival curve between time zero and 2160 days. Many different software programs are available for this purpose, including public domain software, fractional polynominals (eg, in Stata) and curve estimation procedures (eg, in SPSS). To save you time, assume that a good fit is obtained by using the power function:

$$\text{Survival} = 1.017 * \text{revsurti}^{-0.438}$$

Based on the Brameld-Holman method, the correction factor for each index record becomes:

$C_X = 1$ for revsurti $\geq$ 2160 days
$C_X = 0.04/(1.017 * \text{revsurti}^{-0.438})$ for revsurti < 2160 days

Write syntax to calculate and apply these correction factors. Bearing in mind that the equation does not fit the reverse time survival curve perfectly, especially for extreme values of revsurti such as the three index records apparently incident on 1.1.1990, you should constrain the value range of C_X between 0 and 1. **SPSS and SAS users:** Weight the records by C_X and rerun the frequency distribution of inaugural index records by yearinc. **Stata users:** Sum the C_X across inaugural index records by yearinc. Frequency weights in Stata appear to have to be integers to work. One can get around this be summing the C_X for each yearinc for morbseq=1 to get the incident cases in each calendar year. **R Users:** Sum the Cx across inaugural index years by yearinc, and round.

Compare these corrected incident events to those obtained prior to correction. Note how the largest reduction has affected the figures for 1990 (1,450 down to 501 incident cases), whereas the figures from 1994 onwards were unaffected by the correction procedure.

	1990	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
Uncorrected	1,450	1,061	1,459	985	947	971	1.034	891	904	973
Corrected	501	653	1,137	889	938	971	1,034	891	904	973

Capture-Recapture Correction for Period Prevalence Estimation

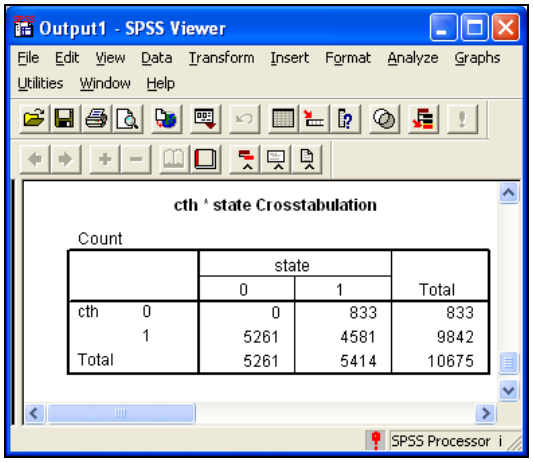
23. Reopen the diabpenult file and turn it upside-down by sorting it in descending order of fileseq. Write syntax that will create six new 0-1 binary variables in the EOR loading area of each inaugural index record where morbseq = 1. Four of the new variables, mbs, pbs, hmds and dth, are each set to '1' to indicate that the individual has been registered in the corresponding data set as evidenced by the penultimate platform file containing the corresponding index record from that data set. A fifth variable, cth, indicates that at least one Commonwealth record exists as evidenced by mbs = 1 or pbs = 1; and the sixth variable, state, indicates that at least one State record exists as evidenced by hmds =1 or dth = 1. Turn the file right-way-up again by sorting fileseq in ascending order. **Stata and R users:** The syntax in Stata to do this works whether or not the file is upside-down.

24. Select the inaugural index records (ie, morbseq = 1) and perform a cross tabulation of cth by state. Your should obtain the result shown opposite. You may now apply the Chapman estimator,

$$N = \frac{(n_1+1)(n_2+1)}{(n_{1.2}+1)} - 1$$

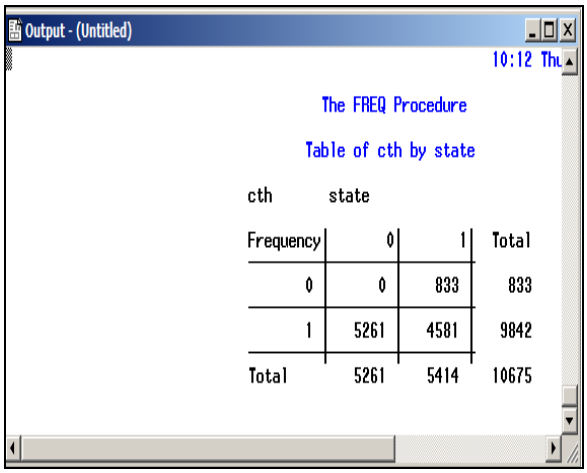
where N is the estimated total number of cases, n_1 is the marginal total of Commonwealth registrants, n_2 is the marginal total of State registrants and n_{12} is the number of individuals registered in both Commonwealth and State data sets. Here, N = 11,631 compared with the observed total count of 10,675, suggesting an ascertainment fraction of 92%.

SPSS



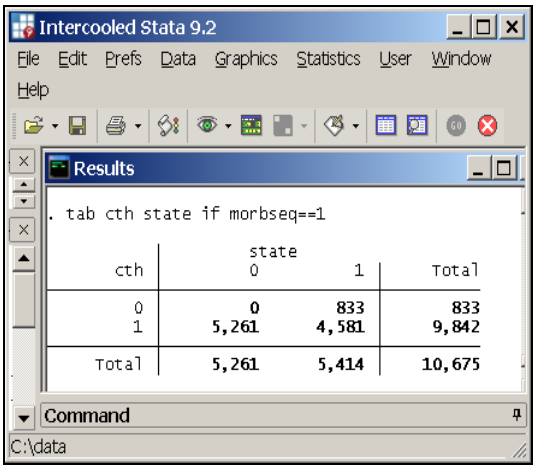
		state 0	state 1	Total
cth	0	0	833	833
	1	5261	4581	9842
Total		5261	5414	10675

SAS



cth	state		
Frequency	0	1	Total
0	0	833	833
1	5261	4581	9842
Total	5261	5414	10675

Stata



cth	state 0	state 1	Total
0	0	833	833
1	5,261	4,581	9,842
Total	5,261	5,414	10,675

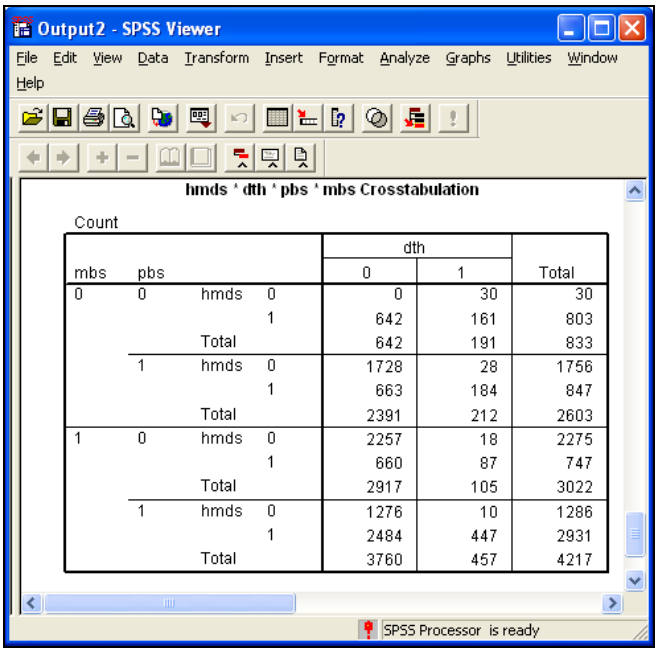
The 95% confidence limits are given by:

$$\underline{N}, \overline{N} = N \mp 1.96\sqrt{\frac{(n_1+1)(n_2+1)(n_1-n_{1.2})(n_2-n_{1.2})}{(n_{1.2}+1)^2(n_{1.2}+2)}}$$

which is this case equate to an interval of 11,534 – 11,728 (ascertainment fraction 91%-93%).

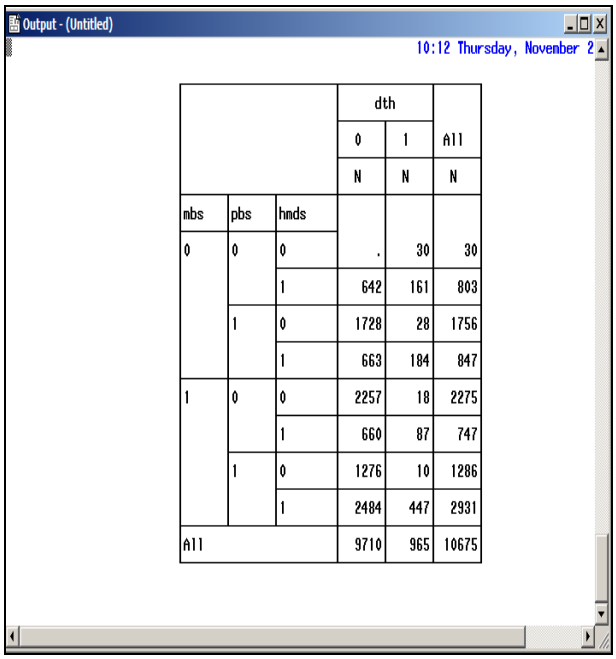
Also perform a cross-tabulation of dth by hmds by pbs by mbs. These results shown below are the multi-way contingency tabular input that may be used in a log-linear analysis to estimate the number of missing cases, taking into account the interactions between the four different data sources due to failure of the assumptions of independence between them. Performing this analysis is beyond the scope of the course. For your information, a Poisson log-linear regression model, fixing the cell structure to account for one missing cell (ie, the unknown number of cases where dth = hmds = pbs = mbs = 0) and arbitrarily specified to include all main effects, all two-way interactions and three of the three-way interactions between the four sources, yielded an estimate of around 1,821 missing cases (ascertainment fraction = 10,675/12,496 = 85%), almost double the number derived from the Chapman estimator (11,631 – 10,675 = 956). Underestimation by the Chapman method is typical and tends to reflect a positive dependence between ascertainment in two sources.

SPSS



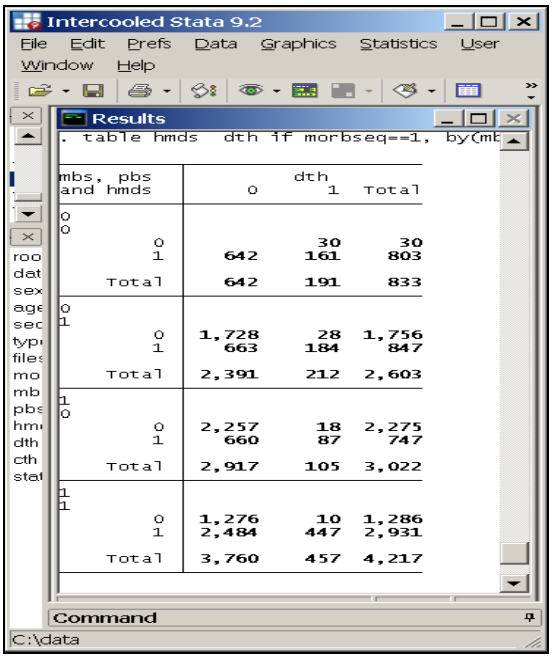
mbs	pbs			dth 0	dth 1	Total
0	0	hmds	0	0	30	30
			1	642	161	803
		Total		642	191	833
	1	hmds	0	1728	28	1756
			1	663	184	847
		Total		2391	212	2603
1	0	hmds	0	2257	18	2275
			1	660	87	747
		Total		2917	105	3022
	1	hmds	0	1276	10	1286
			1	2484	447	2931
		Total		3760	457	4217

SAS



			dth 0	dth 1	All
			N	N	N
mbs	pbs	hmds			
0	0	0	.	30	30
		1	642	161	803
	1	0	1728	28	1756
		1	663	184	847
1	0	0	2257	18	2275
		1	660	87	747
	1	0	1276	10	1286
		1	2484	447	2931
All			9710	965	10675

Stata



mbs, pbs and hmds	0	1	Total
0 0			
0		30	30
1	642	161	803
Total	642	191	833
0 1			
0	1,728	28	1,756
1	663	184	847
Total	2,391	212	2,603
1 0			
0	2,257	18	2,275
1	660	87	747
Total	2,917	105	3,022
1 1			
0	1,276	10	1,286
1	2,484	447	2,931
Total	3,760	457	4,217

References

Brameld KJ, Holman CDJ, Lawrence DM, *et al*. **Improved methods for estimating incidence from linked hospital morbidity data.** *Int J Epidemiol* 2003; 32: 617-624.

Bruno G, LaPorte RE, Merletti F, *et al*. **National diabetes programs. Application of capture-recapture to count diabetes?** *Diabetes Care* 1994; 17: 548-556.

Carrao G, Bagnardi V, Vittadini G, *et al*. **Capture-recapture method to size alcohol related problems in a population.** *J Epidemiol Community Hth* 2000; 54 603-610.

REVIEW 1:
'PRINCIPLES OF ADVANCED LINKED DATA ANALYSIS'

Please check that you are familiar with the following topics:

- Working with medical services and pharmaceutical data ☐
- Backcasting methods for correction of the prevalent pool effect ☐
- Capture-recapture methods for correction of case underascertainment ☐

Please check that you understand the meaning of the following terms:

- Primary and secondary study base ☐
- Pathognomic health services ☐
- Census of cases ☐
- Prevalent pool effect ☐
- Clearance period ☐
- Retrograde survival curve ☐
- Chapman estimator ☐
- Log-linear model ☐
- Multiple recapture census ☐
- Cell structure and structural zero ☐
- Types 1, 2 and 3 data files ☐
- Simple file group and complex file group ☐
- Primary and secondary platform and preliminary processing ☐
- Target condition/service ☐
- Antepenultimate, penultimate and ultimate secondary platform file ☐

Please check that you have demonstrated that you have acquired the following skills:

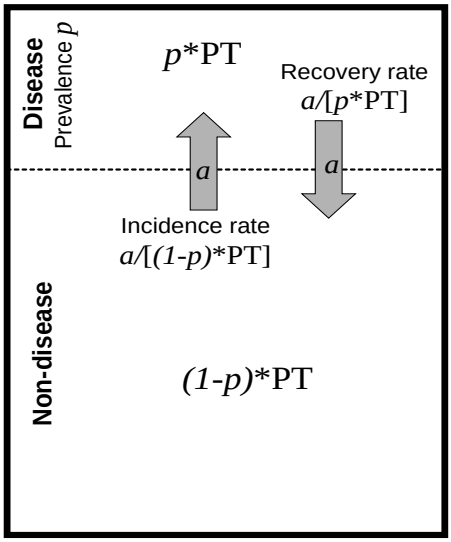
- Construct a secondary platform ☐
- Create a condition/service sequence variable ☐
- Implement the Brameld-Holman Backcasting correction ☐
- Implement capture-recapture estimation ☐

2. ADVANCED TOPICS IN CLINICAL POPULATION DYNAMICS

2.1 General Disease Dynamics in Populations

One-Stage Model of Disease Prevalence

In the simple case where a dynamic population in a closed system is partitioned into non-disease and disease status groups, the population will be in ***dynamic equilibrium*** when the flow of the non-diseased population into the diseased state (ie, arising from the incidence rate, IR) equals the flow of the diseased population back into the non-diseased state (ie, arising from the termination rate, TR, which in a closed system is the recovery rate).



Closed system with total person-time = PT

In a closed system, if the recovering diseased individuals, a, arise out of p*PT person-years at risk, then the average duration spent in the diseased state must be D = (p*PT)/a = 1/TR. It follows that:

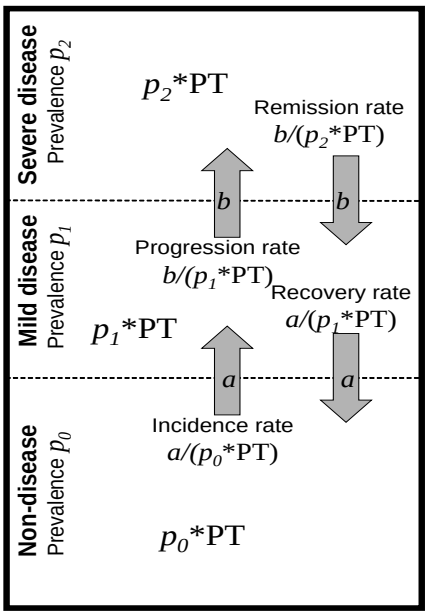
$$
\begin{aligned}
a &= \mathrm{TR}(p.\mathrm{PT}) &&= \mathrm{IR}[(1-p).\mathrm{PT}] \\
&\quad \mathrm{TR}.p &&= \mathrm{IR}(1-p) \\
&\quad p/(1-p) &&= \mathrm{IR}/\mathrm{TR} \\
&\quad &&= \mathrm{IR}.\mathrm{D} \\
&\quad (1-p)/p &&= 1/(\mathrm{IR}.\mathrm{D}) \\
&\quad 1/p - 1 &&= 1/(\mathrm{IR}.\mathrm{D}) \\
&\quad 1/p &&= 1/(\mathrm{IR}.\mathrm{D}) + 1 \\
&\quad &&= 1/(\mathrm{IR}.\mathrm{D}) + (\mathrm{IR}.\mathrm{D})/(\mathrm{IR}.\mathrm{D}) \\
&\quad &&= (1 + \mathrm{IR}.\mathrm{D})/(\mathrm{IR}.\mathrm{D}) \\
&\quad p &&= \mathrm{IR}.\mathrm{D}/(1 + \mathrm{IR}.\mathrm{D})
\end{aligned}
$$

Thus the prevalence of disease, p, is equal to IR.D/(1 + IR.D), which approximates to IR.D if the prevalence is low.

Real populations in research are rarely closed systems. In an open system, there are other means of entering and exiting the two health/disease status groups, such as birth, death, inward and outward migration. However, provided that p remains constant (***steady state prevalence***) the above formulation will still hold true.

Two-Stage Model of Disease Prevalence

The closed and open system models explained above may be extended to cover the situation where the diseased group is further partitioned according to disease severity, eg, without and with complications. The following diagram illustrates this situation:



Closed system with total person-time = PT

Note that under the conditions of dynamic equilibrium across the two partitions between the disease stages:

$$p_2*PT*RemR_2 = b = p_1*PT*ProgR_1$$

and $$p_0*PT*IR_0 = a = p_1*PT*RecovR_1$$

Thus

$$p_0*IR_0 + p_2*RemR_2 = p_1*RecovR_1 + p_1*ProgR_1$$
$$= p_1*TR_1$$
$$= p_1/D_1$$

Therefore

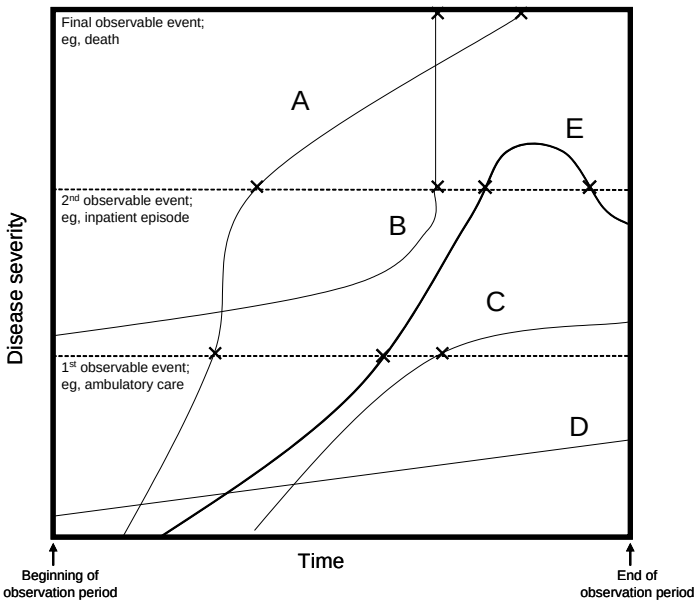
$$D_1(p_0*IR_0 + p_2*RemR_2) = p_1$$

In this instance the prevalence of mild disease, $p_1 = D_1(p_0IR_0 + p_2RemR_2)$ where D_1 is the average duration of time spent in the mild diseased state. The same formula will apply in an open system with steady state prevalences of the three health/disease status groups.

Continuous Disease Severity Models

An even closer approximation to reality is attained by adopting a view of the natural history of disease that allows for the possibility of continuous disease progression. At the individual level, different patients progress to higher levels of disease severity at different rates. This conceptualisation is an especially helpful model in the analysis of linked health data, where population dynamics may be built up from numerous observations made on individuals.

Person-time 'areas', like those shown in the figures above, are composed of the many small time traces of individual people. In the diagram opposite, we see the contributions of five patients to the total person-time observed during a data linkage study. Although disease severity cannot be measured continuously, there are ***sentinel events*** that indicate when a patient has crossed the threshold into a higher disease severity category. Patient A progresses through all the disease stages. Patient B is first observed in the second stage and dies suddenly. Patient C progresses by one stage and remains stable. Individual D never progresses sufficiently to register a sentinel event. Patient E recover sufficiently to re-establish a pattern of ambulatory care.



Advanced Theory Topic C:
2.2 Sentinel Event, Reference Point and Active Disease Models for Characterising Clinical Populations and Patterns of Care

Clinical Populations

The concept of clinical populations builds on the notion of health care episodes that was covered in section 3.10 of the Introductory Course. When health or disease status can be defined in a way that has implications for the planning of clinical services, the enumerated patients are known as a ***clinical population***. Membership of a clinical population is a state that exists between defined ***entry points*** and ***exit points***. For a clinical population to be a practically useful concept these entry and exit events need to be ascertainable.

Example: Patients with end-stage renal failure might enter the clinical population on the date of the first in-hospital record to mention chronic renal failure, when a long-term pattern of renal dialysis is first established, or upon a post-transplant admission in which a failed transplant kidney is removed. Patients might exit from this clinical population on the hospital separation date following a renal transplant, on a date of death or, otherwise, at the end of a study period.

The significance of a clinical population is that it provides a more meaningful denominator for calculation of utilisation measures than the general population, because it removes one of the major sources of variation in utilisation rates, namely differences in the size of the clinical population, eg, between different geographic areas. The size of a dynamic clinical population is expressed in person-time.

Obtaining an accurate fix on biologically and clinically meaningful points of entry and exit is a frequent difficulty encountered in defining the membership of clinical populations. Three methods of dealing with this problem are ***sentinel event models***, ***reference point models*** and ***active disease models***.

Sentinel Event Models

The above example of end-stage renal failure illustrates the sentinel event approach in which specific ***entry events*** and ***exist events***, which are directly ascertainable from available data, are defined as the dates of entry to and exit from the clinical population.

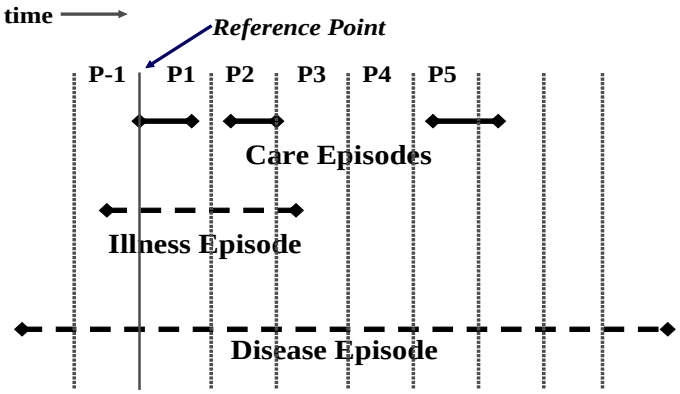
As already foreshadowed, additional sentinel events (***transition events***) may be used to partition the clinical population into strata according to disease severity. Examples of linked data that may be suitable for ascertainment of transition events include a first inpatient episode for a serious complication of a chronic disease (eg, first amputation in peripheral vascular disease) or first filled script representing the commencement of more aggressive pharmacotherapy for a chronic disease (eg, insulin for diabetes mellitus; polypharmacotherapy for asthma or hypertension). Each patient's person-time contribution to the clinical population may be ***partitioned*** using data about transition events, thus building up a picture at the population level of the amounts of person-time under observation in each clinical subgroup. If observed utilisation or outcome events are similarly partitioned then rates of utilisation and outcomes may be compared across disease severity groups; or compared across sociodemographic groups or calendar time groups within patients of similar disease severity. This concept is explained further below.

Reference Point Models

In the situation where there are insufficient or insufficiently reliable transition events to subdivide a clinical population, an alternative approach is to partition the health care episode, and potentially the pre- and post-episode periods, into a number of intervals defined relative to a single ***reference point*** in time, such as date of diagnosis or date of first admission. Each individual patient's person-time might be thereby partitioned into the 'first year after diagnosis', 'second year after diagnosis', and so on; and at the population level the clinical population may be similarly stratified. The choice of intervals,

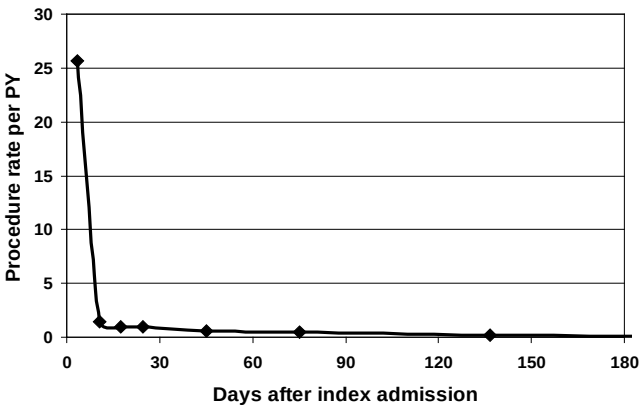
whether in days, months or years, is one for the researcher to make, guided by the natural history of the disease (eg, whether an acute or chronic condition) and the nature of the research questions to be addressed.

Reference point models draw on indirect rather than direct inferences about variations in severity of illness across different intervals. For example, no determination is made as to whether or not active disease is present within each patient in each interval, although it is usually implied that health care needs are greatest immediately after diagnosis. Clinical insights into how disease, illness and care episodes usually unfold in regard to the condition under investigation will be invaluable in deciding how a reference point model, as illustrated conceptually below, should be designed and interpreted. One must also be aware of the possibility of 'reference point drift' over time due, for example, to advances in the use of diagnostic technology.



<u>Example</u>: Patients with ureteric calculus suffer from an acute and painful condition that frequently requires urgent procedural intervention using extracorporeal shock wave lithotripsy, minimally invasive endourological procedures to remove or obliterate the calculus, or, in some cases, open surgery and even nephrectomy. Multiple procedures are often applied in a hierarchical manner if conservative management or the less invasive procedures prove ineffective. A researcher wishing to compare rates of procedural intervention between clinical populations resident in metropolitan and rural areas could use a reference point model, initially to plot out the rates of procedural utilisation in intervals after index admission in order to determine empirical boundaries for the care episode.

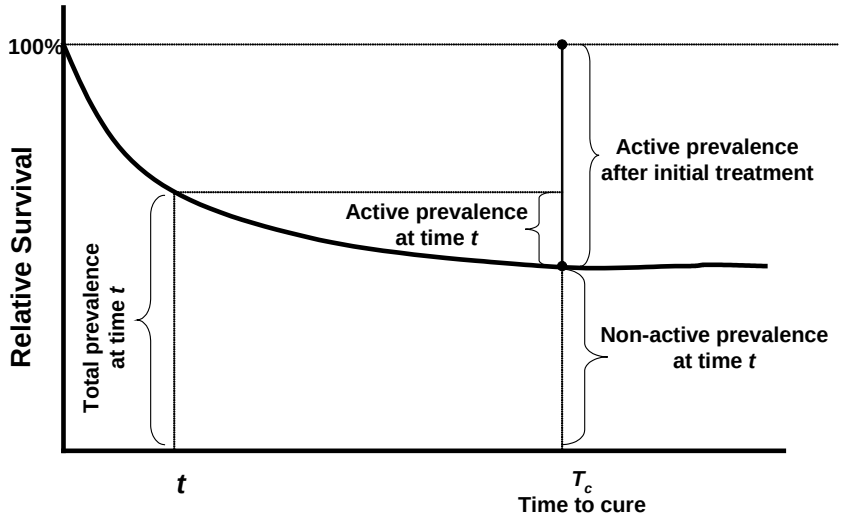
If the result was to be as shown opposite, it would indicate that 99% of all procedures following index admission occur within an interval of 30 days and most of these occurred within the first 7 days. On this basis a clinical population subgroup of 'within 30 days of index admission' could be defined and used as the basis for a comparison of procedure rates between metropolitan and rural patients.



Note that a reference point used to define time-related intervals need not always be at or near the commencement of the interval(s) used in analysis. They may also be used to anchor the end of an interval or the mid-point of an interval. Examples of the former include use of death-anchored retrospective intervals to measure health care utilisation in the last year of life; and use of index admission-anchored retrospective intervals to identify prior comorbid conditions and associated health care utilisation for the purposes of risk adjustment. A reference point may also be a calendar time for all patients (eg, introduction of a new policy) or a specific patient age. Two reference points that partition person-time by two dimension are also possible; eg, calendar date and date of diagnosis.

Active Disease Models

These models are most applicable for diseases that may be rendered inactive by an initial cure, but if not cured will remain active and eventually lead to death or some other observable adverse outcome if the patient is followed for a sufficient length of time. Cancer is especially well suited to this method of analysis, which enables the prevalence of a history of cancer to be partitioned into the prevalence of active cancer and the prevalence of patients with inactive disease. The diagram below illustrates the principles involved in estimating the active prevalence of cancer.



In this model, first proposed by Colman *et al*, the proportion of patients assumed to be cured after initial treatment is estimated by the percentage expected to survive their cancer longer than the time to cure, T_c, which is within 5-15 years for most cancers. T_c is taken to be the follow-up time where the mortality rate in cancer patients approximates to that of an age-sex matched subgroup of the general population based on relative survival (RS) methods. The probability of a cancer patient *i* having active disease at time *t* is then defined as:

$$p_i(t) = 1 - \frac{RS(T_c)}{RS(t)}$$

where $RS(T_c)$ is the relative survival at time to cure and $RS(t)$ is the relative survival at follow-up time t, so that $p_i(t) = 0$ when $t = T_c$. The model is based on the premise that patients who remain prevalent with persistent disease requiring ongoing health care are represented approximately by those who died from the excess risk of mortality before T_c. Active prevalence is therefore the proportion of cancer patients who are destined to die from their cancer (= 1 – proportion destined to survive).

A variation of the active disease model when applied to cancer is to allow for an initial treatment phase, where the prevalence of active cancer is assumed to be 100%, thus effecting a small phase shift of the relative survival curve to the right. An appropriate duration of this initial treatment phase may be determined empirically using a reference point model to determine by what interval initial treatment is complete in a large majority of patients.

Active disease models may be used to partition the numerator of a point prevalence by summing $p_i(t)$ over all of the prevalent cancer patients based on the follow-up time, *t*, for each patient to that point. Person-year clinical population denominators may be similarly weighted using the $p_i(t)$ at the mid-point of each patient's person-time contribution.

References

Brameld KJ, Holman CDJ, Bass AJ, *et al*. **Hospitalisation of the elderly during the last year of life: an application of record linkage in Western Australia 1985-1994.** *J Epidemiol Community Hlth* 1998; 52: 740-744.

Brameld KJ, Holman CDJ, Thomas MAB, *et al*. **Use of a state data bank to measure incidence and prevalence of a chronic disease: end-stage renal failure.** *Am J Kidney Dis* 1999; 34: 1033-1039.

Brameld KJ, Holman CDJ, Threlfall TJ, *et al*. **Increasing "active prevalence" of cancer in Western Australia and its implications for health services.** *ANZ J Public Health* 2002; 26: 164-169.

Brameld KJ, Thomas MAB, Holman CDJ, *et al*. **Validation of linked administrative data on end-stage renal failure: application of record linkage to a 'clinical base population'.** *ANZ J Public Health* 1999; 23: 464-467.

Coldman AJ, McBride ML, Braun T. **Calculating the prevalence of cancer.** *Stat Med* 1992; 11: 1579-1589.

Crombie IK, Davies HTO. **Design, Conduct and Interpretation of Health Services Research.** Chichester: Wiley & Sons, 1996.

Harper AC, Holman CDJ, Dawes VP. **The Health of Populations. An Introduction.** Chapter 5. **Risks II: Epidemiological Methods of Measurement**, pp.77-116. Melbourne: Churchill-Livingstone, 1994.

Holman CDJ, Preen DB, Baynham NJ, *et al*. **A multipurpose Australian comorbidity scoring system performed better than the Charlson index.** *J Clin Epidemiol* 2005; 58: 1006 1014.

Holman CDJ, Wisniewski ZS, Semmens JB, *et al*. **Changing treatments for primary urolithiasis: impact on services and renal preservation in 16 679 patients in Western Australia.** *BJU International* 2002; 90: 7-15.

Hornbrook MC, Hurtado AV, Johnson RE. **Health care episodes: definition, measurement and use.** *Medical Care Review* 1985; 42: 163-218.

Hornbrook MC. **Definition and measurement of episodes of care in clinical and economic studies.** In: Grady ML, Weis KA (Eds), *Conference Proceedings. Cost Analysis Methodology for Clinical Practice Guidelines*, pp.15-45. AHCPR Publication 95-0001, US Department of Health and Human services, 1995.

Kleinbaum DG, Kupper LL, Morgenstern H. **Epidemiologic Research: Principles and Quantitative Methods.** New York: John Wiley & Sons, 1982.

Preen DB, Holman CDJ, Lawrence DM, *et al*. **Hospital chart review provided more accurate comorbidity information than data from a general practitioner survey or an administrative database.** *J Clin Epidemiol* 2004; 57: 1295-1304.

Preen DB, Holman CDJ, Lawrence DM, *et al*. **Modelling comorbidity recency, duration and severity with risk adjustment for mortality and morbidity outcomes with administrative data.** *Aust Epidemiologist* 2004; 11: 12-19.

Preen DB, Holman CDJ, Semmens JB, *et al*. **Length of comorbidity lookback period affected regression model performance of administrative health data.** *J Clin Epidemiol* 2006; 59: 940-946.

Rothman KJ, Greenland S. **Modern Epidemiology. Second Edition.** Philadelphia: Lippincott-Raven, 1998.

TRAINING SESSION 2: HEALTH OCCURRENCES IN PARTITIONED AT RISK POPULATIONS

IMPORTANT NOTICE

The practice data sets provided for training purposes in this course are fictitious. All results based on analysis of these fictitious data are scientifically useless. The course organisers and instructor take no responsibility for any misuse of the spurious results arising from the analysis of these fictitious data sets.

Background to Training Session 2

This training session requires participants to develop a keen sense of clinical population dynamics and a clear conceptualisation of how syntax is written to identify and summate the contributions of individual patients to denominators and numerators of epidemiologic measures used in health services research. You will firstly implement a sentinel events model, which partitions the clinical population of diabetics into three different disease severity subgroups using transition events defined with pharmacotherapeutic criteria. You will estimate the point prevalence of each of the diabetes severity groups on each of two calendar dates. You will then measure and compare rates of health care utilisation across the disease severity groups. The exercise is based on a currently active research program on diabetes mellitus led by Professor Preen and Professor Holman.

Training Session Resources

You will be making use of data files saved from exercise 1. These will consist of your four tagged and sequenced source data files and your ultimate platform file.

If you are using this statistical package …	*… you should use this save file.*
SPSS	Dthdata2.sav MBSdata2.sav PBSdata2.sav diabetes.sav
SAS	Dthdata2.sas7bdat MBSdata2.sas7bdat PBSdata2.sas7bdat diabetes.sas7bdat
Stata	Dthdata2.dta MBSdata2.dta PBSdata2.dta diabetes.dta
R	Dthdata2.rda MBSdata2.rda PBSdata2.rda diabetes.rda

Note that the following additional codes are relevant to this exercise:

Data Set	Coding System	Codes
MBS	*Medicare Benefits Schedule*	GP visits: 1 thru 51, 160 thru 164, 170 thru 173, 193 thru 199, 601 thru 602, 700 thru 712, 720 thru 779, 900 thru 903, 2501 thru 2509, 2517 thru 2526, 2546 thru 2559, 2574 thru 2578, 2721 thru 2727, 5000 thru 5267

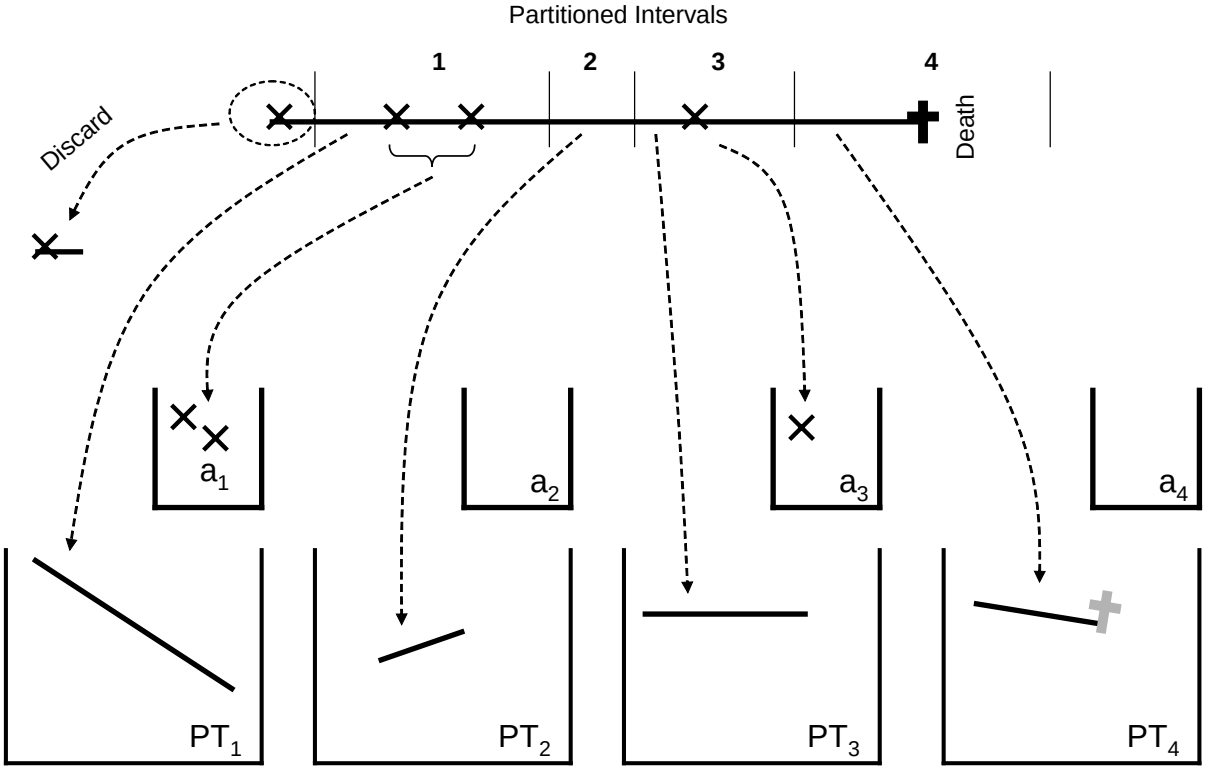
To assist you in syntax preparation, the text of this table is contained in a word document file called AALHD.doc in LMS.

WARNING: You should not rely on these codes in any analysis separate from the course as they may contain errors or be incomplete.

General Concepts Involved in Partitioning

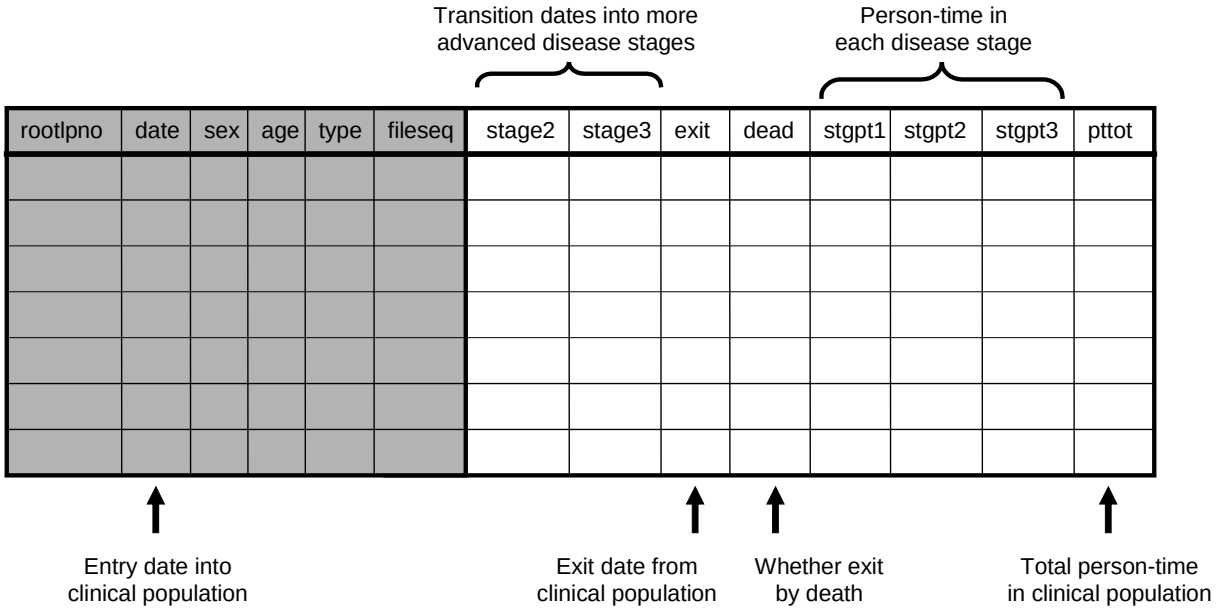
The aim of preparing syntax to solve partitioning problems is to divide each patient's person-time and relevant events into groups, whereby they can be separately enumerated and summed over all of the patients in the data file. The diagram below illustrates this process. The patient's total person-time under observation is split into fragments and the value of each fragment is stored in a repository (actually a variable in the data file). Similarly, the events are divided up and stored in separate repositories/variables.

Partitioning of Individual Patient Time Trace



Partitioning Person-Time Using Sentinel Events

The aim of preparing a platform file for implementation of a sentinel events model is to achieve a result that looks like the following in the case of a clinical population with three disease severity subgroups partitioned by irreversible transition events:



A possible set of steps to achieve this result is as follows:

Step 1: The entry and exit events for the clinical population as a whole, and the irreversible transition events marking the progression of patients between subgroups are clearly defined in terms of the relevant variables, criteria and source files. Typically the entry event will be the date of the inaugural index record, which will already be present in the platform file.

Step 2: One or more source files are searched for transition and exit events and the corresponding dates are internally loaded onto the index records in the source file, which are then saved as a cut down file of relevant index record information. The transition and exit dates are then externally loaded onto the platform file, using the cutdown index source file as a look-up table that is keyed by rootlpno. Often the only relevant exit date information at this stage will be a date of death, which is externally loaded using the source file of deaths directly as a keyed look-up table.

Step 3: For patients with a date of death loaded into exit, a 0-1 binary indicator variable, dead, is set to '1'. For those patients not known to be deceased, exit is censored by assigning the last date of follow-up on which vital status could be reliably ascertained. If outward migration data are available, it may be possible to censor non-deceased patients on known dates of leaving the State. Otherwise, exit for non-deceased patients is assigned the final date of the study observation period.

Step 4: Finally, the person-time spent in each disease severity stage and in the clinical population as a whole are calculated using date functions applied to the dates of boundary events of each stage. An efficient algorithm to achieve this is to start by calculating stgpt3, then calculate (stgpt2 + stgpt3) and subtract stgpt3 to obtain stgpt2. Finally calculate (stgpt1 + stgpt2 + stgpt3), assign this to pttot and subtract (stgpt2 + stgpt3) to obtain stgpt1. The following pseudo-syntax illustrates this approach.

IF stage3 ≠ blank THEN stgpt3 = DAYS(exit – stage3) + 1 ELSE stgpt3 = 0.
IF stage2 ≠ blank THEN stgpt2 = DAYS(exit – stage2) + 1 – stgpt3 ELSE stgpt2 = 0.
Set pttot to DAYS(exit – date) + 1.
Set stgpt1 to pttot – (stgpt2 + stgpt3).

The approach above counts both the dates of entry to and exit from the clinical population as contributing 1 person-day each and, moreover that the higher clinical stage is favoured for the allocation of the 1 person-day from a transition date. Arguably, more accurate approaches are possible, such as an entry date and event-based exit date each contributing only 0.5 person-days (but note that a censored exit date should still contribute 1 whole person-day); and a transition date contributing 0.5 person-days to each of the relevant disease stages.

SPSS

```
COMPUTE stgpt1 = 0.
COMPUTE stgpt2 = 0.
COMPUTE stgpt3 = 0.
IF (stage3 GT 0) stgpt3 = XDATE.TDAY(exit) - XDATE.TDAY(stage3) + 1.
IF (stage2 GT 0) stgpt2 = XDATE.TDAY(exit) - XDATE.TDAY(stage2) + 1 - stgpt3.
COMPUTE totpt = XDATE.TDAY(exit) - XDATE.TDAY(date) + 1.
COMPUTE stgpt1 = totpt - (stgpt2 + stgpt3).
```

SAS

```
stgpt1 = 0;
stgpt2 = 0;
stgpt3 = 0;
if (stage3 > .) then stgpt3 = exit - stage3 + 1;
if (stage2 > .) then stgpt2 = exit - stage2 + 1 - stgpt3;
totpt = exit - date + 1;
stgpt1 = totpt - (stgpt2 + stgpt3);
```

Stata

```
generate stgpt1 = 0
generate stgpt2 = 0
generate stgpt3 = 0
replace stgpt3=exit-stage3+1 if stage3~=.
replace stgpt2=(exit-stage2+1)-stgpt3 if stage2~=.
generate totpt=exit-date+1
replace stgpt1=totpt-(stgpt2 + stgpt3)
```

R

```
stgpt3 <- as.numeric(ifelse(!is.na(stage3),exit - stage3 + 1,0))
stgpt2 <- as.numeric(ifelse(!is.na(stage2),exit - stage2 + 1 - stgpt3,0))
totpt <- as.numeric(exit - date + 1)
stgpt1 <- totpt - (stgpt2 + stgpt3)
```

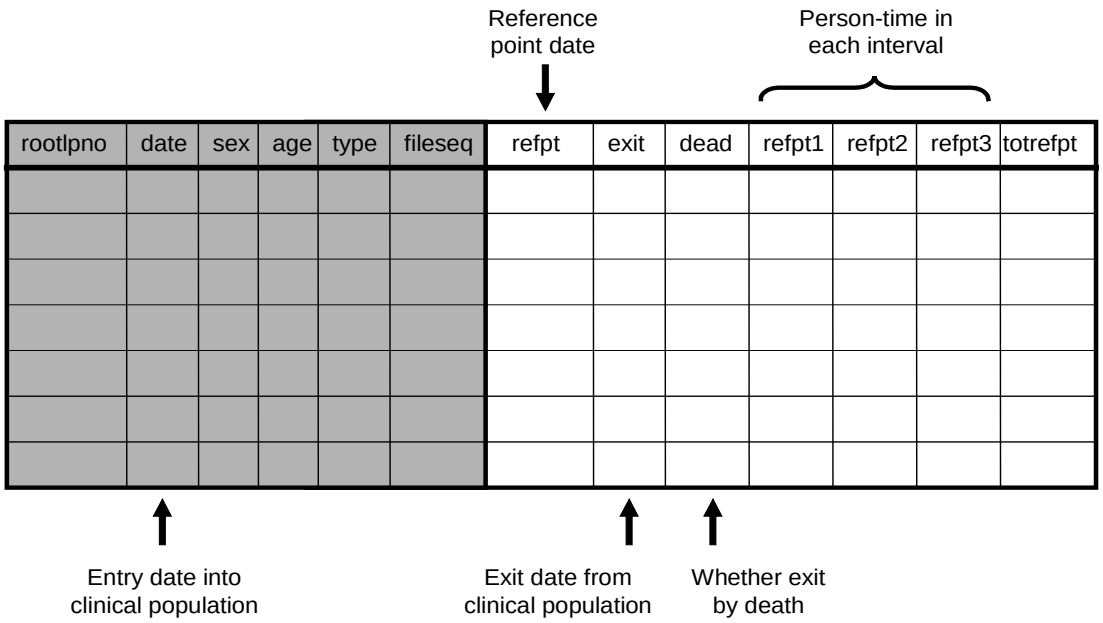
Use of Partitioned Person-Time to Assess Data and Linkage Quality

A valuable spin-off of a person-time analysis partitioned by sentinel events is to identify any person-time estimates that are negative numbers. Negative person-time will be the result of biologically impossible sequences of sentinel events, such as death occurring prior to an entry or transition event; or problems with the way that stages of disease severity have been defined. The former problem will be mostly the result of mismatched deaths. The response of the analyst should be to investigate the

problem(s), going back to sources where necessary, and either (i) correcting the data in the file to the most plausible representation of the true situation; or (ii) adjusting disease severity definitions.

Partitioning Person-Time Using a Reference Point

The aim of preparing a platform file for implementation of a reference point model is to achieve a result that looks like the following in the case of a clinical population with three subgroups based on serial time intervals following the reference point:



A possible set of steps to achieve this result is as follows:

Step 1: The reference point is clearly defined in terms of the relevant variables, criteria and source files. If the reference point is the date of the inaugural index record, it is already present in the platform file and does not require step 2 to create a separate refpt variable.

Step 2: If necessary, one or more source files are searched for events that define the date of the reference point for each patient, which is loaded onto the index records in the source file and saved as a cut-down file of index record information. The reference point data are then externally loaded onto the platform file, using the cutdown index source file as a look-up table that is keyed by rootlpno.

Step 3: If not already in the platform file, the date of death is loaded into exit and the 0-1 binary indicator variable, dead, is set to '1' for those who are deceased. For those patients not known to be deceased, exit is censored by assigning the last date of follow-up on which vital status could be reliably ascertained. If outward migration data are available, it may be possible to censor non-deceased patients on known dates of leaving the State. Otherwise, exit for non-deceased patients is assigned the final date of the study observation period.

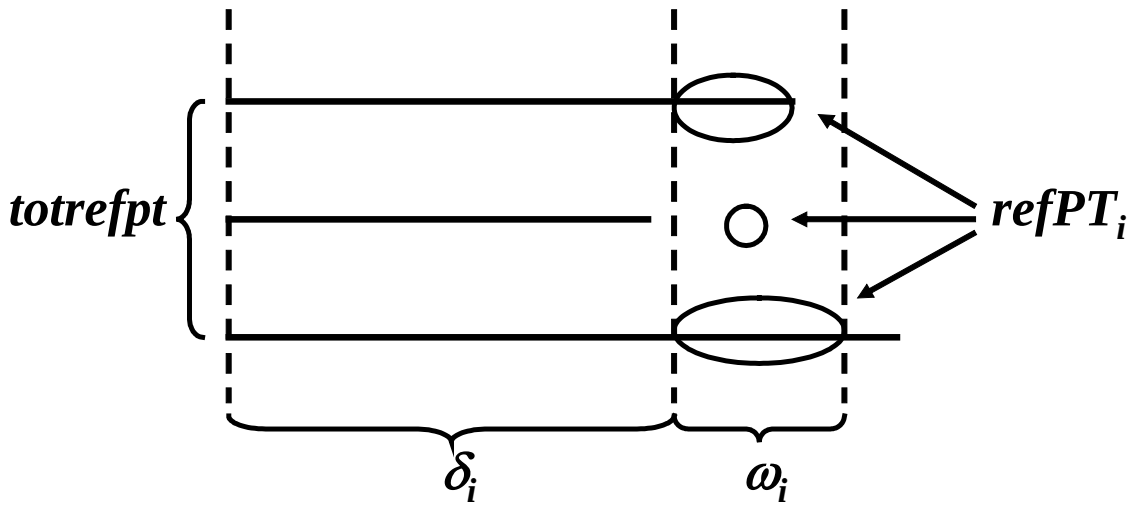
Step 4: Finally, the person-time spent in each interval, anchored by the reference point, is calculated. An efficient algorithm to achieve this is to calculate the total person-time available between the reference point refpt and the exit date and assign this value to totrefpt. The person-time available within reference interval i of time width ω_i is then defined as:

$$refPT_i = totrefpt - \delta_i \quad \text{if} \quad 0 \le (totrefpt - \delta_i) \le \omega_i$$

$$refPT_i = 0 \quad \text{if} \quad (totrefpt - \delta_i) < 0$$

$$\text{and} \quad refPT_i = \omega_i \quad \text{if} \quad (totrefpt - \delta_i) > \omega_i$$

where δ_i is the time between refpt and the commencement of interval i.



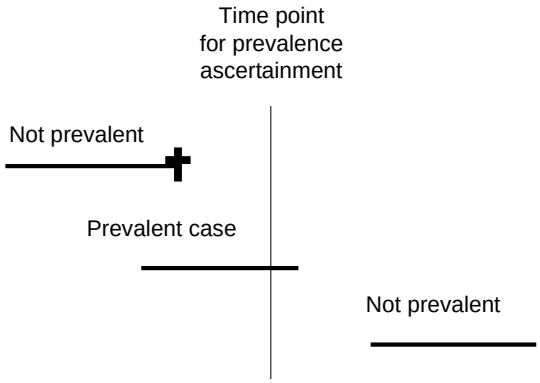
The following pseudo-syntax illustrates this approach where there are three 365-day reference intervals starting at the reference point.

```
Set totrefpt to DAYS(exit – refpt) + 1.
Set refpt1 to totrefpt.
IF refpt1 > 365 THEN = 365.
Set refpt2 to totrefpt – 365.
IF refpt2 < 0 THEN refpt2 = 0 ELSE IF refpt2 > 365 THEN refpt2 = 365.
Set refpt3 to totrefpt – 730.
IF refpt3 < 0 THEN refpt3 = 0 ELSE IF refpt3 > 365 THEN refpt3 = 365.
```

Enumerating Point Prevalent Cases

To determine if a particular patient is a prevalent case at a particular point in time requires two initial specifications: (i) the clinical definition of what prevalent cases are to be enumerated; eg whether restricted to a specified stage of disease severity or at a specified interval after diagnosis; and (ii) the point in time as a date, age or other time-scale value when prevalent cases are to be enumerated.

The first specification defines the relevant segment of person-time for each patient and the second specification defines a time point that must be overlapped by the patient's person-time trace to warrant the patient being counted as a prevalent case. The following diagram illustrates the principle.



Patients that meet the criteria are typically tagged with a 0-1 binary indicator variable, set to '1' to indicate those prevalent at the specified time point. In the following pseudo-syntax, the aim is to tag

patients who are prevalent with stage 3 disease on 30 June 2006, using a platform file that has already been partitioned with sentinel events along the lines explained above:

```
IF stage3 ≠ blank AND stage 3 ≤ DATE(30.6.2006) AND exit ≥ DATE(30.6.2006)
   THEN s3prev06 = 1 ELSE s3prev06 = 0.
```

The following pseudo-syntax could be used when the definition of a prevalent case is a patient in the third year after diagnosis on 30 June 2006, using a platform file that has already been partitioned with a reference point:

```
IF DATE(DAYS(refpt) + 730)) ≤ DATE(30.6.2006)
   AND DATE(DAYS(refpt) + 1094)) ≥ DATE(30.6.2006) AND exit ≥ DATE(30.6.2006)
   THEN r3prev06 = 1 ELSE r3prev2006 = 0.
```

As with person-time estimations (see above), more accurate approaches are possible where an entry, transition or exit date falls exactly on the point prevalence date. For example, a patient transitioning from stage 2 to stage 3 on 30 June 2006 could contribute 0.5 to the prevalent case count for each stage, thus avoiding double counting of the whole case in both stages.

Enumerating Numerators of Rates within Partitioned Person-Time

The method of counting numerators of utilisation or outcome rates based on events occurring within partitioned person-time, whether partitioned using sentinel events or a reference point, bears analogy with the method of enumerating point prevalent cases. Relevant events must be defined, identified and tagged. Each event must have an identifiable date of occurrence. This date is treated like the time point in a point prevalence estimation problem. It is then a matter of determining which of the patient's person-time segments, whether a disease stage segment or interval segment, overlaps the date of the event. Whichever person-time segment overlaps the event determines in which numerator repository/variable the event is counted.

Overview of Exercise 2

Steps 1-3. Identify transition events to stage 2 in PBSdata file and load onto diabetes file.

Steps 4-6. Identify transition events to stage 3 in PBSdata file and load onto diabetes file.

Step 7. Load deaths from Dthdata file onto diabetes file and set exit dates.

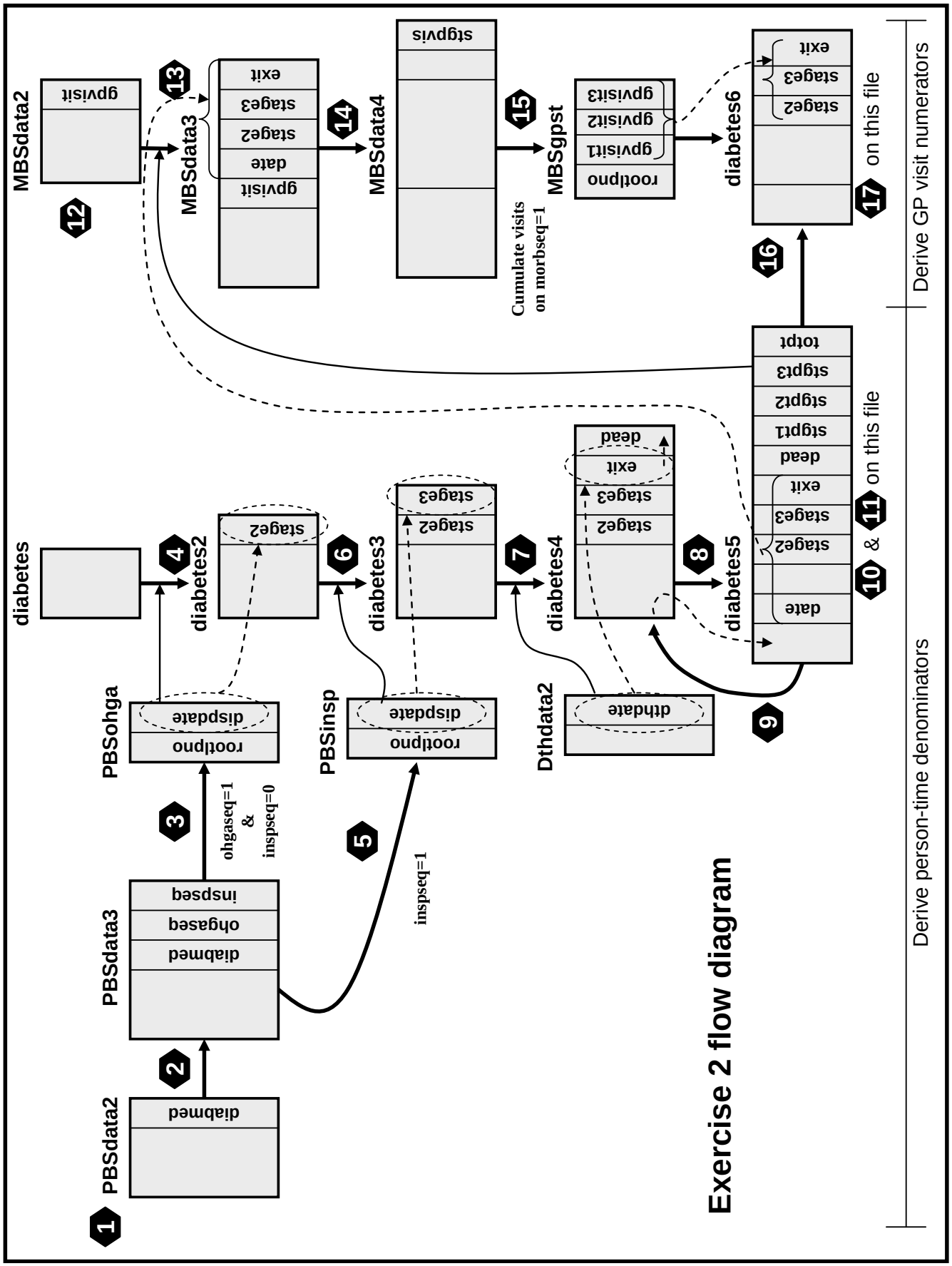
Steps 8-10. Derive person-time partitioned by stage.

Step 11. Estimate point prevalence in 1994 and 1999.

Steps 12-14. Load transition dates from diabetes file onto MBSdata file and identify/classify GP visits.

Steps 15-16. Accumulate GP visits by stage onto index records in MBSdata and load onto diabetes file.

Step 17. Derive GP visit counts & rates partitioned by stage.

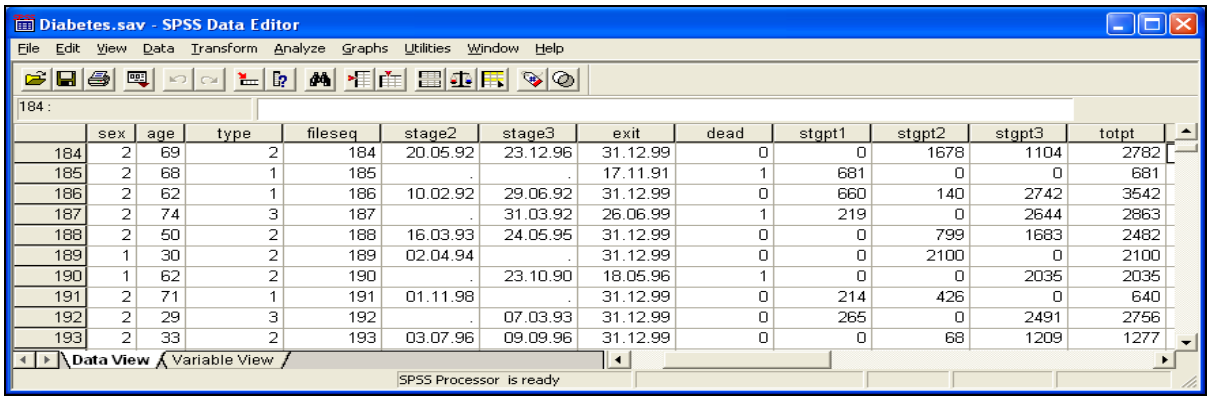

Exercise 2 flow diagram
Derive person-time denominators
Derive GP visit numerators
1
PBSdata2
diabmed
2
PBSdata3
diabmed
ohgaseq
inspseq
3
ohgaseq=1
&
inspseq=0
PBSohga
rootlpno
dispdate
4
diabetes
diabetes2
stage2
5
inspseq=1
PBSinsp
rootlpno
dispdate
6
diabetes3
stage2
stage3
7
Dthdata2
dthdate
diabetes4
stage2
stage3
exit
dead
8
diabetes5
9
10 & 11 on this file
date
stage2
stage3
exit
dead
stgpt1
stgpt2
stgpt3
totpt
12
MBSdata2
gpvisit
13
MBSdata3
gpvisit
date
stage2
stage3
exit
14
MBSdata4
stgpvis
15
Cumulate visits
on morbseq=1
MBSgpst
rootlpno
gpvisit1
gpvisit2
gpvisit3
diabetes6
stage2
stage3
exit
16
17 on this file

Exercise Instructions

Partitioning Person-Time by Disease Severity Using Sentinel Events

1. Open your saved, tagged and sequenced PBSdata2 source file. You will recall that you had tagged the scripts for oral hypoglycaemic agents as diabmed = '1' and insulin preparations as diabmed = '2'. We will use the dispensing date of the first insulin script as the sentinel event marking entry into 'stage 3' diabetes. If there is an earlier dispensing date of a first script for an oral hypoglycaemic agent, we will use it as the sentinel event marking entry into 'stage 2' diabetes. A patient who has an inaugural index date that precedes either of these dates will commence their clinical diabetic history in 'stage 1'.
2. Create two new index sequence variables, ohgaseq using the first script for an oral hypoglycaemic agent as the index record, and inspseq using the first script for an insulin preparation as the index record. Write your syntax so that only records in a block prior to the index record are numbered with a '0'. Save your source file with these changes as PBSdata3.
3. Now cut down the file to contain only the rootlpno and dispdate for records where (ohgaseq = 1 and inspseq = 0) and save this cut down file as PBSohga. Can you see why it was important to include the additional selection criterion of inspseq = 0?
4. Open your diabetes platform file and use a merge procedure to externally load dispdate using PBSohga as a look-up table keyed by rootlpno. Change the name of dispdate in the platform file to stage2. Save your platform file in this form as diabetes2.
5. Open the PBSdata3 file again and this time cut it down to contain only rootlpno and dispdate for records where inspseq = 1 and save this cut down file as PBSinsp.
6. Open your diabetes2 platform file again and use a merge procedure to externally load dispdate using PBSinsp as a look-up table keyed by rootlpno. This time change the name of dispdate in the platform file to stage3. Save your platform file in this form as diabetes3.
7. Now externally load the dates of any deaths using the source file Dthdata2 as a look-up table keyed by rootlpno. Rename dthdate to exit and create a new binary indicator variable, dead, which is set to '1' when a date of death is present and otherwise set to '0'. For records where dead is set to '0', or for those whose death occurred after 31 December 1999, you should change to date value of exit to 31 December 1999, being the last date of the study observation period and hence the censoring date. **Stata users:** Specify the nokeep option of merge to prevent Dthdata2 cases not found in the diabetes3 file from being included. The Stata merge will import rootlpno from death records that are not in the diabetes3 file (51 of them) if this option is not specified. Save your platform file in this form as diabetes4.
8. Use the methods explained in the briefing to this training session to calculate the lengths of person-time that each patient spent in each of the three stages of diabetes, as well as the total person-time that each patient contributed to the clinical population during all stages combined. Call the stage-specific person-time variables stgpt1, stgpt2 and stgpt3, and the total person-time variable totpt. If your syntax performed correctly, your platform file should look like this:

SPSS



	sex	age	type	fileseq	stage2	stage3	exit	dead	stgpt1	stgpt2	stgpt3	totpt
184	2	69	2	184	20.05.92	23.12.96	31.12.99	0	0	1678	1104	2782
185	2	68	1	185	.	.	17.11.91	1	681	0	0	681
186	2	62	1	186	10.02.92	29.06.92	31.12.99	0	660	140	2742	3542
187	2	74	3	187	.	31.03.92	26.06.99	1	219	0	2644	2863
188	2	50	2	188	16.03.93	24.05.95	31.12.99	0	0	799	1683	2482
189	1	30	2	189	02.04.94	.	31.12.99	0	0	2100	0	2100
190	1	62	2	190	.	23.10.90	18.05.96	1	0	0	2035	2035
191	2	71	1	191	01.11.98	.	31.12.99	0	214	426	0	640
192	2	29	3	192	.	07.03.93	31.12.99	0	265	0	2491	2756
193	2	33	2	193	03.07.96	09.09.96	31.12.99	0	0	68	1209	1277

SAS

VIEWTABLE: Aalhd.Diabetes

	Patient gender	Patient age in years	date	type	fileseq	stage2	stage3	Date of Exit from Clinical Population	Death Flag	stgpt1	stgpt2	stgpt3	totpt
184	2	69	20/05/1992	2	184	20/05/1992	23/12/1996	31/12/1999	0	0	1678	1104	2782
185	2	68	06/01/1990	1	185	.	.	17/11/1991	1	681	0	0	681
186	2	62	21/04/1990	1	186	10/02/1992	29/06/1992	31/12/1999	0	660	140	2742	3542
187	2	74	25/08/1991	3	187	.	31/03/1992	26/06/1999	1	219	0	2644	2863
188	2	50	16/03/1993	2	188	16/03/1993	24/05/1995	31/12/1999	0	0	799	1683	2482
189	1	30	02/04/1994	2	189	02/04/1994	.	31/12/1999	0	0	2100	0	2100
190	1	62	23/10/1990	2	190	.	23/10/1990	18/05/1996	1	0	0	2035	2035
191	2	71	01/04/1998	1	191	01/11/1998	.	31/12/1999	0	214	426	0	640
192	2	29	15/06/1992	3	192	.	07/03/1993	31/12/1999	0	265	0	2491	2756
193	2	33	03/07/1996	2	193	03/07/1996	09/09/1996	31/12/1999	0	0	68	1209	1277
194	1	73	13/04/1990	1	194	17/01/1993	.	31/12/1999	0	1010	2540	0	3550
195	1	63	11/02/1992	2	195	.	11/02/1992	31/12/1999	0	0	0	2881	2881

Stata

Data Browser

Preserve | Restore | Sort | << | >> | Hide | Delete...

sex[184] = 2

	sex	age	sequence	type	morbseq	fileseq	stage2	stage3	exit	dead	stgpt1	stgpt2	stgpt3	totpt
184	female	69	1	2	1	184	20 May 92	23 Dec 96	31 Dec 99	Alive	0	1678	1104	2782
185	female	68	1	1	1	185	.	.	17 Nov 91	Dead	681	0	0	681
186	female	62	1	1	1	186	10 Feb 92	29 Jun 92	31 Dec 99	Alive	660	140	2742	3542
187	female	74	1	3	1	187	.	31 Mar 92	26 Jun 99	Dead	219	0	2644	2863
188	female	50	1	2	1	188	16 Mar 93	24 May 95	31 Dec 99	Alive	0	799	1683	2482
189	male	30	1	2	1	189	02 Apr 94	.	31 Dec 99	Alive	0	2100	0	2100
190	male	62	1	2	1	190	.	23 Oct 90	18 May 96	Dead	0	0	2035	2035
191	female	71	1	1	1	191	01 Nov 98	.	31 Dec 99	Alive	214	426	0	640
192	female	29	1	3	1	192	.	07 Mar 93	31 Dec 99	Alive	265	0	2491	2756
193	female	33	1	2	1	193	03 Jul 96	09 Sep 96	31 Dec 99	Alive	0	68	1209	1277

9. This is a good time to check the file for any negative person-time values. You will find that there are 12 out of 10,675 cases that have problems in this regard, all due to an impossible date of death (NB: look at all pt variables and not just totpt). In the real-life situation, these problems would need to be investigated. But for the purpose of this training exercise you should handle them by assuming that the deaths are mismatches. In each instance, change the value of dead to '0' and the date value of exit to 31 December 1999. You should now rerun your syntax from question 8 and save your updated platform file as diabetes5.

10. A convenient way of summing stage-specific and total person-times over the entire clinical population is to use a descriptive statistics procedure in your statistical software package to provide the sum for each person-time variable. The means are also of interest, being the average observed durations of each stage and the overall average follow-up time from entry to exit from the clinical population. Minimum and maximum values are also useful and can serve as a quality control function. For example, no person-time should exceed the study observation period and every patient should have at least 1 person-day of follow-up. Check that you obtained the results shown opposite and then save your platform file. The clinical population contains 17.8 million person-days of observation, 42% in stage 1, 40% in stage 2 and 18% in stage 3.

SPSS

Output7 - SPSS Viewer

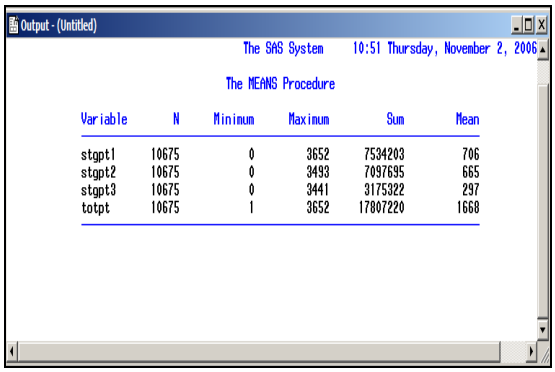
File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean
stgpt1	10675	0	3652	7534203	705.78
stgpt2	10675	0	3493	7097695	664.89
stgpt3	10675	0	3441	3175322	297.45
totpt	10675	1	3652	17807220	1668.12
Valid N (listwise)	10675				

SPSS Processor is ready

SAS



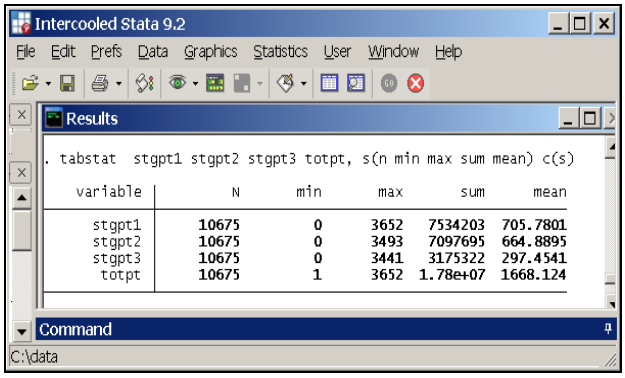
Output - (Untitled)

The SAS System 10:51 Thursday, November 2, 2006

The MEANS Procedure

Variable	N	Minimum	Maximum	Sum	Mean
stgpt1	10675	0	3652	7534203	706
stgpt2	10675	0	3493	7097695	665
stgpt3	10675	0	3441	3175322	297
totpt	10675	1	3652	17807220	1668

Stata



Intercooled Stata 9.2

File Edit Prefs Data Graphics Statistics User Window Help

Results

```
. tabstat  stgpt1 stgpt2 stgpt3 totpt, s(n min max sum mean) c(s)
```

variable	N	min	max	sum	mean
stgpt1	10675	0	3652	7534203	705.7801
stgpt2	10675	0	3493	7097695	664.8895
stgpt3	10675	0	3441	3175322	297.4541
totpt	10675	1	3652	1.78e+07	1668.124

Command

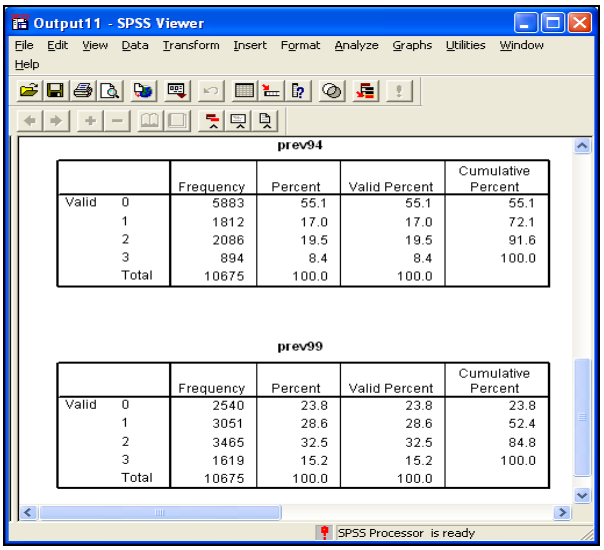
C:\data

Estimating Point Prevalence

11. Your next objective is to enumerate the prevalent cases of diabetes in each of the three stages on 30 June 1994 and 30 June 1999. Create two new variables in the platform file, prev94 and prev99, which should each take on the values of '0' (not prevalent), '1' (prevalent in stage 1), '2' (prevalent in stage 2) or '3' (prevalent in stage 3) for the two time points in question.

    This task goes beyond the method explained in the briefing session, but only in as much as the results of the three enumerations of prevalent cases at each time point are combined in the same prevalence indicator variables. Run frequency distributions to obtain the counts of prevalent cases in each stage at each time point and check that your results are shown opposite. Assume that your linked data cover an entire geographically defined population with census estimates of 115,200 in 1994 and 128,571 in 1999. What are the prevalence estimates for total diabetes? In which severity groups have there been the largest relative and absolute increases? **Hint:** See your syntax solutions book for the results.

SPSS



Output11 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

prev94

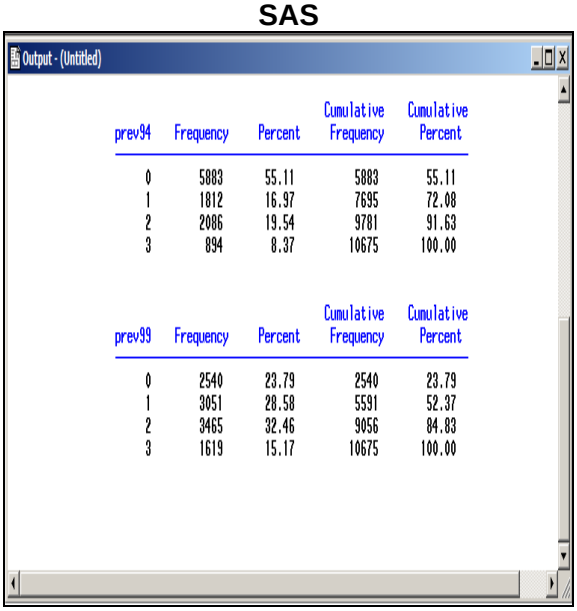
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	5883	55.1	55.1	55.1
	1	1812	17.0	17.0	72.1
	2	2086	19.5	19.5	91.6
	3	894	8.4	8.4	100.0
	Total	10675	100.0	100.0	

prev99

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	2540	23.8	23.8	23.8
	1	3051	28.6	28.6	52.4
	2	3465	32.5	32.5	84.8
	3	1619	15.2	15.2	100.0
	Total	10675	100.0	100.0	

SPSS Processor is ready

SAS

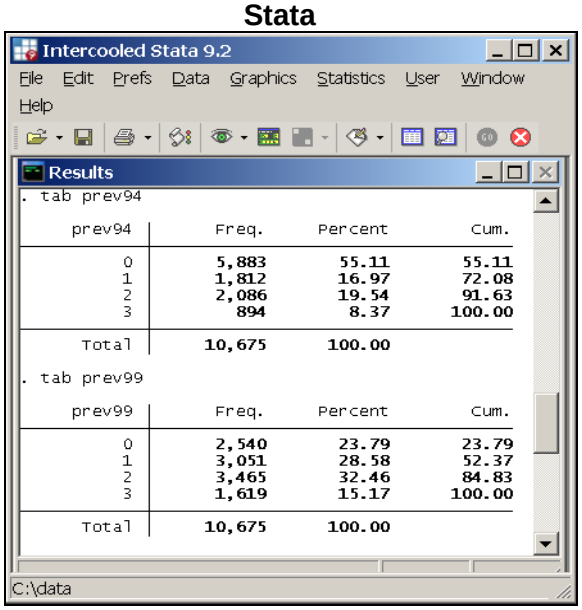


Output - (Untitled)

prev94	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5883	55.11	5883	55.11
1	1812	16.97	7695	72.08
2	2086	19.54	9781	91.63
3	894	8.37	10675	100.00

prev99	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2540	23.79	2540	23.79
1	3051	28.58	5591	52.37
2	3465	32.46	9056	84.83
3	1619	15.17	10675	100.00

Stata



Intercooled Stata 9.2

File Edit Prefs Data Graphics Statistics User Window Help

Results

```
. tab prev94
```

prev94	Freq.	Percent	Cum.
0	5,883	55.11	55.11
1	1,812	16.97	72.08
2	2,086	19.54	91.63
3	894	8.37	100.00
Total	10,675	100.00	

```
. tab prev99
```

prev99	Freq.	Percent	Cum.
0	2,540	23.79	23.79
1	3,051	28.58	52.37
2	3,465	32.46	84.83
3	1,619	15.17	100.00
Total	10,675	100.00	

C:\data

Utilisation Rate of Primary Medical Care by Disease Severity

12. Open your MBSdata2 source file and tag all records that mention a MBS item that represents a GP visit by creating a new variable, gpvisit. Set gpvisit to '1' for a tagged record and '0' otherwise. Run a frequency distribution to confirm that you have tagged 745,336 visits.

13. Use a merge procedure to externally load the following variables from the platform file, using the latter as a look-up table keyed by rootlpno: date, stage2, stage3 and exit. Save as MBSdata3.

14. Each tagged GP visit should now be classified according to whether it occurred during disease stages 1, 2 or 3. Technically, this requires essentially the same logic that you applied in writing you syntax for question 11. Create a new variable, stgpvis, which should take on the values of '0' (not diabetic at that time), '1' (visit in stage 1), '2' (visit in stage 2) or '3' (visit in stage 3). Run a frequency distribution to confirm that you have identified 151,196 GP visits in stage 1, 177,935 in stage 2 and 76,852 in stage 3, while 339,353 visits occurred before the patient became diabetic. Save as MBSdata4.

15. Create three new variables, gpvisit1, gpvisit2 and gpvisit3 and internally load the sums of the GP visits in the three disease stages into the appropriate variable on the first record in each block.

    HINT for SPSS users: It is best to turn the file upside-down by sorting in descending order of fileseq and to accumulate running tallies of the numbers of GP visits of each type in the three new variables. Remember to turn your file the right way up upon completing the task.

    HINT for SAS users: Use the retain statement to set up appropriate counters.

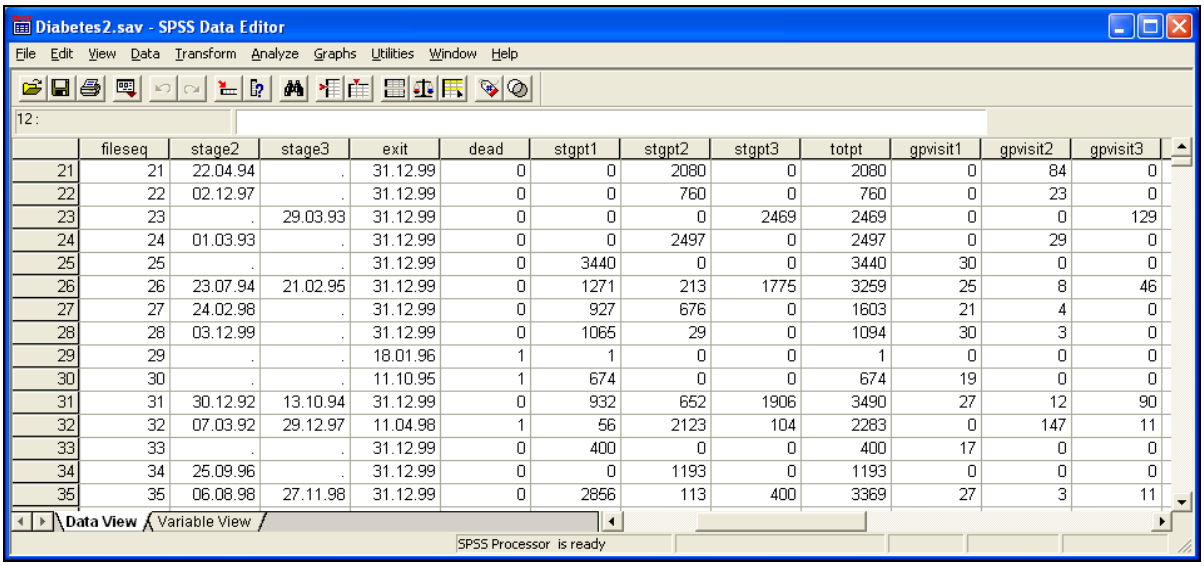
    HINT for Stata users: One way is to generate three temporary variables containing the sum of gpvisit by rootlpno and conditioned by stgpvis and then use egen to create gpvisit1, gpvisit2 and gpvisit3 as the maximum of the temporary variables by rootlpno.

    HINT for R users: One way is to turn the file upside-down, as for SPSS, and then create the three variables separately using the function ave() in conjunction with the function cumsum() each time.

    Cut down the file to contain only the rootlpno, gpvisit1, gpvisit2 and gpvisit3 for records where morbseq = 1 and save this cut down file as MBSgpst. Can you think why you should select on morbseq = 1 (ie, first record in each block) rather than on servseq =1 (ie, index record).

16. Open your diabetes5 platform file again and use a merge procedure to externally load gpvisit1, gpvisit2 and gpvisit3, using MBSgpst as a look-up table keyed by rootlpno. There are 47 cases with missing values for gpvisit1, gpvisit2 and gpvisit3; these should be recoded to zero. Check that your platform file looks like the screen below and save it in this form as diabetes6.

SPSS



	fileseq	stage2	stage3	exit	dead	stgpt1	stgpt2	stgpt3	totpt	gpvisit1	gpvisit2	gpvisit3
21	21	22.04.94	.	31.12.99	0	0	2080	0	2080	0	84	0
22	22	02.12.97	.	31.12.99	0	0	760	0	760	0	23	0
23	23	.	29.03.93	31.12.99	0	0	0	2469	2469	0	0	129
24	24	01.03.93	.	31.12.99	0	0	2497	0	2497	0	29	0
25	25	.	.	31.12.99	0	3440	0	0	3440	30	0	0
26	26	23.07.94	21.02.95	31.12.99	0	1271	213	1775	3259	25	8	46
27	27	24.02.98	.	31.12.99	0	927	676	0	1603	21	4	0
28	28	03.12.99	.	31.12.99	0	1065	29	0	1094	30	3	0
29	29	.	.	18.01.96	1	1	0	0	1	0	0	0
30	30	.	.	11.10.95	1	674	0	0	674	19	0	0
31	31	30.12.92	13.10.94	31.12.99	0	932	652	1906	3490	27	12	90
32	32	07.03.92	29.12.97	11.04.98	1	56	2123	104	2283	0	147	11
33	33	.	.	31.12.99	0	400	0	0	400	17	0	0
34	34	25.09.96	.	31.12.99	0	0	1193	0	1193	0	0	0
35	35	06.08.98	27.11.98	31.12.99	0	2856	113	400	3369	27	3	11

SAS

VIEWTABLE: Aalhd.Diabetes

	fileseq	stage2	stage3	Date of Exit from Clinical Population	Death Flag	stgpt1	stgpt2	stgpt3	totpt	prev94	prev99	gpvisit1	gpvisit2	gpvisit3
21	21	22/04/1994	.	31/12/1999	0	0	2080	0	2080	2	2	0	84	0
22	22	02/12/1997	.	31/12/1999	0	0	760	0	760	0	2	0	23	0
23	23	.	29/03/1993	31/12/1999	0	0	0	2469	2469	3	3	0	0	129
24	24	01/03/1993	.	31/12/1999	0	0	2497	0	2497	2	2	0	29	0
25	25	.	.	31/12/1999	0	3440	0	0	3440	1	1	30	0	0
26	26	23/07/1994	21/02/1995	31/12/1999	0	1271	213	1775	3259	1	3	25	8	46
27	27	24/02/1998	.	31/12/1999	0	927	676	0	1603	0	2	21	4	0
28	28	03/12/1999	.	31/12/1999	0	1065	29	0	1094	0	1	30	3	0
29	29	.	.	18/01/1996	1	1	0	0	1	0	0	0	0	0
30	30	.	.	11/10/1995	1	674	0	0	674	1	0	19	0	0
31	31	30/12/1992	13/10/1994	31/12/1999	0	932	652	1906	3490	2	3	27	12	90
32	32	07/03/1992	29/12/1997	11/04/1998	1	56	2123	104	2283	2	0	0	147	11
33	33	.	.	31/12/1999	0	400	0	0	400	0	1	17	0	0
34	34	25/09/1996	.	31/12/1999	0	0	1193	0	1193	0	2	0	0	0
35	35	06/08/1998	27/11/1998	31/12/1999	0	2856	113	400	3369	1	3	27	3	11
36	36	07/03/1992	.	31/12/1999	0	0	2856	0	2856	2	2	0	74	0

Stata

Data Browser

Preserve | Restore | Sort | << | >> | Hide | Delete...

stage3[21] = .

	fileseq	stage2	stage3	exit	dead	stgpt1	stgpt2	stgpt3	totpt	gpvisit1	gpvisit2	gpvisit3
21	21	22 Apr 94	.	31 Dec 99	Alive	0	2080	0	2080	0	84	0
22	22	02 Dec 97	.	31 Dec 99	Alive	0	760	0	760	0	23	0
23	23	.	29 Mar 93	31 Dec 99	Alive	0	0	2469	2469	0	0	129
24	24	01 Mar 93	.	31 Dec 99	Alive	0	2497	0	2497	0	29	0
25	25	.	.	31 Dec 99	Alive	3440	0	0	3440	30	0	0
26	26	23 Jul 94	21 Feb 95	31 Dec 99	Alive	1271	213	1775	3259	25	8	46
27	27	24 Feb 98	.	31 Dec 99	Alive	927	676	0	1603	21	4	0
28	28	03 Dec 99	.	31 Dec 99	Alive	1065	29	0	1094	30	3	0
29	29	.	.	18 Jan 96	Dead	1	0	0	1	0	0	0
30	30	.	.	11 Oct 95	Dead	674	0	0	674	19	0	0
31	31	30 Dec 92	13 Oct 94	31 Dec 99	Alive	932	652	1906	3490	27	12	90
32	32	07 Mar 92	29 Dec 97	11 Apr 98	Dead	56	2123	104	2283	0	147	11
33	33	.	.	31 Dec 99	Alive	400	0	0	400	17	0	0
34	34	25 Sep 96	.	31 Dec 99	Alive	0	1193	0	1193	0	0	0
35	35	06 Aug 98	27 Nov 98	31 Dec 99	Alive	2856	113	400	3369	27	3	11

17. Sum the stage-specific GP visits using the descriptive statistics procedure in your statistical software, requesting also the means, minima and maxima. You should obtain the results shown opposite. Now divide each stage-specific numerator by its corresponding stage-specific person-time denominator to measure the utilisation rate of primary medical care by severity of diabetes. What are the rates per person-year? **Hint:** See your syntax solutions book for the results.

SPSS

Output2 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean
gpvisit1	10675	0	479	151196	14.16
gpvisit2	10675	0	336	177935	16.67
gpvisit3	10675	0	414	76852	7.20
Valid N (listwise)	10675				

SPSS Processor is ready

SAS

Output - (Untitled)

The SAS System 10:51 Thursday, November

The MEANS Procedure

Variable	N	Minimum	Maximum	Sum	Mean
gpvisit1	10675	0	479	151196	14
gpvisit2	10675	0	336	177935	17
gpvisit3	10675	0	414	76852	7

Stata

Intercooled Stata 9.2

File Edit Prefs Data Graphics Statistics User Window Help

Results

```
. tabstat  gpvisit*, s(n min max sum mean) c(s)
```

variable	N	min	max	sum	mean
gpvisit1	10675	0	479	151196	14.16356
gpvisit2	10675	0	336	177935	16.66838
gpvisit3	10675	0	414	76852	7.199251

Command

C:\data

REVIEW 2:
'CLINICAL POPULATION DYNAMICS USING LINKED DATA'

Please check that you are familiar with the following topics:

- One-stage and two-stage models of disease prevalence ☐
- Continuous disease severity models ☐
- Clinical populations ☐
- Sentinel event and reference point models ☐
- Active disease model ☐

Please check that you understand the meaning of the following terms:

- Dynamic equilibrium in a dynamic population ☐
- Sentinel events ☐
- Entry, transition and exit events ☐
- Partitioning person-time ☐
- Reference point ☐
- Couplet and intra-couplet interval ☐
- Concordant and mixed couplets ☐

Please check that you have demonstrated that you have acquired the following skills:

- Partition person-time ☐
- Enumerate point prevalent cases ☐
- Enumerate numerators of rates within partitioned person-time ☐

3. ADVANCED TOPICS IN LONGITUDINAL RESEARCH DESIGN

3.1 Optimising Validity in Epidemiologic Research

The validity or truth of the results of epidemiologic research is often separated into two separate considerations:

- ***Internal validity*** is a lack of ***systematic error*** or ***bias*** in empirically observed measures when used as estimates of corresponding theoretical measures in the population about which direct inferences are drawn from the study (the study ***domain***).
- ***External validity*** or ***scientific generalisability*** is a lack of effect modification in theoretical measures between the study domain and populations beyond the domain and about which indirect inferences or theoretical principles of a more universal nature are drawn from the study.

Neither of these is concerned with ***precision***, which is the lack of ***random error*** due to chance variation in empirically observed measures. As a general rule, the research designer will place a premium on the optimisation of internal validity, if necessary, at the expense of trade-offs in precision, external validity or both. The reason is that precision is measurable (with a confidence interval) and the challenges of external validity can be considered explicitly based on criteria of biologic plausibility and the degree of difference between the domain and target population for generalisation. The level of error arising from bias, however, is difficult to quantify and its implications for the interpretation of a study will often remain a matter for conjecture.

Example: Restriction in non-experimental research design may be used to improve internal validity at the expense of both precision and external validity. A study of the effects of alcohol consumption on the incidence rate of lung cancer could be undertaken restrictively in never-smokers. This restriction would eliminate confounding by tobacco smoking, but at the cost of a smaller sample size available for analysis (loss of precision) and greater uncertainty as to whether the results might apply also to smokers (loss of scientific generalisability).

In arriving at longitudinal designs for the analysis of linked health data, the analyst should be familiar with the common sources of bias that would diminish the internal validity of results and the means to reduce these sources of bias, wherever possible, through intelligent study design. Some of the most important considerations relate to how group contrasts are formed, how exposures are measured and how outcomes are ascertained.

3.2 Overview of Types of Bias

Classification of Bias

Biases may be classified according to the direction of the error they produce or the stages of the research process where they occur.

Direction of bias: ***Bias towards the null*** occurs when the systematic error results in an empirical effect measure that is less extreme than the true value. ***Bias away from the null*** is the opposite. Another classification based on direction is to distinguish ***negative bias*** from ***positive bias***. There is confusion as to whether these terms mean the same as the null-referenced classification when the primary effect under study is purportedly protective. One approach is to regard a positive bias as one that increases the value of the empirical effect measure regardless of whether the apparent effect is protective or causal.

Stage of research process: There are several classifications (Sackett, Choi, Maclure and Schneeweiss, Kleinbaum *et al*, Steineck and Ahbolm), but in regard to non-experimental research all are variations of the following trilogy:

- ***Selection bias*** arising from definition and selection of the study population.
- ***Information bias*** arising from data collection.
- ***Confounding*** arising from the association between different determinants in the population.

Some commentators regard confounding as a source of error separate from bias in that it is not an artefact of study design, but a pre-existing state of affairs within the actual study base. The classification used here is largely based on the paper by Delgado-Rodriguez and Llorca, which includes a number of additional and unique forms of bias that arise in the context of the execution of an intervention in experimental research. Types of bias frequently encountered in research and their relevance to linked administrative health data analysis are overviewed in the following tables. Types of bias of particular relevance to this course are indicated by shading.

Selection Bias

This is a distortion in the estimate of effect resulting from the manner in which subjects are selected for the study population.

Type of Selection Bias	Nature of Selection Bias	Relevance to, and Control in, Data Linkage Studies
Sampling frame and ascertainment biases		
Coverage bias	A primary study base sampling frame for the non-exposed is incomplete relative to the ideal; eg, the electoral roll.	Ensure that all comparisons are undertaken within a complete primary study base sampling frame.
Health care access bias	Patients admitted to a particular institution are unrepresentative due to clinician interests (popularity bias), institutional reputation (centripetal bias), case complexity (referral filter bias) or sociodemographic or locational reasons (diagnostic/treatment access bias).	Use linked data from all health care institutions serving the study base.
Length-bias sampling	Prevalent patients include more low severity disease than incident cases due to different average durations of disease.	Use first-time patients and a clearance period.
Detection bias in case-control studies	Cases of disease are more likely to be detected in persons who are exposed due to anticipation of disease.	Restrict the domain of the study to a primary study base with more homogenous case detection efforts.
Neyman bias (survival bias)	Prevalent cases have increased exposure to putative causal factors that are also prognostic.	Use first-time patients and a clearance period.
Spectrum bias	Only definite disease presentations are included; or only patients with no comorbidity (purity diagnostic bias).	Avoid spectrum-restricted index case definitions.
Survivor treatment bias (time-dependent bias; immortal time bias)	Patients who survive longer have more opportunity to be exposed to treatment.	Use exposure opportunity-equivalent longitudinal designs; or time-dependent covariates in Cox regression or a nested case-control study with incidence-density sampling; or multi-stage modelling.
Control selection biases in case-control studies		
Berkson's bias	Occurs in hospital-based case-control studies where hospitalisation differs between cases and controls and is influenced by exposure.	Avoid disease association studies based on coincident mentions of diseases on the same records.
Exclusion bias	Exclusion of controls with conditions related to exposure not implemented in cases.	Avoid differential exclusions across comparison groups.
Inclusion bias	Conditions defining controls in a hospital-based case-control study are related to exposure.	Careful selection of control conditions.
Matching (incl. friend control bias and relative control bias)	In case-control studies, matching introduces a selection bias.	Perform matched analysis (eg, conditional logistic regression)

Type of Selection Bias	Nature of Selection Bias	Relevance to and Control in Data Linkage Studies
Selection biases arising during study implementation		
Competing risks	The outcome competes with other possible events that are related to exposure; eg, death from a specific condition.	Use longitudinal designs with survival analysis methods to censor patients lost to follow-up from competing events.
Loss from follow-up	Exclusion due to loss from follow-up or withdrawal is related to either exposure or outcome.	Improve data linkage tracing systems. Restrict study population to a low migration group. Perform sensitivity analysis.
Missing information	Exclusion due to incomplete information on exposure where the likelihood of complete information relates to the outcome or vice versa.	Improve data linkage tracing systems. Restrict study population to a low migration group. Perform sensitivity analysis.
Response bias (incl. healthy volunteer effect)	Persons volunteering to be exposed have an *a priori* different risk of outcome.	Unusual in data linkage studies. Use methods analogous to intention to treat or instrumental variables.
Biases in reporting and systematic reviews		
Post hoc analysis	A global hypothesis is evaluated using several measures or methods and only the evidence of association is reported.	Avoid multiple tests for a global hypothesis.
Publication bias (incl. language bias)	Results providing evidence of association are more likely to be published and more likely published in English.	Detect using a funnel plot.
Citation bias	Highly cited articles (usually with associations) are more easily found and included in systematic reviews.	Detect using a funnel plot.

Information Bias

This refers to flaws in measuring exposures or outcomes.

Type of Information Bias	Nature of Information Bias	Relevance to and Control in Data Linkage Studies
Misclassification bias		
Differential misclassification bias	Misclassification of exposed/non-exposed or diseased/non-diseased is different between groups.	A ubiquitous problem. Bias may be either direction. Use more accurate measurements.
Non-differential misclassification bias	Misclassification is the same across groups	A ubiquitous problem. Bias towards the null for binary data. Use more accurate measurements.
Detection bias in longitudinal studies	Exposed persons are more likely to have the outcome detected due to anticipation of disease.	Use a non-exposed group with a similar propensity to experience case detection efforts.
Observer bias	Knowledge of outcome status influences the classification of exposure or vice versa.	Unusual in data linkage studies. Blind the observer as to outcome or exposure status.
Recall bias	Presence of the outcome influences perception of its cause (rumination bias), search for exposure (exposure suspicion bias); exposure may influence perception of outcome (participant expectation bias).	Unusual in data linkage studies. Blind the participant as to the study hypothesis.
Reporting bias	Participants give answers to please the researcher (obsequiescence bias); existence of a case triggers family information (family aggregation bias); sensitive information is withheld (underreporting bias); mode for mean bias for socially undesirable behaviours.	Unusual in data linkage studies. Blind the participant as to the study hypothesis.
Regression to the mean		
Regression dilution bias	A variable with an extreme value on its first assessment will tend toward the mean on a later assessment.	Relevant if 'high-risk' individuals are followed after an intervention to reduce risk. Requires correction.

Type of Information Bias	Nature of Information Bias	Relevance to and Control in Data Linkage Studies
Other information biases		
Hawthorne effect	Change in outcome caused by knowledge of being observed.	Unusual in data linkage studies. Caution in publicising prospective non-experimental research touching on provider or consumer behaviour.
Lead time bias	Follow-up of two groups does not commence at comparable times.	Ensure equivalent time-zero for follow-up.
Protopathic bias (incl. sick quitter bias and temporal ambiguity bias)	Exposure is influenced by early or subclinical stages of a disease outcome; eg, increased exposure to a medication or service; reduced exposure to a risk factor.	Use a delay in outcome ascertainment after completion of exposure measurement.
Verification bias	Propensity to execute a gold standard test is influenced by the results of a preliminary test.	Perform gold standard test to verify all preliminary test results.
Will Rogers phenomenon	Improvements in diagnostic tests cause an apparent shift in staging to more advanced disease.	Standardise staging methods across groups.

Confounding

Confounding is mixing of effects. It occurs when a risk factor for an effect in the non-exposed is associated with the exposure of interest in the population from which the effect derives, without being affected by the exposure or the disease (in particular, without being an intermediate step in the causal pathway between exposure and effect).

Type of Confounding	Nature of Confounding	Relevance to and Control in Data Linkage Studies
Susceptibility bias (classical 'confounding')	People who are particularly susceptible to the outcome are also prone to be exposed or not exposed for other reasons.	A ubiquitous problem. Restriction, stratification, matching in cohort studies. Risk adjustment.
Confounding by group (ecological fallacy)	The association observed at the group level is not the same as observed between variables at the individual level because exposure of each community is correlated with outcome risk in the non-exposed.	Unusual in data linkage studies. Use individual-level research designs.
Confounding by indication	An intervention is indicated by factor(s) associated with risk of the outcome.	Common in data linkage studies. Restriction, stratification, matching in cohort studies. Risk adjustment.
Healthy worker effect	Those who are employed remain healthier than those who leave employment.	Of relevance to occupational health data linkage studies. Avoid comparison groups that include the unemployed.

Specific Biases in the Execution of Interventions

These additional sources of bias arise in experimental studies.

Type of Bias	Nature of Bias	Relevance to and Control in Data Linkage Studies
Allocation of intervention bias	The intervention is differentially assigned to the population as may occur in a non-randomised trial.	These types of bias are unusual in data linkage studies, except where a data linkage system is used in the follow-up of trial participants. Methods of control are best applied in the design of randomisation, design of the intervention delivery processes and by doing an intention to treat analysis.
Compliance bias	Degree of adherence or compliance influences outcome assessment.	
Contamination bias	Intervention-like activities find their way into the non-intervention group.	
Lack of intention to treat analysis	On protocol analysis of subjects where remaining on protocol is risk-related.	

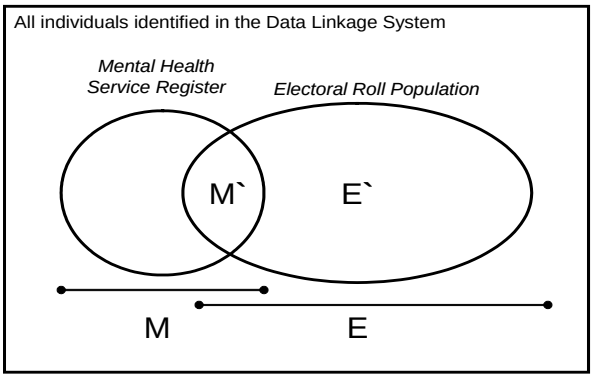
Advanced Theory Topic E:
3.3 Complex Longitudinal Designs Using Domain Restrictions

A number of the biases listed above are amenable to reduction through adoption of ***comparability principles*** (comparability of populations, effects and information), the essence of which is ***counterfactual*** thinking. The aim is to compare groups that are equivalent in all possible ways except for the presence of different levels of exposure. The web of links in a comprehensive data linkage system facilitates several different restriction strategies to improve comparability.

Domain Restriction Strategies Operative on the Exposed

Coverage bias can be reduced by selecting the exposed group from within the same primary study base sampling frame of which the non-exposed group is a (non-exposed) census or random sample.

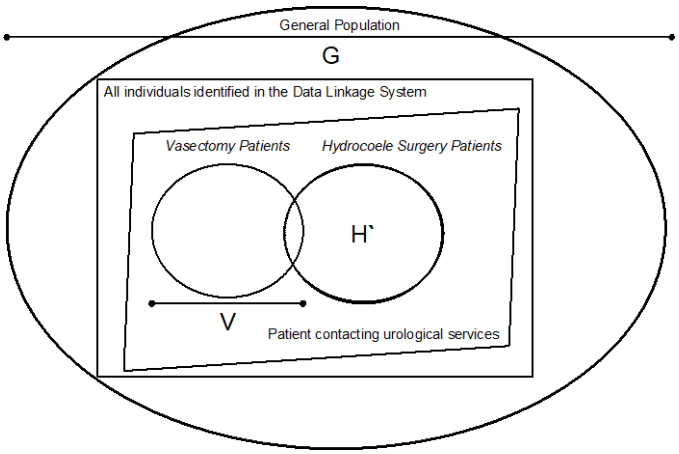
Example: The internal validity of a comparison of primary medical care utilisation between users of mental health services (the 'exposure') and a non-exposed group selected from the electoral roll would be enhanced by restricting the mental health patients to those also found on the electoral roll. From the diagram, a contrast between M` and E` will be of superior comparability to that between M and E` or, worse still, M and E.



Domain Restriction Strategies Operative on the Non-Exposed

Detection bias in a cohort study can be reduced by selecting a non-exposed group that arises from a restricted domain that is homogenous with respect to interactions with the health system whereby detection opportunities might arise.

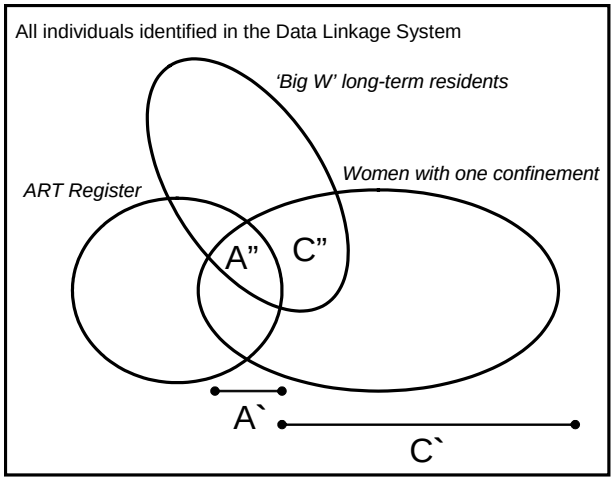
Example: The internal validity of a comparison of incidence rates of prostate cancer between vasectomised and non-vasectomised men would be enhanced by restricting the non-exposed group to men undergoing another elective urological procedure such as surgery for a hydrocoele. From the diagram, a contrast between V and H` will be of superior comparability to a contrast between V and (G-V) and, worse still, V and G.



Domain Restriction Strategies Operative on Both Exposed and Non-Exposed

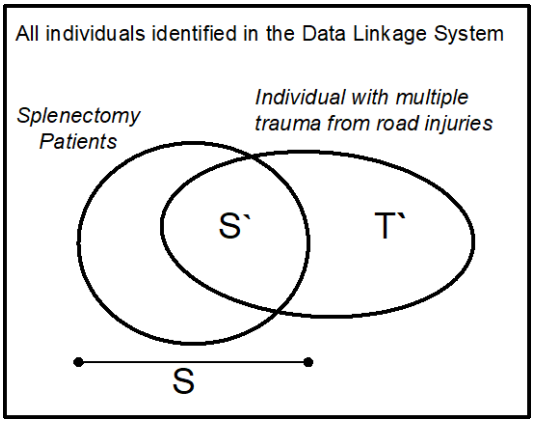
Bias due to loss to follow-up can be reduced by selecting exposed and non-exposed groups from the same primary study base that is defined on the basis of a low likelihood of loss to follow-up regardless of outcome status.

Example: The internal validity of a comparison of long-term cancer incidence rates in women who have had one confinement in labour, exposed and not exposed to Assistive Reproductive Technology (ART) interventions would be enhanced by further restricting the study base to women who have maintained Medicare residential address registration only in Western Australia during the study period ('Big W' status). From the diagram, a contrast between A" and C" will be of superior comparability to a contrast between A` and C`.



Another use of a domain restriction operant on both exposed and non-exposed groups is in the control of confounding by choosing a domain in which confounding factors have no or less 'natural' variation. This is the same as the idea of performing a study of the association of lung cancer to alcohol restricted to non-smokers to prevent any confounding by tobacco smoking.

Example: The internal validity of a comparison of incidence rates of pneumococcal pneumonia in splenectomised and non-splenectomised individuals would be enhanced by restricting the study base to individuals who experience multiple serious injuries from road trauma, with the exposed group being those undergoing splenectomy at that time. Serious road trauma is a major cause of splenectomy and is also associated with social disadvantage and risk-taking behaviour. The latter factors are also associated with pneumococcal pneumonia. By restricting the study to road trauma victims, the splenectomised and non-splenectomised individuals will be rendered more comparable because variation in social disadvantage and risk-taking behaviour will be reduced. From the diagram, a contrast between S` and T` will be of superior comparability to a contrast between S and any broader sample from the data linkage system or general population.



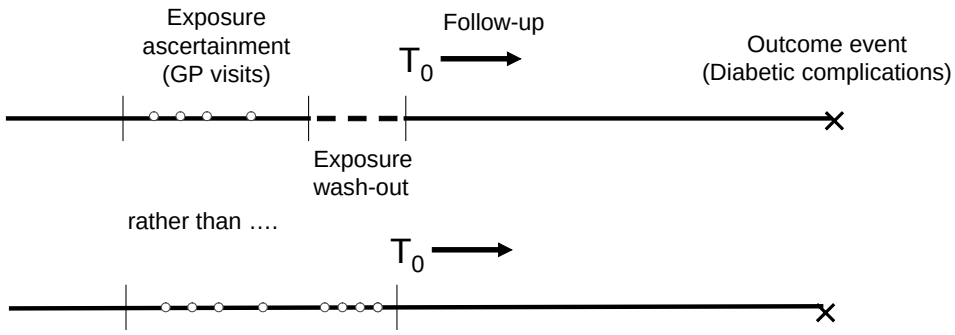
Advanced Theory Topic F:
3.4 Complex Longitudinal Designs Using Time Restrictions

Another means to improve internal validity is to restrict the time interval for either ascertainment of exposure or outcome so as to render them more comparable.

Time Restriction Strategies Operative on Exposure Ascertainment

Protopathic bias can be reduced by restricting the period over which health care exposure is ascertained so as to exclude an appropriate time interval (of ***exposure wash out***) just prior to time-zero for follow-up where protopathic bias might otherwise be operative. The method requires that those enrolled in the cohort at time-zero have no history of the outcome.

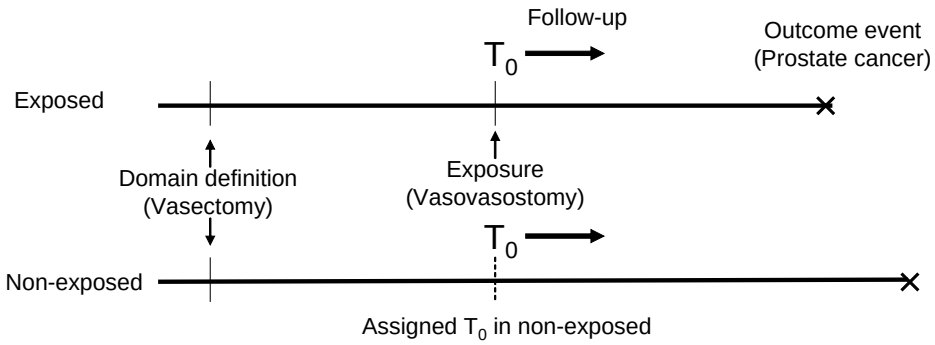
Example: Internal validity of a study of the preventive effect of GP visits on the rate of complications of diabetes would be improved by restricting the time interval for exposure ascertainment to a three-year period ending six months prior to time-zero for follow-up. This would reduce protopathic bias whereby GP visits shortly before time-zero might be the result of early symptoms or signs of emerging complications that manifest within the first year of follow-up. Complications arising in the exposure wash-out period render those diabetics ineligible for follow-up.



Time Restriction Strategies Operative on Outcome Ascertainment

Survivor treatment bias can be reduced by restricting the period over which the outcome is ascertained so that it is no longer correlated with the period over which exposure is ascertained.

Example: The internal validity of a study of the preventive effect of vasovasostomy (vasectomy reversal) on the incidence rate of prostate cancer in vasectomised men would be improved by restricting the time interval for outcome ascertainment by taking time-zero as the date of vasovasostomy in the exposed group and assigning to the non-exposed group an equivalent distribution of time-zeros for commencement of follow-up. This compares favourably with a design where time-zero was taken as the date of vasectomy in both the exposed and non-exposed, thereby creating an artificial correlation between the follow-up time available for both exposure and outcome ascertainment. Note that the time axis in the diagram below is relative time and not calendar time.



The above scenario also lends itself to the use of a time-dependent determinant covariate in Cox regression analysis as originally proposed by Crowley and Hu in comparing outcomes in heart transplant recipients with those in patients waiting for a heart transplant. Alternatively, it is possible to use a time-dependent exposure in a nested case-control study with incidence density sampling. Another approach is multi-stage model analysis as described by Suissa. The design solution suggested here is similar to the use of drug exposure risk windows in unexposed comparator groups in pharmaco-epidemiology.

A paper published by van Walravan and others in 2004 reported that 19% of published clinical observation studies were affected by immortal time bias and that in 5% of the studies, correction of the bias may have materially altered the study results.

Advanced Theory Topic G:
3.5 Nested Case-Control Studies

The nested case-control study is not a means to improve internal validity, but to improve the cost-efficiency of a cohort study or to reduce its incursion on privacy. It is the design of choice when there is an essential exposure variable missing from the administrative data that must be sought through supplementary data collection such as chart review. It may also be preferred for ethical reasons to reduce the volume of health data released from administrative data collections. As with all case-control studies, the essence of the design is that it is a sampling strategy used to determine the distribution of exposure within the underlying study base from which the cases arose.

Example: Patients undergoing cataract surgery have a small risk of the potentially devastating complication of endophthalmitis. A study of the preventive effect of pre-operative antibiotic eye drops, using administrative hospital data, is hampered by lack of information in these routine data on use of antibiotic drops. The additional item of information must be obtained by hospital chart review. Rather than review the charts of thousands of cataract surgery patients, it would be far more efficient to perform a nested case-control study where only the cases of endophthalmitis and a five-fold larger sample of controls from the entire cohort at time-zero are selected for chart review. The results would be much the same as if the entire cohort was used.

Typically a nested case-control study will enrol all available cases and a sample from the study base that is up to 4-5 times larger than the number of cases. Sampling controls beyond a ratio of five starts to yield discouraging diminishing returns for effort. This is because the formula:

$$F = \frac{r}{1+r}$$

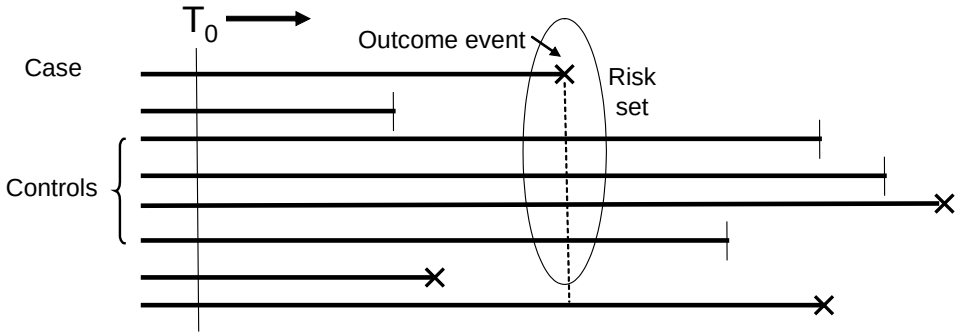
where r is the ratio of controls to cases, defines the fraction, F, of the maximum possible statistical precision that would be achieved if there were an infinite number of controls per case. At r=5 the researcher achieves 83% of maximum possible precision compared with 86% at r=6.

Methods of Control Sampling in Nested Case-Control Studies

Case-cohort sampling: Controls are selected from the entire cohort at time-zero. It is possible for a case to also be a control. Because the controls in this instance are sampled from counts forming the denominators of exposed and non-exposed cumulative incidence proportions, the odds ratio obtained from the case-control 2×2 table is an estimate of the cumulative incidence ratio.

Incidence density sampling: Controls and cases are selected in ***risk sets*** consisting of 1 case and r controls. The controls in each risk set are selected from members of the cohort who are still at risk at the time of occurrence of their corresponding case. Thus a case may be a control in a previous risk set, but not in its own risk set, another at the same time or those occurring afterwards. One individual may also be a control in multiple risk sets. Restricting controls to be used only once and to those who never become cases typically results in biased estimation. Because the controls in this instance are sampled from person–time forming the denominators of exposed and non-exposed incidence rates, the

odds ratio obtained from the case-control 2×2 table is an estimate of the incidence rate ratio. See the diagram below.



Nested Case-Control Study with Linked Data

Ng JQ, Morlet N, Bulsara MK, *et al*. **Reducing the risk for endophthalmitis after cataract surgery: population-based nested case-control study: Endophthalmitis Population Study of Western Australian sixth report.** *J Cataract Refractive Surg* 2007; 33: 269-280.

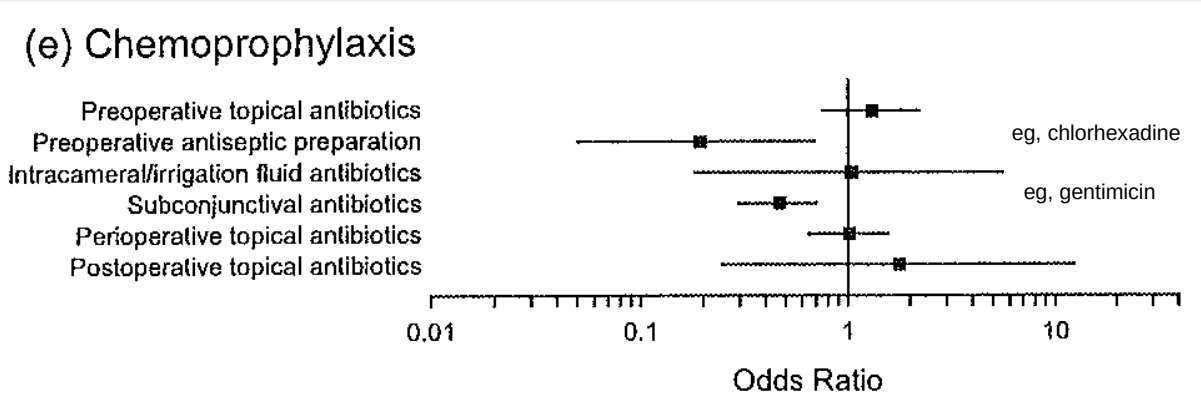
PURPOSE: To characterize operative and nonoperative risks for the potentially blinding complication of endophthalmitis after cataract surgery.

SETTING: Ophthalmology services for the whole state of Western Australia.

METHODS: This retrospective population-based nested case-control study in Western Australia covered the period between 1980 and 2000. For each of the 205 cases of endophthalmitis, 4 time-matched controls were randomly selected from all cataract operations performed in the state. Conditional logistic regression was used to estimate multivariate adjusted odds ratios (ORs) and 95% confidence intervals (CIs) for the risk factors of interest.

RESULTS: Wound location, suturing the wound, and type of cataract operation did not affect the risk for postoperative endophthalmitis. Antibiotic prophylaxis reduced the risk, but only if given as a subconjunctival injection (OR, 0.46; 95% CI, 0.29-0.70). The risk for endophthalmitis was greater with same-day surgery (OR, 2.27; 95% CI, 1.52-3.41) than with admission the day before surgery. A concurrent eyelid procedure was a substantial risk. Surgeons within 2 years of obtaining specialist qualifications were more likely to have a case of endophthalmitis, although this was partly the result of more posterior capsule breaches. Posterior capsule breach increased the risk when it occurred in private hospitals (OR, 13.57; 95% CI, 4.00-45.99), but not in public hospitals.

CONCLUSIONS: Nonoperative as well as operative factors are important in the prevention of endophthalmitis. Subconjunctival injection of antibiotics appears to be beneficial for endophthalmitis prophylaxis. The model showed that active risk management strategies designed to optimize hospitalization and chemoprophylaxis may reduce the incidence of endophthalmitis by up to 81%.



References

Armstrong BK, White E, Sartacci R. **Principles of Exposure Measurement in Epidemiology.** Oxford: Oxford University Press, 1992.

Borgmeier I, Holman CDJ. **Does vasectomy reversal protect against prostate cancer?** *Ann Epidemiol* 2004; 14: 748-749.

Choi BCK. **Bias, overview.** In: Gail MH, Benichou J (Eds). *Enclyclopedia of Epidemiologic Methods*, pp.74-82. Chicester: John Wiley & Sons, 2000.

Collett D. **Modelling Survival Data in Medical Research.** London: Chapman & Hall, 1994.

Crowley J, Hu M. **Covariance analysis of heart transplant survival data.** *J Am Stat Assoc* 1977; 72: 27-36.

Delgado-Rodriguez M, Llorca J. **Bias.** *J Epidemiol Community Health* 2004; 58: 635-641.

Diggle PJ, Liang K-Y, Zeger SL. **Analysis of Longitudinal Data.** Oxford: Clarendon Press, 1994.

Einarsdottir K, Preen DB, Sanfilippo FM, *et al*. **Mortality in Western Australian seniors with chronic respiratory diseases: a cohort study.** *BMC Pub Health* 2010; 10: 385, doi:10.1186/1471-2458-10-385.

Einarsdottir K, Preen DB, Emery J, *et al*. **Regular primary care decreases the likelihood of mortality and hospitalisation in older epilepsy patients.** *Med Care* 2010; 48: 472-476.

Einarsdottir K, Preen DB, Emery JD, *et al*. **Regular primary medical care lowers hospitalisation and mortality in seniors with chronic respiratory disease.** *J Gen Int Med* 2010; 25: 766-773.

Gail MH, Benichou J (Eds). **Encyclopedia of Epidemiologic Methods.** Chicester: John Wiley & Sons, 2000.

Hofler M. **The Bradford-Hill considerations on causality counterfactual perspective.** *Emerging Themes Epidemiol* 2005; 2: 11.

Langholz B, Goldstein L. **Risk set sampling in epidemiologic cohort studies.** *Stat Sci* 1996; 11:35-53.

Levesque LE, Hanley JA, Kezouh A, *et al*. **Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes.** *BMJ* 2010; 340: b5087, doi:10.1136/bmj.b5087

Maclure M, Schneeweiss S. **Causation of bias: the episcope.** *Epidemiology* 201; 12: 114-122.

Maldonado G, Greenland S. **Estimating causal effects.** *Int J Epidemiol* 2002; 31: 422-429.

Ng JQ, Morlet N, Bulsara MK, *et al*. **Reducing the risk for endophthalmitis after cataract surgery: population-based nested case-control study: Endophthalmitis Population Study of Western Australian sixth report.** *J Cataract Refractive Surg* 2007; 33: 269-280.

Qun Mai, Holman CDJ, Sanfilippo FM, *et al*. **Do Users of mental health services lack access to general practitioner services?** *Med J Aust* 2010; 192: 501-506.

Sackett DL. **Bias in analytical research.** *J Chron Dis* 1979; 32: 51-63.

Shafer G. **Commentary: estimating causal effects.** *Int J Epidemiol* 2002; 31: 434-435.

Steineck G, Ahbolm A. **A definition of bias founded on the concept of the study base.** *Epidemiology* 1992; 3: 447-482.

Suissa S. **Immortal time bias in pharmacoepidemiology.** *Am J Epidemiol* 2008; 167: 492–499.

Tamim H, Monfared AA, LeLorier J. **Application of lag-time into exposure definitions to control for protopathic bias.** *Pharmacoepidemiol Drug Safety* 2007; 16: 250-258.

van Walraven C, Davis D, Forster A, *et al*. **Time-dependent bias was common in survival analyses published in leading clinical journals.** *J Clin. Epidemiol* 2004; 57: 672–682.

TRAINING SESSION 3:
IMPLEMENTING COMPLEX LONGITUDINAL DESIGNS

IMPORTANT NOTICE

The practice data sets provided for training purposes in this course are fictitious. All results based on analysis of these fictitious data are scientifically useless. The course organisers and instructor take no responsibility for any misuse of the spurious results arising from the analysis of these fictitious data sets.

Background to Training Session 3

The aim of this training session is to acquaint you with methods of analysis of linked data files that require random sampling procedures in order to implement complex longitudinal designs. The exercise mirrors a study by Borgmeier and Holman, which used a time-restriction operative on outcome assessment to reduce survivor treatment bias. In that study the subject matter was vasovasostomy (ie, vasectomy reversal) and prostate cancer. However, you will implement an analogous design to compare the rates of diabetic complications in female diabetics of reproductive age (an additional domain restriction operative on exposed and non-exposed) who are and are not admitted to hospital for a condition related to pregnancy, childbirth (including normal childbirth) and the puerperium.

Warning on Random Numbers: Because the exercise involves the use of a pseudo-random number generator, participants should anticipate the likelihood of some departures of their results from the print screens displayed and figures given by the instructor. The results quoted were stable in SPSS provided that the specified pseudo-random generator seeds (given in the instructions) were used at each step on every occasion. Whether these random seeds will produce the same random numbers in other versions of SPSS, or in SAS, Stata or R, is untested at this time.

Training Session Resources

You will be making use of your updated platform file saved from Exercise 2 as well as a file of diabetic complications, diabcomp and your tagged and sequenced HMDSdata file.

If you are using this statistical package …	*… you should use this save file.*
SPSS	Diabetes6.sav diabcomp.sav HMDSdata2.sav
SAS	Diabetes6.sas7bdat diabcomp.sas7bdat HMDSdata2.sas7bdat
Stata	Diabetes6.dta diabcomp.dta HMDSdata2.dta
R	Diabetes6.rda diabcomp.rda HMDSdata2.rda

Note that the following additional codes are relevant to this exercise:

Data Set	Coding System	Codes
HMDS		**Complications of diabetes mellitus:**
	ICD-9-CM	Diagnosis codes 112.1, 354.X, 355.X, 356.8, 362.34, 365.X, 366.X, 369.X, 401.X thru 405.X, 410.X thru 414.X, 428.X, 429.2, 429.3, 429.9, 430.X thru 438.X, 440.X, 443.X, 444.X, 447.1, 459.8, 459.9, 580.X thru 586.X, 590.X, 595.X, 599.0, 681.X, 682.X, 707.X, 729.2, 730.17, 784.3X, 785.4, 791.0, V45.1, V56.X
	ICD-10-AM	B37.3, G45.X, G56.X, G57.X, G58.7, G60.8, H25.X, H26.X, H28.X, H34.0, H40.X, H42.8, H54.X, I20.X thru I22.X, I24.X, I25.X, I50.0, I50.1, I50.9, I51.6, I51.7, I51.8, I51.9, I60.X thru I67.X, I69.X, I70.X, I73.X, I74.X, I77.1, I87.2, I99.X, L03.XL89.X, L97.X, L98.4, M54.10, M54.11, M54.19, M79.2, M86.37, M86.47, M87.57, M87.67, M87.87, N00.X, N01.X, N03.X thru N05.X, N07.X, N08.X, N10.X, N11.8 thru N12.9, N15.1, N15.9, N16.X thru N18.X, N28.8, N30.X, N39.0, N77.1, R47.0, R02.X, R80.X, Z49.X, Z99.2
		Procedure codes
	ICD-9-CM	84.1, 84.3
	ICD-10-AM	44338-00, 44358-00, 44361-00, 44361-01, 44364-00, 44364-01, 44367-00, 44367-01, 44367-02, 44376-00
		Conditions of pregnancy, childbirth and puerperium
	ICD-9 and *ICD-9-CM*	630.X thru 677.X
	ICD-10-AM	O00.X thru O99.X (1st character is the capital letter 'O')

The list of diabetic complications is based on a paper by Davis *et al* and covers circulatory, visual, renal, neuropathic and other conditions. Rather than participants wasting time in implementing these complication codes, in this exercise you are given a linkable file containing the relevant data.

WARNING: You should not rely on these codes in any analysis separate from the course as they may contain errors or be incomplete.

General Concepts Involved in Sampling within Linked Files

Sampling within files of administrative health data take on several different forms and purposes, including:

- ***Random sampling*** as in a nested case-control study with case-cohort sampling.
- ***Matched sampling*** as in cohort study with a fixed ratio of non-exposed individuals to exposed individuals within each sex and five-year age group (***category matching***) or nearest age (***individual matching***).
- ***Incidence density sampling*** as in a nested case-control study with incidence density sampling.
- ***Combinations*** of the above; eg, random, category matched, incidence density sampling.

Random Sampling

The most efficient way to undertake random sampling in large data files is through the use of a ***pseudo-random number generator***. This uses a formula to generate numbers which behave very like genuine random numbers and are widely used for simulations of random processes and statistical methods. In most cases a good pseudo-random number generator seems to work as you would expect a genuine random generator to work. Statistical software packages offer a wide range of such generators based on user-selected distributions. The most useful of these for the purpose of this course is the ***uniform distribution***. Pseudo-random number generators require priming by a ***seed*** number, which is either provided by the user, provided as a default by the software, seeded from a previous pseudo-random number generation or read from an area of computer RAM with highly variable states. The following are the uniform distribution pseudo-random number generator functions in SPSS, SAS, Stata and R:

Package:	SPSS	SAS	Stata	R
Function to return a pseudo-random number between 0 and 1 drawn from a uniform distribution.	UNIFORM(1). Set the random number seed under the transform menu to a new number at the start of each session.	ranuni(seed);	set seed generate newvar=uniform()	set.seed(<seed>) runif(n,min,max) e.g. runif(100,0,1)

To draw a random sample without ***replacement*** within a file of administrative health data requires the planning of two necessary entities: (i) the ***sampling frame*** of all of the records eligible for selection; and (ii) the ***sampling fraction***, $f = n/N$, where N is the total size of the sampling frame and n is the desired number of random selections. For example, in a nested case-control study with case-cohort sampling, index records representing all members of the cohort will be the sampling frame. If the size of that frame is 1,000 patients from which 40 cases arise and the investigator decides on a control-case ratio of 5:1, then the desired number of controls is 200. Thus the sampling fraction is 200/1,000 = 0.2. To yield the random sample, each record in the sampling frame is assigned a uniformly distributed random number between 0 and 1 in a new variable, such as sample. Those selected are where sample $\leq f$.

Random Assignment

This concept is closely related to random sampling without replacement. The need for random assignment may arise, for example, when a time restriction strategy is operative on outcome ascertainment in order to reduce survivor treatment bias. Here dates of exposure in exposed individuals need to be randomly assigned to non-exposed individuals. For example, if there are 101 exposed and 1,556 non-exposed individuals in the data set, how does one randomly allocate the 101 exposure dates to the latter group of 1,556?

One method of achieving a random assignment is to separate the records into exposed and non-exposed files and to perform a ***random sort*** of the records in each file. This can be achieved by assigning every record a uniform random number and sorting the files on these random numbers. In our example, the non-exposed file is then divided serially into 15 blocks of 101 records, which are each serially assigned numbers 1 to 101, and a final block of 41 records assigned numbers 1 to 41. These numbers are used as a linkage key to assign exposure dates from the randomly sorted exposure file. It is important that the exposure file has also been randomly sorted to ensure that the final block of 41 records, in the case of the example, is assigned a uniform spread of exposure dates rather than the earliest 41 dates as might happen if the exposure file was sorted according to chronological order.

Category Matched Sampling

To give effect to a category-matched design, it is first necessary to enumerate all of the ***cross-categories*** that are possible given the matching factors. The number of possible combinations, T, will amount to:

$$T = \prod_{i=1}^{k} m_i$$

where m_i is the number of categories of the ith matching factor and k is the total number of matching factors. The identities of cross-category combinations take the form of a matrix with k dimensions. Thus, if group matching is performed on two categories of sex, nine categories of 10-year age group and two categories of place of residence (metro/country) there will be a three-dimensional matrix with $2\times9\times2 = 36$ cross-categories.

Each matching cross-category is given a ***group number***; eg, number 10 might represent males aged 40-49 years in the country. Every record in the sampling frame is then assigned its appropriate group number and the frequencies of records in each group are obtained, stratified as appropriate to the nature of the study design by exposure or disease status. Typically, the intention of the researcher will be to draw a random sample from within each group. This becomes a matter of restricting the

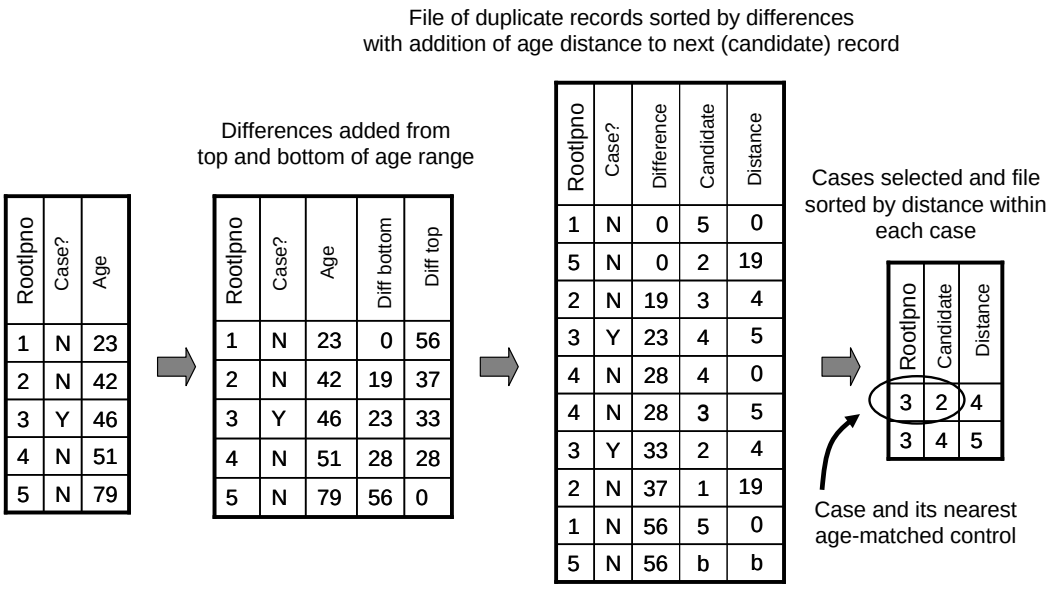
sampling frame to each group in turn and setting the sampling fraction in each instance to yield the desired ratio of controls to cases, or non-exposed to exposed individuals, depending on the design. Note that the capacity to implement category matching schemes with a large number of strata may become limited by sample size in sparse data sets.

Individual Matched Sampling

Individual matched sampling, such as that based on nearest age or nearest date of service, is a more complex procedure. The implementation of individual matching strategies for linked administrative data files will usually involve sorting the records in the sampling frame according to values of the matching factor to bring the candidate records for selection into close proximity with the index record for each match. However, a further complication is introduced according to whether the sampling rules call for unidirectional or bidirectional matched sampling (see below). Also, depending on whether the sampling is undertaken with or without replacement, records may be sampled only once (where the computing syntax is known as a ***greedy algorithm***) or more than once. Rules of ***precedence*** are also necessary when index records effectively 'compete' for matched counterparts.

Unidirectional individual matching: This is relevant where, for example, the record with the nearest subsequent date of service (but not the nearest prior date of service) is selected as the first available record following the index record in a file sorted by the matching factor.

Bidirectional individual matching: This is relevant where, for example, the record with the nearest age in either direction is selected. A handy 'trick' is to establish the range of possible values of the matching factor; eg, 23-79 years of age. The matching factor is then re-expressed on each record as the difference from the bottom of the range; and a duplicate set of records is added to the file with the matching factor re-expressed as the difference from the top of the range. The file is sorted on all of these differences and nearest matches (in both directions) may then be found by moving through the sorted file in one direction, provided that records are not matched to themselves or to a previously drawn record, if these possibilities are excluded by the sampling rules. The record number of and ***sampling distance*** to the nearest matched record is internally loaded onto the index records which are saved as a cut-down file with each index record represented twice. This file is sorted by sampling distance within index record number. The first record within each couplet of index records will have the smaller sampling distance and should reveal the identity of the best individually matched record. Rules of precedence for candidate records of equal sampling distance above and below an index record can be defined and implemented. The following diagram illustrates this ***dual representation method*** of bidirectional individual matched sampling.



RootIpno	Case?	Age
1	N	23
2	N	42
3	Y	46
4	N	51
5	N	79

RootIpno	Case?	Age	Diff bottom	Diff top
1	N	23	0	56
2	N	42	19	37
3	Y	46	23	33
4	N	51	28	28
5	N	79	56	0

RootIpno	Case?	Difference	Candidate	Distance
1	N	0	5	0
5	N	0	2	19
2	N	19	3	4
3	Y	23	4	5
4	N	28	4	0
4	N	28	3	5
3	Y	33	2	4
2	N	37	1	19
1	N	56	5	0
5	N	56	b	b

RootIpno	Candidate	Distance
3	2	4
3	4	5

Incidence Density Sampling

A method of incidence density sampling suitable for a nested case-control study is to sort the file containing the sampling frame for all cases by descending total length of follow-up time, whether ended due to becoming a case, censoring from loss to follow-up or the last date of the study. A fileseq sequence variable is assigned and a separate sampfram variable is created. The value of sampfram is the fileseq value of the last preceding record to have a follow-up time at least one day after the current record. It follows that for each case record the value of sampfram defines the sampling frame of individuals still contributing person-time on the case's incident date and from whom controls should be selected. For example, a case with fileseq = 3,442 might have a sampfram = 3,440. This indicates that the first 3,440 records are available for sampling of controls. Record number 3,441 is not available because it exited the person-time at risk on the same follow-up day as the case. If five controls need to be selected, five uniform 0-1 random numbers are generated and each is multiplied by 3,441 with the results truncated down to the nearest whole numbers (the very rare result of exactly 3,441 is rounded down to 3,440). If the same number is drawn twice, a redraw takes place. These five numbers are the fileseq values of the randomly selected controls in the risk set for that case.

Overview of Exercise 3

Steps 1-2. Load dates of first diabetic complication dates onto diabetes platform and create binary indicator.

Steps 3-4. Load aug date from platform file onto the HMDSdata file.

Steps 5-7. Tag pregnancy-related records in HMDSdata, sequence post-aug pregnancy records, cut-down to dates for 1st-time.

Steps 8-9. Load pregnancy dates onto diabetes platform and restrict study domain to women aged 15-39 at aug date.

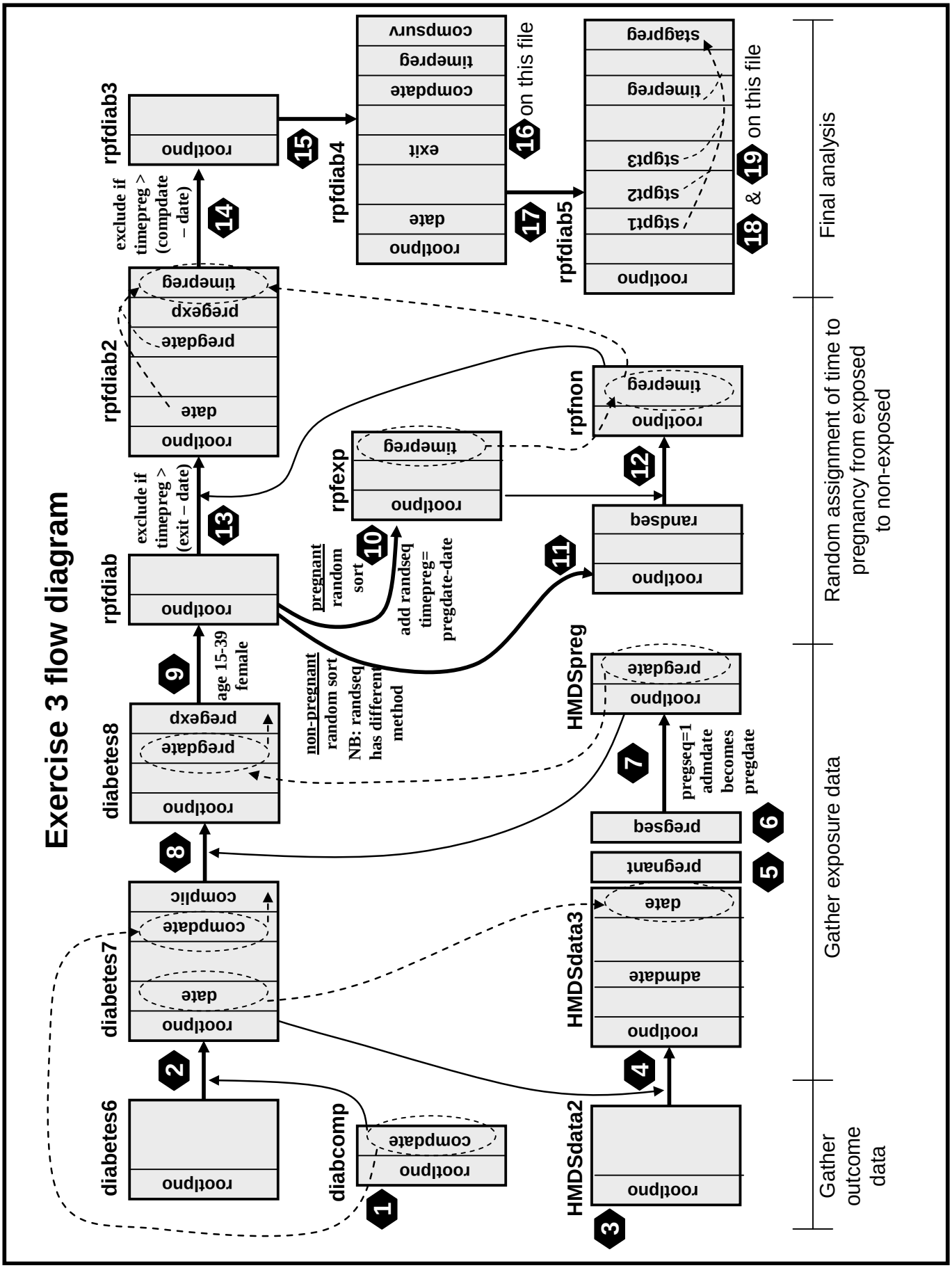
Steps 10. Create derivative file of women with a pregnancy, randomly sort, calculate time to pregnancy from date and cut-down.

Steps 11. Create derivative file of women with no pregnancy, randomly sort, apply block numbering and cut-down.

Steps 12-13. Assign (load) time to pregnancy from exposed to non-exposed women; then load this information onto platform; exclude non-exposed women with an impossible virtual pregnancy date.

Steps 14-15. Exclude women who developed complications prior to their real or virtual pregnancy date and calculate survival time to complication.

Steps 16-19. Perform survival analysis on pregnant/non-pregnant women.


Exercise 3 flow diagram
diabetes6
rootpno
diabetes7
rootpno
date
compdate
complic
diabcomp
rootpno
compdate
diabetes8
rootpno
pregdate
pregexp
rpfdiab
rootpno
rpfdiab2
rootpno
date
pregdate
pregexp
timepreg
rpfdiab3
rootpno
rpfdiab4
rootpno
date
exit
compdate
timepreg
compsurv
rpfdiab5
rootpno
stgpt1
stgpt2
stgpt3
timepreg
stagpreg
HMDSdata2
rootpno
HMDSdata3
rootpno
admdate
date
pregant
pregseq
HMDSpreg
rootpno
pregdate
rpfexp
rootpno
timepreg
rpfnon
rootpno
timepreg
randseq
rootpno
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18 & 19
exclude if timepreg > (compdate – date)
exclude if timepreg > (exit – date)
age 15-39 female
pregnant random sort
add randseq timepreg= pregdate-date
non-pregnant random sort NB: randseq has different method
pregseq=1 admdate becomes pregdate
on this file
on this file
Final analysis
Random assignment of time to pregnancy from exposed to non-exposed
Gather exposure data
Gather outcome data

Exercise Instructions

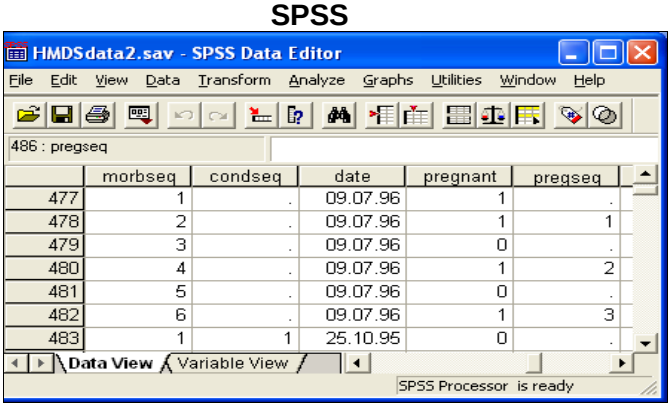
Identification of Complications of Diabetes Mellitus

1. For the purpose of this exercise we will define a diabetes-related complication as either (i) death from any cause; or (ii) a hospital admission for any of the conditions identified by Davis *et al* as diabetes-related complications and represented by mention of a hospital morbidity diagnosis or procedure code listed in your briefing session. To save you time the file diabcomp has been prepared for you in advance. Open this file and note that it contains only two variables: rootlpno and compdate. The file is limited to those individuals who experienced a diabetes–related complication during the study period. For these individuals, compdate is their date of first hospital admission for any diabetes-related complication or, if no such admission occurred, their date of death. It is important to note that not all designated diabetes-related complications (eg, ischaemic heart disease) will have been caused by diabetes and thus is it possible for the incident date of a complication to precede the inaugural index date of a diabetes diagnosis.
2. Open the updated diabetes6 platform file and using a merge procedure load compdate onto the end of the records using the diabcomp file as an external look-up table keyed by rootlpno. Add a binary indicator variable, complic, which is assigned the value '1' if a complication occurred, but is otherwise set to '0'. Run a frequency distribution on complic to confirm that 6,196 of the 10,675 patients had complications of diabetes. Save the updated platform file as diabetes7.

Implementation of Domain and Time Restrictions in a Longitudinal Design

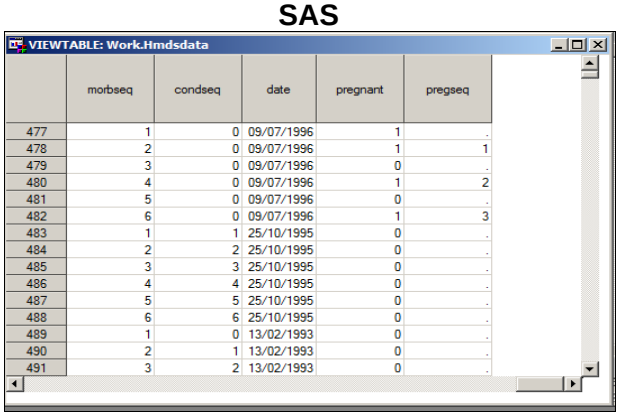
3. Open your previously sequenced HMDSdata2 file.
4. Load the inaugural index date variable from your platform file onto the end of each record, using the diabetes7 file as an external look-up table keyed by rootlpno.
5. Tag all records that mention a pregnancy, childbirth or puerperium diagnosis code, using a new variable, pregnant. Set pregnant to '1' for a tagged record and '0' otherwise.
6. Create a pregseq variable that numbers only the tagged pregnancy records within each block starting with the first tagged record on or after the inaugural index date. This ensures that only pregnancy admissions occurring after the development of diabetes are enumerated. Check that your file looks like the screen shown opposite with record number 477 being an example of a pregnancy admission occurring prior to the inaugural index date. Save as HMDSdata3.

SPSS



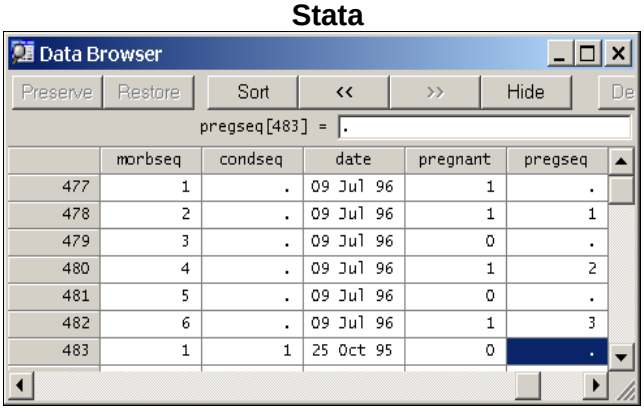
	morbseq	condseq	date	pregnant	pregseq
477	1	.	09.07.96	1	.
478	2	.	09.07.96	1	1
479	3	.	09.07.96	0	.
480	4	.	09.07.96	1	2
481	5	.	09.07.96	0	.
482	6	.	09.07.96	1	3
483	1	1	25.10.95	0	.

SAS



	morbseq	condseq	date	pregnant	pregseq
477	1	0	09/07/1996	1	.
478	2	0	09/07/1996	1	1
479	3	0	09/07/1996	0	.
480	4	0	09/07/1996	1	2
481	5	0	09/07/1996	0	.
482	6	0	09/07/1996	1	3
483	1	1	25/10/1995	0	.
484	2	2	25/10/1995	0	.
485	3	3	25/10/1995	0	.
486	4	4	25/10/1995	0	.
487	5	5	25/10/1995	0	.
488	6	6	25/10/1995	0	.
489	1	0	13/02/1993	0	.
490	2	1	13/02/1993	0	.
491	3	2	13/02/1993	0	.

Stata



	morbseq	condseq	date	pregnant	pregseq
477	1	.	09 Jul 96	1	.
478	2	.	09 Jul 96	1	1
479	3	.	09 Jul 96	0	.
480	4	.	09 Jul 96	1	2
481	5	.	09 Jul 96	0	.
482	6	.	09 Jul 96	1	3
483	1	1	25 Oct 95	0	.

7. Rename the variable admdate to pregdate and cut the file down vertically so that it contains only first post-diabetes tagged records (pregseq = 1), and horizontally so it contains only rootlpno and pregdate. Exclude records with missing data. Check that you are on track by confirming that your cut down file contains 213 records. If all is well, save the cut down HMDS source file with the filename HMDSpreg.

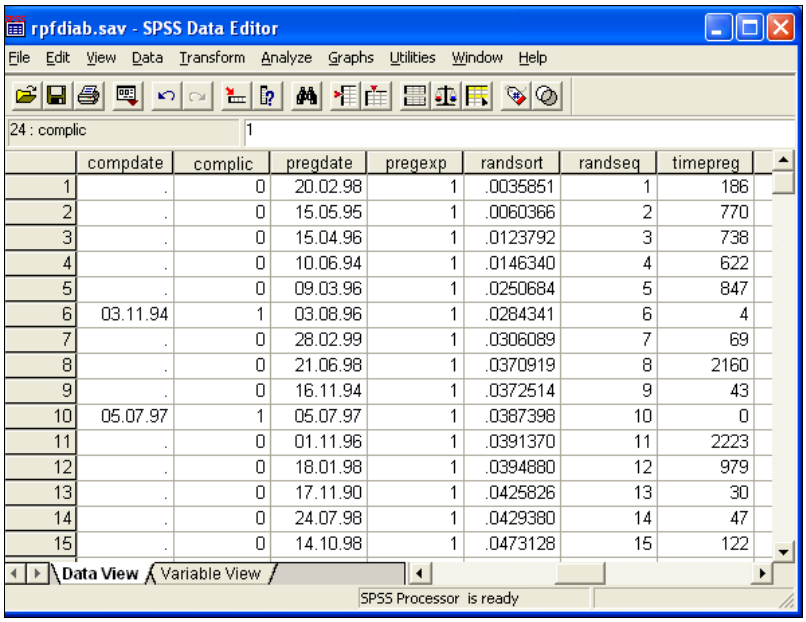
8. Open your diabetes7 platform file again and this time load pregdate onto the end of each record, using HMDSpreg as an external look-up table keyed by rootlpno. Add a binary indicator variable, pregexp, which is assigned the value '1' if a pregnancy admission (ie, the exposure) occurred, but is otherwise set to '0'. You now have all of the basic information you need on your platform file to proceed with the analysis. Save as diabetes8.

9. You should apply the domain restriction to women of 'reproductive age'. There are many ways this could be implemented, but we will simply select women aged 15-39 years at the time of index diagnosis using age. Better approaches would have been to impose the age restriction at the date of exposure or pseudo exposure or even to match on age. With up to 10 years of follow-up, this will impose an upper age limit of 49 years. Delete the males with diabetes and the women of younger or older ages from the file. Check that 716 records remain and save this domain-restricted platform as rpfdiab.

10. The next step is to create and save two separate derivative files for women in the domain who were and were not exposed to a pregnancy admission. Start by deleting the non-exposed women leaving 200 women of 'reproductive age' who had a pregnancy admission after their diabetes diagnosis. Set the seed in your random number generator to 2,000,000 and assign a random variable, randsort, using a uniform random distribution from 0 to 1.

    Sort the file in ascending order of randsort. Create a new file sequence variable in the usual manner to number the records from 1 to 200, but call it randseq. Calculate a new variable, timepreg, as the number of days elapsed between the inaugural index date and the time of first pregnancy admission, pregdate. If your random number generator has reacted to the seed of 2,000,000 the same way as mine did, your file should look like the screen shown opposite.

SPSS



	compdate	complic	pregdate	pregexp	randsort	randseq	timepreg
1	.	0	20.02.98	1	.0035851	1	186
2	.	0	15.05.95	1	.0060366	2	770
3	.	0	15.04.96	1	.0123792	3	738
4	.	0	10.06.94	1	.0146340	4	622
5	.	0	09.03.96	1	.0250684	5	847
6	03.11.94	1	03.08.96	1	.0284341	6	4
7	.	0	28.02.99	1	.0306089	7	69
8	.	0	21.06.98	1	.0370919	8	2160
9	.	0	16.11.94	1	.0372514	9	43
10	05.07.97	1	05.07.97	1	.0387398	10	0
11	.	0	01.11.96	1	.0391370	11	2223
12	.	0	18.01.98	1	.0394880	12	979
13	.	0	17.11.90	1	.0425826	13	30
14	.	0	24.07.98	1	.0429380	14	47
15	.	0	14.10.98	1	.0473128	15	122

If your are satisfied, delete all of the variables except rootlpno, randseq and timepreg, then save this cut-down file of exposed women of reproductive age as rpfexp.

SAS

VIEWTABLE: Aalhd.Rpfexp

	Date of 1st complication or death	complic	pregdate	pregexp	randsort	randseq	time
1	.	0	20/02/1998	1	0.0035850769	1	186
2	.	0	15/05/1995	1	0.0060365694	2	770
3	.	0	15/04/1996	1	0.0123791946	3	738
4	.	0	10/06/1994	1	0.014634049	4	622
5	.	0	09/03/1996	1	0.0250684274	5	847
6	03/11/1994	1	03/08/1996	1	0.0284340712	6	4
7	.	0	28/02/1999	1	0.0306089344	7	69
8	.	0	21/06/1998	1	0.0370919216	8	***
9	.	0	16/11/1994	1	0.0372513612	9	43
10	05/07/1997	1	05/07/1997	1	0.0387397669	10	0
11	.	0	01/11/1996	1	0.0391370407	11	***
12	.	0	18/01/1998	1	0.0394879817	12	979
13	.	0	17/11/1990	1	0.0425825669	13	30
14	.	0	24/07/1998	1	0.0429379628	14	47
15	.	0	14/10/1998	1	0.0473128394	15	122
16	.	0	15/02/1995	1	0.0479127574	16	***

Stata

Data Browser

Preserve | Restore | Sort | << | >> | Hide | Delete...

pregdate[1] = 14247

	compdate	complic	pregdate	pregexp	randsort	randseq	timepreg
1	.	0	03 Jan 99	1	.0073832	1	2133
2	07 Sep 96	1	11 Mar 96	1	.0158828	2	52
3	.	0	15 Sep 95	1	.0272813	3	31
4	.	0	31 Dec 91	1	.0348112	4	194
5	.	0	18 Jan 98	1	.0392386	5	979
6	.	0	17 Dec 90	1	.0399714	6	81
7	.	0	16 Apr 96	1	.0411559	7	48
8	.	0	06 Jun 96	1	.0437911	8	739
9	04 Jun 94	1	04 Mar 92	1	.0516727	9	19
10	.	0	04 Oct 99	1	.0527438	10	14
11	.	0	09 Mar 96	1	.0707721	11	847
12	.	0	01 Jan 91	1	.0773118	12	116
13	.	0	18 Jun 95	1	.0805826	13	36
14	.	0	22 Nov 94	1	.0835781	14	1713
15	.	0	13 May 94	1	.0899882	15	54

11. Open rpfdiab again and delete the exposed women leaving 516 who had no pregnancy admission after their diabetes diagnosis. Follow similar steps as above: (i) set the random number generator seed to 3,000,000 (not 2,000,000); (ii) assign randsort as a uniform random number between 0 and 1; and (iii) sort the file in ascending order of randsort. A randseq variable is now created, but the method is different here than in the file of exposed women. The first 200 records are numbered sequentially from 1 to 200; the counter is then reset to 1 and numbering begins again such that the 400th record is again numbered 200; the last 116 records are numbered 1 to 116.
12. Sort the file of non-exposed women by ascending order of randseq. Load timepreg onto each record using the rpfexp as an external look-up table keyed by randseq. Resort the file by rootlpno, delete all of the variables except rootlpno and timepreg and save the file as rpfnon.
13. Open rpfdiab again and load timepreg using rpfnon as an external look-up table keyed by rootlpno. This will place a virtual 'time to pregnancy' value on every record for non-exposed women. Assign a real value to timepreg (= days between date and pregdate) in the exposed women and compare the means and standard deviations of timepreg between the two groups (see opposite).

SPSS

Output12 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Report

Time to Pregnancy Admission

Pregnancy	Mean	N	Std. Deviation
0	464.42	516	629.912
1	458.80	200	623.526
Total	462.85	716	627.705

SPSS Processor is read

SAS

Output - (Untitled)

The SAS System 14:18 Thursday, November 16,

The MEANS Procedure

Analysis Variable : timepreg

pregexp	N Obs	N	Mean	Std Dev
0	516	516	464.4	629.9
1	200	200	458.8	623.5

Stata

Intercooled Stata 9.2

File Edit Prefs Data Graphics Statistics User Window Help

Results

```
Summary for variables: timepreg
     by categories of: pregexp (Pregnancy Admission Flag)

pregexp |         N      mean        sd
--------+----------------------------------
      0 |       516   458.157  626.8577
      1 |       200   458.805  623.5255
--------+----------------------------------
  Total |       716   458.338  625.4929
```

Command

C:\data

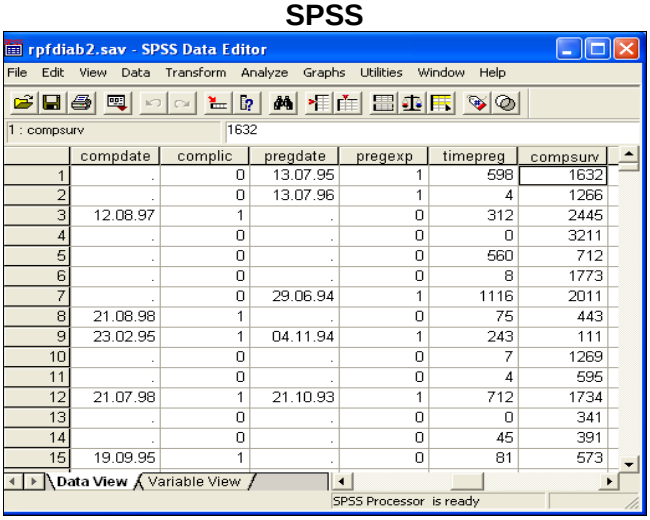
In my file the distributions of time to real or virtual pregnancy in the exposed and non-exposed women were closely alike as one would hope.

At this point an important exclusion process is undertaken. A number of non-exposed women will have been assigned virtual times to pregnancy that exceed their total length of available follow-up. On average, such women will have relatively short follow-up times, thus reducing the prospects for exposure and complications to be observed – part of the reason for survivor treatment bias. These non-exposed women should be excluded from the analysis. Identify them by creating a flag variable which is set to '1' if timepreg exceeds the number of days between date and exit; otherwise set the flag to '0'. Run a frequency distribution on flag. In my file there were 54 non-exposed women who were flagged. Exclude them from the file (leaving 662 records in my file; 655 in Stata after excluding 61) and delete the now superfluous flag variable. Save as rpfdiab2.

14. Another restriction that should be imposed at this time is to exclude women who developed diabetes-related complications prior to the date of pregnancy admission in the exposed and the equivalent virtual date (represented by date + timepreg) in the non-exposed. Flag any records that violate this restriction (there were 73 in my file: 14 exposed and 59 non-exposed; 66=14+52 in Stata) and exclude them, delete the superfluous flag and save as rpfdiab3 (my file had 589 records). The purpose of this restriction is to reduce what is known as 'sick quitter bias'. In this case women who previously developed diabetic complication may have been less inclined or able to become pregnant. The similar distributions of time to virtual or real pregnancy helps to ensure that comparable clearance periods are applied to both groups. In a real study it would be desirable for this clearance period to extend back in time for 1-3 years so that one would have greater confidence that the outcome events being compared were truly first-time complications.

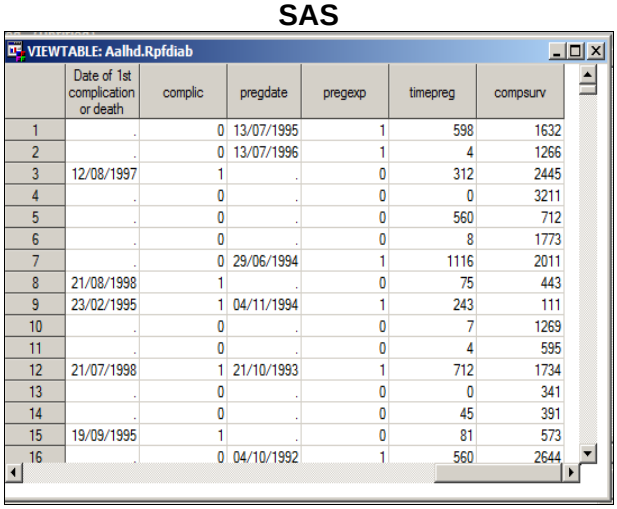
15. In order to perform a survival analysis you require an outcome status variable and a time variable. The appropriate binary outcome status variable is already available in your file: complic. Create a compsurv time variable which is initially set to the days from date to compdate for those with complications and from date to exit for those without complications. Time-zero is then moved forward to the real or virtual date of pregnancy admission by reducing compsurv by timepreg. By this stage my file looked like this:

SPSS



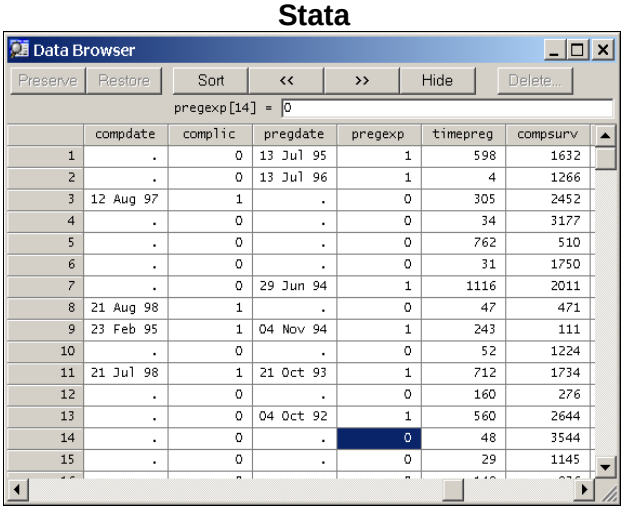
	compdate	complic	pregdate	pregexp	timepreg	compsurv
1	.	0	13.07.95	1	598	1632
2	.	0	13.07.96	1	4	1266
3	12.08.97	1	.	0	312	2445
4	.	0	.	0	0	3211
5	.	0	.	0	560	712
6	.	0	.	0	8	1773
7	.	0	29.06.94	1	1116	2011
8	21.08.98	1	.	0	75	443
9	23.02.95	1	04.11.94	1	243	111
10	.	0	.	0	7	1269
11	.	0	.	0	4	595
12	21.07.98	1	21.10.93	1	712	1734
13	.	0	.	0	0	341
14	.	0	.	0	45	391
15	19.09.95	1	.	0	81	573

SAS



	Date of 1st complication or death	complic	pregdate	pregexp	timepreg	compsurv
1	.	0	13/07/1995	1	598	1632
2	.	0	13/07/1996	1	4	1266
3	12/08/1997	1	.	0	312	2445
4	.	0	.	0	0	3211
5	.	0	.	0	560	712
6	.	0	.	0	8	1773
7	.	0	29/06/1994	1	1116	2011
8	21/08/1998	1	.	0	75	443
9	23/02/1995	1	04/11/1994	1	243	111
10	.	0	.	0	7	1269
11	.	0	.	0	4	595
12	21/07/1998	1	21/10/1993	1	712	1734
13	.	0	.	0	0	341
14	.	0	.	0	45	391
15	19/09/1995	1	.	0	81	573
16	.	0	04/10/1992	1	560	2644

Stata

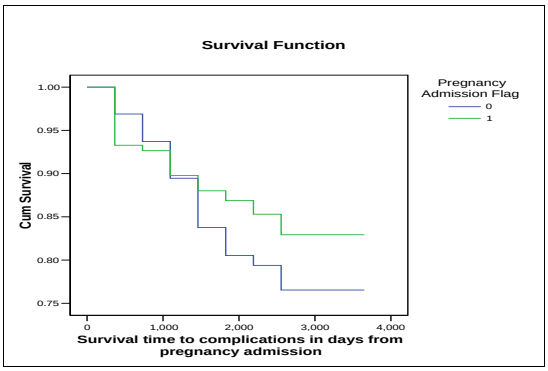


	compdate	complic	pregdate	pregexp	timepreg	compsurv
1	.	0	13 Jul 95	1	598	1632
2	.	0	13 Jul 96	1	4	1266
3	12 Aug 97	1	.	0	305	2452
4	.	0	.	0	34	3177
5	.	0	.	0	762	510
6	.	0	.	0	31	1750
7	.	0	29 Jun 94	1	1116	2011
8	21 Aug 98	1	.	0	47	471
9	23 Feb 95	1	04 Nov 94	1	243	111
10	.	0	.	0	52	1224
11	21 Jul 98	1	21 Oct 93	1	712	1734
12	.	0	.	0	160	276
13	.	0	04 Oct 92	1	560	2644
14	.	0	.	0	48	3544
15	.	0	.	0	29	1145

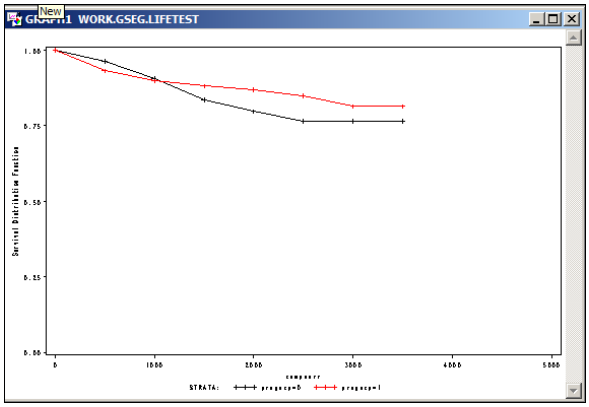
Save the file as rpfdiab4.

16. Perform an actuarial survival analysis with complic as the outcome status variable and compsurv as the time variable, comparing survival plots between women exposed and not exposed to a pregnancy admission based on pregexp. Notice how the pregnant group have a higher risk of complications within the first 2-3 years, but after that their survival from complications falls away less rapidly than the non-pregnant group.

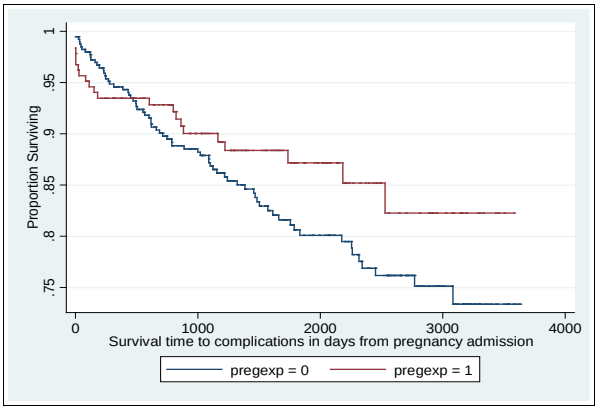
SPSS



SAS



Stata



17. To see if a hunch about confounding by age and disease severity stands up to empirical scrutiny, we need to create a suitable disease severity covariate, stagpreg (=1, 2, 3) based on stage of diabetes at the time of real or virtual pregnancy. Devise an algorithm to correctly classify the women as falling into disease stages 1, 2 and 3 using the variables: timepreg, stgpt1, stgpt2 and stgpt3. **Hint**: for women in stage 2 at the time of real or virtual pregnancy, timepreg will be > stgpt1 and ≤ (stgpt1 + stgpt2). I found 417 women in stage 1, 85 in stage 2 and 87 in stage 3 at the time of real or virtual pregnancy admission (417, 90 and 82 respectively in Stata). Save as rpfdiab5.
18. Perform a series of Cox regressions with complic as the status variable and compsurv as the time variable. In the first model just place pregexp as the only determinate. Add age as a continuous covariate to the second model, and add stagpreg (with age and pregexp) as a categorical covariate in the third model. What happens to the estimates of the IRR of complications comparing pregnant to non-pregnant women as more covariates are added. My IRRs went from 0.79 in the unadjusted model to 0.86 after adjustment for age, and 0.87 after further adjustment for stage of diabetes (0.63, 0.68 and 0.73 respectively in Stata).
19. Now reapply the third model above to a subset of records in the file based on the criterion that timepreg is greater than 270 days; and reapply the model again to the alternative set of records where timepreg is less than or equal to 270 days. What do you notice about the two IRR's for pregnancy exposure arising from these two different strata? What is the technical name given to this phenomenon? Can you think of a biological explanation? My results were IRR = 1.46 for timepreg > 270 and IRR=0.65 for timepreg ≤ 270 (1.20 and 0.52 respectively in Stata).

References

Borgmeier I, Holman CDJ. **Does vasectomy reversal protect against prostate cancer?** *Ann Epidemiol* 2004; 14: 748-749.

Davis WA, Knuiman MW, Hendrie D, *et al*. **Determinants of diabetes-attributable non–blood glucose–lowering medication costs in type 2 diabetes: the Fremantle Diabetes Study.** *Diab Care* 2005; 28: 329-336.

REVIEW 3: 'COMPLEX LONGITUDINAL DESIGNS USING LINKED DATA'

Please check that you are familiar with the following topics:

- Direction of bias ☐
- Types of selection bias, information bias and confounding ☐
- Use of domain restrictions to improve internal validity ☐
- Use of time restrictions to improve internal validity ☐
- Nested case-control studies: case-cohort and incidence density sampling ☐

Please check that you understand the meaning of the following terms:

- Bias and precision ☐
- Internal validity, external validity scientific generalisability ☐
- Coverage bias and survivor treatment bias ☐
- Differential and non-differential misclassification error ☐
- Detection bias and protopathic bias ☐
- Confounding by indication ☐
- Risk set ☐
- Random, (category and individual) matched and incidence density sampling ☐
- Unidirectional and bidirectional individual matching ☐
- Pseudo-random number generation ☐
- Uniform distribution ☐
- Sampling frame, sampling fraction, sampling distance and replacement ☐

Please check that you have demonstrated that you have acquired the following skills:

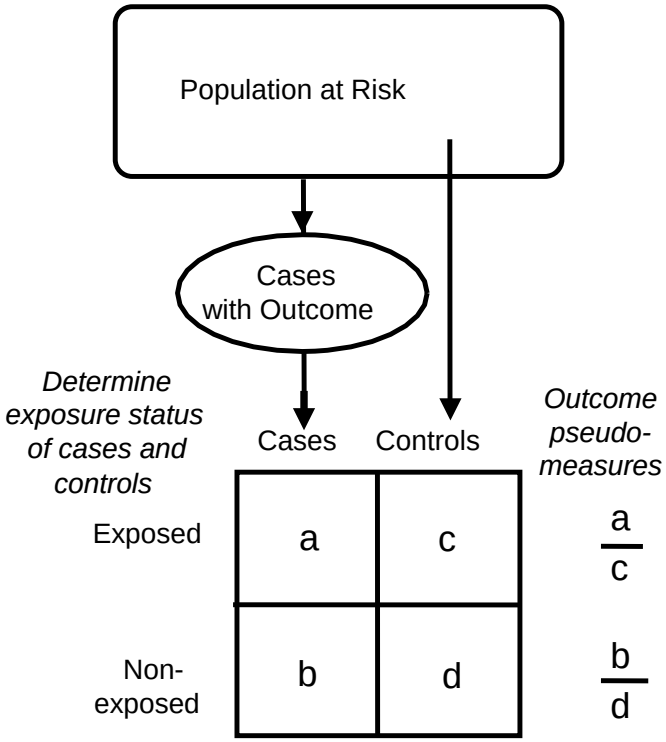
- Random assignment of virtual exposure information ☐
- Implementation of domain and time restrictions in complex longitudinal designs ☐

4. ADVANCED TOPICS IN CASE DISTRIBUTIONS STUDIES

4.1 Revision of the Case-Control Sampling Strategy

Those who have undertaken the introductory course, or an epidemiology course informed by Mietinen's insights into study design, will be aware that the 'case-control study' is not a distinct design in its own right, but a sampling strategy that may be applied in a wide range of situations to estimate the size of relative effect measures.

The essence of the case-control sampling strategy consists in operational terms of a complete enumeration of cases arising from a population at risk and classified with respect to exposure status; combined with a sample, known as the controls, drawn from the population at risk and also classified according the exposure status. The role of the controls is to provide information on the distribution of exposure within the population at risk from which the cases arose (see opposite). A common misunderstanding is to incorrectly construe the role of the controls as a group of non-diseased individuals available to compare against the cases (the traditional view adopted by Cornfield in 1951). This misconception arises from what Miettinen called the ***trohoc fallacy*** – the belief that a case-control design is a cohort study done backwards.



From Miettinen's 1985 text:

> *"The prevailing concept of the case-control approach is that it involves the comparison of patients* (cases) *with a group of controls that consists of persons who are free of the disease under study. In this text the outlook is somewhat different. … the conceptual focus of research is … an occurrence relation. Thus, concerns for comparisons are confined to those between categories of determinants* (meaning a contrast between exposed vs non-exposed and not diseased vs non-diseased) …
>
> *With the cases in the study base identified and information on them obtained, one is in a position to know the numerators of the compared rates. The remaining concern is not that of comparing these cases with non-cases, but that of learning about the denominators of the empirical rates. … For these reasons, the notion of case-noncase comparison as the essence of case-referent studies should give way to one of sampling, because the experience of the study base is captivated by ascertaining and examining the cases that occur in it, and by obtaining a sample of the base itself."*

And from Miettinen's 1999 paper:

> *"According to our dictionary* [meaning Last's *Dictionary of Epidemiology] … case-referent study … is one in which the past history of exposure to a suspected risk factor is compared between cases and controls, the latter being persons who … do not have the disease … It thus represents quite a reversal of the 'cohort' study concept, so that the 'trohoc' term is quite apposite …*

From here, alas, benightedness sets it. There is no understanding that, whereas an etiologic study generally is to address incidence in relation to etiologic determinants … follow-up of a study population of persons who have not yet experienced the outcome event is inescapable. … And it defines 'nested case-control study', as though some 'case-control' studies were not 'nested' in a population …

… the cases do not constitute a study group but a rate numerator series, and from this would immediately flow the need, not for a control group but for a denominator series in the meaning of a fair sample of the study base … And given this understanding … quasi-rates specific to the contrasted [exposure] histories would be understood to be immediate results. Thus, concerns for comparability that focus on comparisons between the numerator and denominator series would be understood to be absurd, never providing for understanding, for example, the confounding alternative to causality."

Correct sampling of the controls is a critical design consideration in a case-control scheme. The primary consideration is that controls should provide an unbiased representation of the distribution of exposure status within the population at risk (of becoming a case) *as formulated in the denominators of the epidemiologic measures chosen for comparison.* The significance of the latter part of this statement lies in the fact that exposure distributions within populations at risk differ according to how one accounts for the observation of population experience. For example, the distribution of exposure to statins (cholesterol-lowering drugs) in a cohort of diabetic patients will differ according to whether the population is described in terms of a head count at time-zero or the accumulation of person-time over a given time interval such as ten years. If statins reduce mortality in diabetics, one would expect the ratio of statin-exposed to non-exposed diabetics at time-zero to be lower than the ratio of statin-exposed to non-exposed diabetic person-time over ten years of follow-up. Thus, according to the nature of the relative effect measure one chooses to estimate, whether a cumulative incidence ratio or an incidence rate ratio, a different control sampling strategy will become necessary. This is the basis for the distinction between ***case-cohort sampling*** and ***incidence density sampling*** in nested case-control studies. We can now generalise the first of these to ***case-base sampling*** in any scheme used to estimate the ratio between two proportions. These situations include application to a cross-sectional prevalence survey data set, where controls are randomly selected from the entire population of survey participants and cases are those with the attribute to be enumerated in the numerators of prevalence measures compared between exposed and non-exposed.

Thus, with case-base sampling, the ratio of exposed to non-exposed controls provides an estimate of the ratio of exposed to non-exposed head counts in the denominators of proportions, ie, $c/d \approx N_1/N_0$. This relationship is used to estimate the cumulative incidence ratio (or prevalence ratio or other ratio of proportions) comparing the risk in the exposed with the risk in the non-exposed.

$$\begin{aligned} CIR = CI_1/CI_0 &= (a/N_1) \div (b/N_0) \\ &= (a/b) \div (N_1/N_0) \\ &= (a/b) \div (c/d) \\ &= ad/bc \quad \text{known as the cross-product or odds ratio (OR).} \end{aligned}$$

With incidence density sampling, the ratio of exposed to non-exposed controls provides an estimate of the ratio of exposed to non-exposed person-time, i.e., $c/d \approx PT_1/PT_0$. This relationship is used to estimate the incidence rate ratio comparing the rate in the exposed with the rate in the non-exposed.

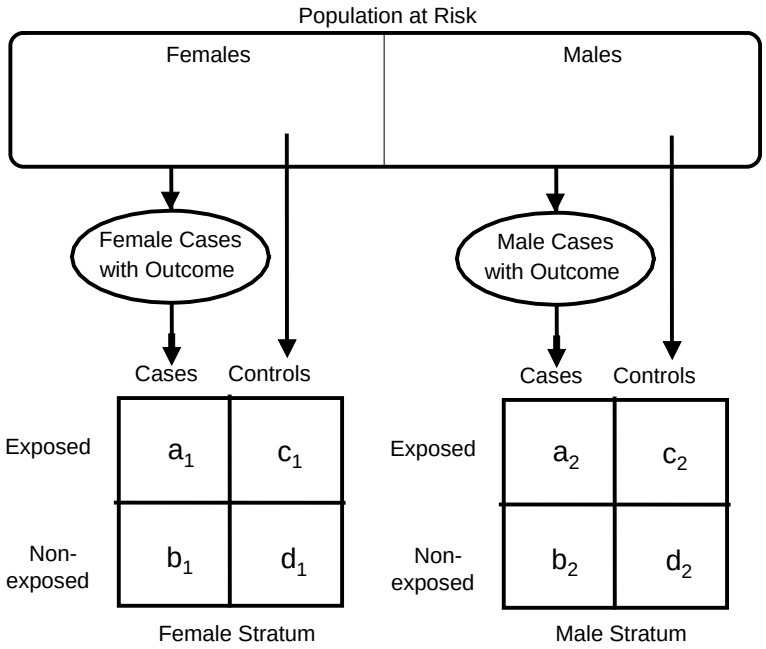
$$\begin{aligned} IRR = IR_1/IR_0 &= (a/PT_1) \div (b/PT_0) \\ &= (a/b) \div (PT_1/PT_0) \\ &= (a/b) \div (c/d) \\ &= ad/bc \quad \text{known also as the cross-product or odds ratio (OR).} \end{aligned}$$

Miettinen's clarification of the case-control sampling strategy has led to new insights in epidemiologic study design, including a new stable of designs known as the case-distribution studies. These may have no control subjects at all in the traditional sense. The following is a list of implications of Miettinen's reformulation of case-control sampling:

- A return to the effect of the exposure contrast as the focus of aetiologic research.
- Obsolescence of Cornfield's 'rare disease assumption'.
- Distinction between study bases from cohorts, dynamic populations and cross-sections of either.
- Rational rules about eligibility of controls with disease.
- Basis for case-control studies on diseases that may occur more than once.
- Control sampling schemes for different effect measures: incidence density sampling for the IRR; case-base sampling for the CIR or PR.
- Correct sampling of controls in hospital-based case-control studies: 2º vs 1º study base.
- Clarification of the nature of matching in case-control designs.
- New stable of case-distribution studies, which do not have separate controls in the traditional sense: eg, case-crossover study, case-time-control study and others.

4.2 Individually Matched Case-Control Designs

It is useful to revise another case-control topic by way of foundation principles for the understanding of case distribution designs, and this is the nature of matching in case-control studies. Here the population at risk, also known as the ***study base***, is partitioned into a number of separate strata defined by possible values of the matching factor(s). With category matching, the groups are kept fairly broad. Take the simplest possible form of category matching which involves only one matching factor with two possible values; eg, gender. For female cases, controls are sampled from the female stratum within the population at risk; and a similar sampling rule applies for male cases (see diagram below). The study proceeds as if it were two separate studies running in parallel, one in females and the other in males. It is only when the odds ratios from the two strata are pooled into a weighted average that the results from the two strata are combined, much the same as a meta-analysis of two separate case-control studies.



$OR_1 = (a_1d_1)/(b_1c_1)$ $OR_2 = (a_2d_2)/(b_2c_2)$

Pooled OR = $\Sigma w_i OR_i / \Sigma w_i$

In an individually matched case-control study, the values of matching factors are very finely graduated such that each case arises from a particularly narrow stratum within the population at risk, and it together with its matching control(s) are likely to be the only information drawn from that stratum into the study data. In the simplest situation of one individually matched control per case, each case-control pair may be viewed as forming a unique stratum within the data set. Thus for a dichotomous exposure, each stratum may yield one of only four possible 2×2 table configurations:

	Case	Control
Exposed	1	1
Non-exposed	0	0

Case & Control Exposed = Type $v_{1.1}$

	Case	Control
Exposed	1	0
Non-exposed	0	1

Case Exposed, Control Non-exposed = Type $v_{1.0}$

	Case	Control
Exposed	0	1
Non-exposed	1	0

Case Non-exposed, Control Exposed = Type $v_{0.1}$

	Case	Control
Exposed	0	0
Non-exposed	1	1

Case & Control Non-Exposed = Type $v_{0.0}$

Because the Mantel-Haenszel (M-H) estimator of a pooled odds ratio takes the form of:

$$OR = \sum_{i=1}^{k} \frac{a_i d_i}{T_i} / \sum_{i=1}^{k} \frac{b_i c_i}{T_i}$$

where for each of k strata, a_i, b_i, c_i and d_i are the cell entries and T_i is the table total, it may be seen that each table type $v_{1.0}$ contributes one unit to the numerator and each table type $v_{0.1}$ contributes one unit to the denominator of the M-H estimator, whereas the contributions of the other two table configurations amount to zero. This is the basis for the formation of the odds ratio as $v_{1.0}/v_{0.1}$ when data from an individually one-on-one matched case control study are presented as number of pairs in the following format:

		Control	
		Exposed	Non-exposed
Case	Exposed	$v_{1.1}$	$v_{1.0}$
	Non-exposed	$v_{0.1}$	$v_{0.0}$

As shown by Miettinen in the 1960s, the M-H estimator of the odds ratio arising from a pair-matched design is also the conditional maximum likelihood estimator of the OR.

4.3 Case Distribution Designs

Case distribution designs are a form of case-control study with the distinguishing feature of not using separate control subjects drawn from the population at risk to represent its distribution according to exposure status. They make use of the principle that the function of controls in a case-control study is to use c/d to estimate the ratio of PT_1/PT_0 or N_1/N_0. In fact, any method that produces a valid estimate of the distribution of the population at risk according to exposure status can provide the basis for effect measurement. For example, if we believe that the distribution of blood group in a population is unrelated to gender, there is no reason why a case-control study of the effect of blood group on the aetiology of prostate cancer could not use women as controls rather than men, even though it is not biologically possible for a woman to become a case.

Initially five different designs of case distribution studies were described:

(1) ***case-crossover study*** – Maclure's 1991 *Myocardial Infarction Onset Study* in which cases were used as their own controls to study the effect on MI risk of physical exertion and other exposures with transient effects.

(2) ***N-of-1 case-crossover study*** – described by Maclure in 1996 as the non-experimental equivalent of the N-of-1 randomised controlled trial, but also the minimalist version of the case-crossover study in which both the exposed and non-exposed 'population' experience arises from one individual.

(3) ***case-time-control study*** – Suissa's 1995 asthma study, which examined use of beta-agonists during 365 days before asthma hospitalisation or death compared with 365-730 days before. To account for underlying trends in natural progression of disease a reference group of general asthmatics was also enrolled and beta-agonist exposures were ascertained in similar time windows.

(4) ***case-genotype study*** – genetic association study designs described by Khoury and Flanders in 1996, which did not have separate controls, but relied upon laws of genetics to establish underlying exposure distributions in the population at risk.

(5) ***residential case-specular study*** – described by Zaffanella and others in 1998 as a method for the study of magnetic field from electric transmission wires and childhood cancer, where case residences were compared with virtual control (specular) residences created as a reflection on the other side of the street.

The first two types of study are especially valuable to the field of health services research. Brameld and Holman in 2003 have also since described and applied analogous longitudinal crossover designs: (1) ***cohort-crossover study*** and (2) ***cohort-comparison-crossover study*** – as methods for the estimation of readmission risk following surgical procedures. Similar methods had been applied in the study of DVT risk following long-haul air travel.

The principal advantage of most of these designs is that controls are particularly closely matched for potential confounders to the cases. This is readily apparent in the designs where a case serves as its own control by being observed at a different period in time. In this respect the case-crossover study, for example, is a form of pairwise individually matched case-control study. Several forms of case distribution study are well suited to the use of linked administrative health data to study short-term adverse events, such as in medication safety research or post-procedural risks. A common problem with administrative health data is insufficient information on potential confounders. The prospect of much of this problem disappearing with the close matching on potential confounders that occurs in a case distribution design has significant attractions.

For the sake of completeness, I also mention ***sequence symmetry analysis***, proposed by Petri *et al* in 1988 as a method to assess adverse drug reactions using the sequences of pairs of drugs in administrative data on prescriptions. It is a form of case-crossover analysis (specifically, a reverse-direction matched-pair case-crossover design where the control observations come after the case observations) albeit that the authors did not conceive it in those terms.

Advanced Theory Topic H:
4.4 Case-Crossover Study

The key feature of this design is that each case acts as its own control. The simplest case-crossover study is closely analogous to a matched-pair case-control design in which only one individual observed during two periods, known as the ***case window*** and the ***control window*** (rather than two individuals) contributes information to each stratum. Another way of looking at a general case-crossover design is as meta-analysis of a series of N-of-1 case-crossover studies. It requires a paradigm shift from control persons to control windows.

Necessary Conditions and Assumptions in a Case-Crossover Study

The success of the design is dependent on a number of important conditions and assumptions:

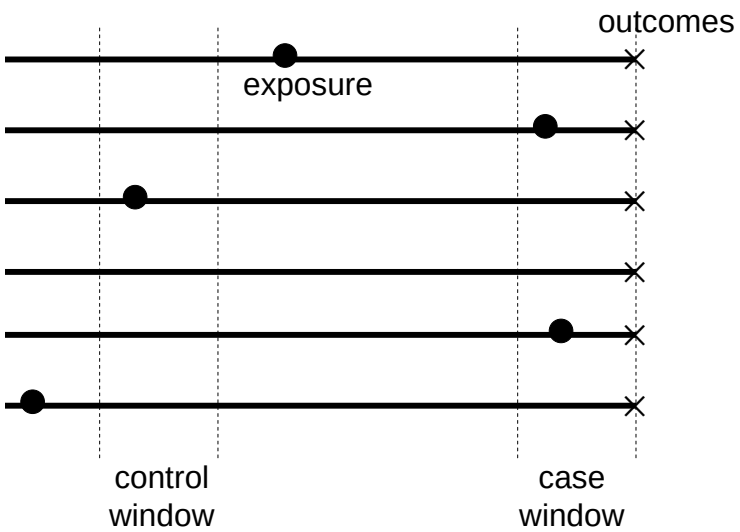
1. ***Acute caseness***: The onset of the outcome must be relatively abrupt because the design depends on a well delineated time of onset that can be related to immediately antecedent exposures.
2. ***Crossover***: At least some individuals must have crossed at least once from lower to higher exposure or vice versa. The minimum crossover is a one-time, irreversible unidirectional change in exposure status, such as death of a spouse in older people who are unlikely to remarry.
3. ***Transience of effect***: The effect of exposure must be time-limited, otherwise no difference in exposure will be expected between the case and control windows. In a case-crossover design a harmful exposure is known as a ***trigger***; and a protective exposure is a ***deterrent***.

Sampling the Study Base in a Case-Crossover Study

Duration and timing of case window: The ***hazard period*** is the interval after a trigger in which an increased risk of the outcome is caused by that trigger. The duration and timing of the case window should be such as to cover the range of possible trigger times with hazard periods overlapping the date of the outcome. Commencement of the hazard period may be delayed by an ***induction period***.

Duration and timing of the control window: The two most important attributes of the control window are that (i) it should <u>not</u> cover the range of possible trigger times with hazard periods overlapping the date of the outcome; and (ii) it should provide a representative sample of the person-time in the study base. Several methods have been described for selection of control windows to sample the distribution of exposed and non-exposed person-time in the study base:

- *Pair-matched interval approach*: A single control window is defined that is believed to meet the two criteria. For example, 25-24 hours pre-outcome as a control window vs 1-0 hours as the case window for a trigger with a very short and immediate hazard period. In the illustration below there are three discordant pairs of windows, two exposed in the case window and one in the control window. The odds ratio is therefore 2/1 = 2.0.



- *Multiple intervals approach*: Multiple control windows are defined. As in a case-control study, extending the ratio of control to case windows beyond 4-5 may not warrant the diminishing return on improved precision.
- *Usual frequency approach*: This approach contrasts exposure in the case period with the expected exposure based on each individual's usual frequency of exposure over a much longer period of time; eg, over the last year compared with a case window of the last hour before the outcome. Thus, if the usual frequency of exposure is twice per month over one year (8,766 hours), the odds of being exposed in any randomly selected one-hour period will be 24:8,742.

Note that for case and control windows that are widely separated, seasonal variation in exposure may lead to confounding by seasonality. In this circumstance, matching of the case and control windows by season may be achieved either by moving them close together in the same season of the same year, or by separating them further so as to relate to the same season in different years.

Only in rare situations, where outcome is not believed to affect future exposure, is it possible to use control windows that occur after the case windows. In those rare circumstances a balanced combination of pre- and post-outcome control windows may be effective in controlling time-trend bias (see below).

Analysis of a Case-Crossover Study

This takes the same form as the analysis of a matched-pair or R-to-1 matched case-control study. For a paired design, discordant pairs of case and control windows are enumerated and the Mantel-Haenszel odds ratio is used as an estimate of effect. *Modern Epidemiology* (formula 15-24 in the 2nd Edition) provides a variance of the log of the M-H odds ratio for use calculating confidence limits. Alternatively, analysis can be undertaken by conditional logistic regression.

Deep vein thrombosis and air travel: record linkage study

C W Kelman, M A Kortt, N G Becker, Z Li, J D Mathews, C S Guest, C D J Holman

Abstract

Objective To investigate the time relations between long haul air travel and venous thromboembolism.
Design Record linkage study using the case crossover approach.
Setting Western Australia.
Participants 5408 patients admitted to hospital with venous thromboembolism and matched with data for arrivals of international flights during 1981-99.
Results The risk of venous thromboembolism is increased for only two weeks after a long haul flight; 46 Australian citizens and 200 non-Australian citizens had an episode of venous thromboembolism during this so called hazard period. The relative risk during this period for Australian citizens was 4.17 (95% confidence interval, 2.94 to 5.40), with 76% of cases (n=35) attributable to the preceding flight. A "healthy traveller" effect was observed, particularly for Australian citizens.
Conclusions The annual risk of venous thromboembolism is increased by 12% if one long haul flight is taken yearly. The average risk of death from flight related venous thromboembolism is small compared with that from motor vehicle crashes and injuries at work. The individual risk of death from flight related venous thromboembolism for people with certain pre-existing medical conditions is, however, likely to be greater than the average risk of 1 per 2 million for passengers arriving from a flight. Airlines and health authorities should continue to advise passengers on how to minimise risk.

pulmonary embolism after long flights has brought the issue to public attention.

The incidence of venous thromboembolism ranges from 1000-2000 per million person years for deep vein thrombosis and 500-1000 per million person years for pulmonary embolism. The incidence of venous thromboembolism increases with age. Around 1-2% of venous thromboembolic events result in death.[7 8] The post-thrombotic syndrome—a long term disabling consequence of deep vein thrombosis—develops in 10-30% of patients with venous thromboembolism.[9]

International air travel has increased to around 1.56 billion person trips each year.[10] At any one time an estimated 4000 Australians are on international flights, and more than 30 000 make short domestic flights each day.

Since 1970, Australia has kept electronic data on arrivals and departures of international travellers. The state of Western Australia uses record linkage under well developed protocols to protect patient privacy.[11] Most Western Australian residents live in Perth, and flight times from there to other major airports are long. We investigated the relation between international air travel and venous thromboembolism by linking Western Australian hospital data with records on air travel.

Participants and methods

Data included coded personal identifiers, age, sex, arrival and departure dates, and nationality of the traveller (Australian citizen or non-Australian citizen). Non-Australian citizens included temporary residents

Commonwealth Department of Health and Ageing, GPO Box 9848, Canberra, ACT 2601, Australia
C W Kelman *medical adviser*
M A Kortt *assistant director*
J D Mathews *deputy chief medical officer*
National Centre for Epidemiology and Population Health, Australian National University, Canberra, ACT 0200
N G Becker *professor of biostatistics*
Z Li *postdoctoral fellow*
C S Guest *visiting fellow*
School of Population Health, University of Western Australia, Perth, WA 6009, Australia
C D J Holman *chair in public health*
Correspondence to: C W Kelman christopher.kelman@health.gov.au

bmj.com 2003;327:1072

Advanced Theory Topic I:
4.5 Case-Time-Control Study

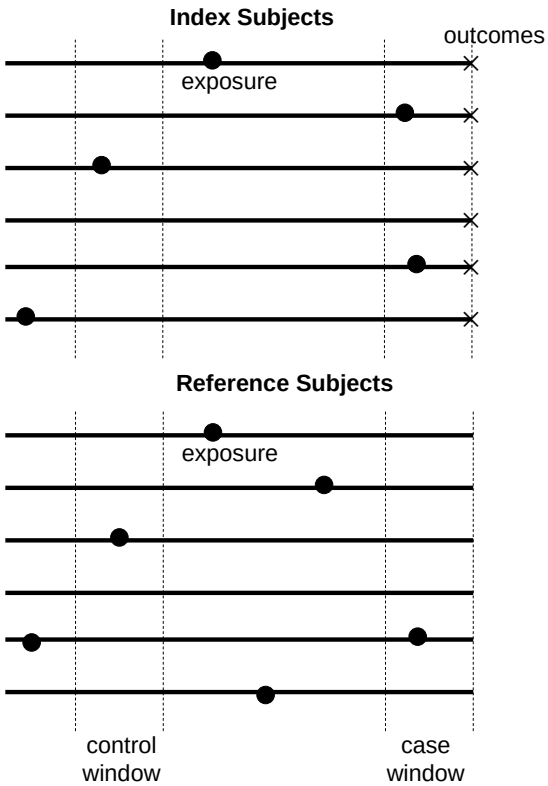
The case-time-control design is motivated by one of the limitations of the case-crossover design when applied to relatively long or widely separated case and control windows. The problem is known as ***time-trend bias***. This may occur if there are correlated underlying time trends in both exposure and outcome that are independent of the effect of exposure on outcome. This problem is especially common in studies of adverse outcomes of disease interventions such as pharmacotherapies, where confounding may be driven by underlying time trends in the propensity for treatment to be given due to:

- changes in medical practice concerning adoption of the intervention;
- changes in medical practice concerning the dosage of intervention;
- protopathic bias due to worsening in the severity of the patient's illness;
- protopathic bias due to increasing comorbidity in the patient.

Any of these may cause a systematic difference in propensity for exposure in the case window relative to the control window.

Structure and Analysis of the Case-Time-Control Design

The case-time-control design aims to control time-trend bias by adding a second group of ***reference subjects***, on whom exposure observations are made in case and control windows, the same as they are in the ***index subjects***. The design is illustrated as follows:



As explained by Suissa in 1995 and again in 1998, the analysis of data from a case-time-control design necessitates the use of a conditional logistic regression model with an interaction term. In practice, the index and reference subjects are treated as providing matched pairs of case window and control window exposure observations. If the variable E is a 0,1 binary indicator of exposure in each window and the variable G is a 0,1 binary indicator of index subject status, then in the conditional logistic regression model:

$$Ln(odds\ of\ caseness) = \beta_0 + \beta_1 E + \beta_2 E.G$$

the antilog of coefficient β_2 will provide an estimate of the effect of exposure on the OR over and above the time period effect. This situation arises because only the time period effect can account for any apparent exposure effect in reference subjects (β_1), whereas both the time period effect and the true, independent effect of exposure accounts for the total apparent exposure effect in index subjects ($\beta_1 + \beta_2$). The model allows β_2 to be estimated independently of β_1.

Limitations of the Case-Time Control Design

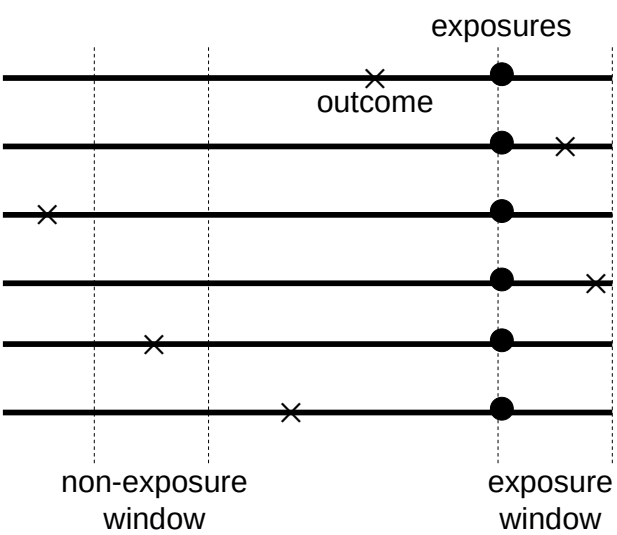
The most important limitation of the design arises from the necessary assumption that the period effect, separate from the true exposure effect, is the same in index and reference subjects. This assumption may be violated, for example, if the outcome (eg, death) leads to index subjects being observed during a period when severity of illness is worsening at a faster rate than in the reference subjects. As pointed out by Greenland in 1996, under such circumstances the adjustment for the time-trend bias can introduce a new bias, which in relatively extreme circumstances can exceed the original time-trend bias itself.

Advanced Theory Topic J:
4.6 Crossover Longitudinal Designs

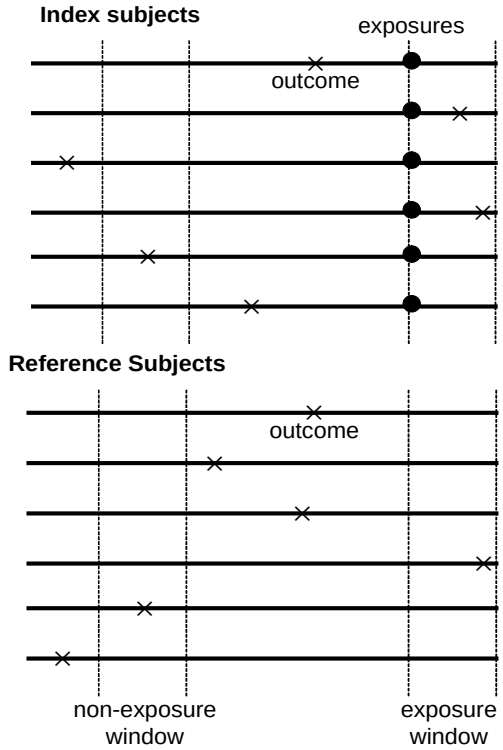
The practical distinction between a case-crossover study and a cohort-crossover study is that subjects in the first design are selected on the basis of caseness and in the second design they are selected on the basis of exposure. However, despite the analogy, the cohort-crossover is not a form of case distribution design.

A cohort-crossover design is not a case-control sampling strategy because the relative sizes of the ***exposure window*** and ***non-exposure window***, being set by the investigator, are not used to estimate the distribution of exposure status in the study base. Rather, it may be more helpful to view the design as a cohort study with a correlated data structure, because each individual is represented twice: once as an exposed subject and once as an unexposed subject. Follow-up commences at the start of each exposure/non-exposure window and either cumulative incidence or incidence rate effect measures may be derived. Unlike the case-crossover design, the cohort-crossover design is capable of estimating both absolute and relative effect measures.

Cohort-Crossover Design



Cohort-Comparison-Crossover Design



It is not essential to the cohort-crossover design that the effect of exposure is transient. However, if the exposure effect is transient, it may be possible to control any time-trend bias by using a balanced selection of non-exposure windows both pre- and post- the exposure window. Otherwise, it is possible to perform a cohort-comparison-crossover study, which introduces a reference group to account for period effects. The reference subjects undergo outcome ascertainment in the same non-exposure and exposure windows as the index subjects. The analysis requires attention to the correlated data structure using ***robust variance estimation*** procedures such as the ***Huber-White method***. Analogous to the case-time-control study, estimation of the exposure effect above and beyond the period effect is possible using a regression model term for the interaction between exposure and index/reference subject status. According to the nature of the effect measure and whether the outcome may occur more than once in a single window, the appropriate regression models are Cox regression, Poisson regression or logistic regression.

4.7 Self-Controlled Case-series

A newer research methodology using cases as their own controls is the self-controlled case-series design, which has become increasingly popular in the last 5+ years, especially in the area of pharmacoepidemiology. This method is similar to a case-crossover study in that is only includes people with the outcome of interest, comparing the risk of outcomes in different exposure periods but specifically comparing outcomes in exposed and non-exposed periods. It also uses all person-time available rather than sampling selected control windows. The study design allows for either single or cumulative events to be examined in the different exposure windows, which differentiates it from a case-crossover study. In the latter scenarios conditional Poisson regression is the primary analytical approach. Time varying covariates can also then be adjusted in the analysis. The general structure of the self-controlled case-series design is shown below.

A comparison of selected features of the case-crossover and self-controlled case-series study designs is included in the table below.

Method	Effect Measure	Study Population	Data used in analysis	Main assumptions
Self-controlled case series	IRR	Case patients who experienced at least one event during each observation period	Every exposure and every event for each case patient	Events do not alter the probability of subsequent exposure and events
Case-crossover	Exposure OR (approximating IRR)	Case patients who experienced at least one event during each observation period	Every exposure and the first event for each case patient	There are no time-trends in the occurrence of exposure.

References

Cadarette SM, Maclure M, Delaney JAC *et al*. **Control yourself: ISE-endorsed guidance in the application of self-controlled study designs in pharmacoepidemiology.** *Pharmacoepidemiol Drug Saf* 2020; 30: 671-84.

Cornfield J. **A method of estimatng comparative rates from clinical data: applications to cáncer of the breast, lung, and cérvix.** *JNCI* 1951; 11: 1269-1275.

Figueiras A, Carracedo-Martinez E, Saez M, *et al*. **Analysis of case-crossover designs using longitudinal approaches : A simulation study.** *Epidemiol* 2005; 16: 239-246.

Greenland S. **Confounding and exposure trends in case-crossover and case-time-control designs.** *Epidemiol* 1996; 7: 231-9.

Greenland S. **A unified approach to the analysis of case-distribution (case-only) studies.** *Stats Med* 1999; 18: 1-15.

Hallas J, Pottegard A**. Use of self-controlled designs in pharmacoepidemiology.** *J Int Med* 2014; 275: 581-9.

Hernandez-Diaz S, Hernan MA, Meyer K, *et al*. **Case-crossover and case-time-control designs in birth defects epidemiology.** *Am J Epidemiol* 2003; 158: 385-391.

Kelman CW, Kortt MA, Becker NG, *et al*. **Deep vein thrombosis and air travel: record linkage study.** *BMJ* 2003; 327: 1072-1076.

Khoury MJ, Flaunders WD. **Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls?** *Am J Epidemiol* 1996; 144: 207-13.

Kleinbaum DG, Kupper LL, Morgenstern H. **Epidemiologic Research: Principles and Quantitative Methods.** Belmont, California: Wadsworth, 1982

Maclure M, Mittleman MA. **Should we use a case-crossover design?** *Ann Rev Public Health* 2000; 21: 193-221.

Maclure M, Mittleman MA. **N-of-1 case-crossover design.** *Am J Epidemiol* 1996; 143: S28.

Maclure M. **The case-crossover design: a method for studying transient effects on the risk of acute events.** *Am J Epidemiol* 1991; 133: 144-153.

Maclure M, Fireman B, Nelson JC *et al*. **When should case-only designs be used for safety monitoring of medical products?** *Pharmacoepidemiol Drug Saf* 2012; 21: 50-61.

Miettinen OS. **The matched pair design in the case of all-or-none responses.** *Biometrics* 1969; 22: 339-352.

Miettinen OS. **Estimability and estimation in case-referent studies.** *Am J Epidemiol* 1976: 103: 226-235.

Miettinen OS. **Design options in epidemiologic research.** *Scan J Work Env Health* 1982; 8 (suppl. 1): 7-14.

Miettinen OS. **The "case-control study": valid selection of subjects.** *J Chron Dis* 1985; 543-548.

Miettinen OS. **Theoretical Epidemiology. Principles of Occurrence Research in Medicine.** New York: Wiley, 1985.

Mittleman MA, Maclure M, Robins J. **Control sampling strategies for case-crossover studies: an assessment of relative efficiency.** *Am J Epidemiol* 1995; 142: 91-98.

Petri H, DeVet HCW, Naus J, *et al*. **Prescription sequence analysis: a new and fast method for assessing certain adverse reactions of prescription drugs in large populations.** *Stat Med* 1988; 7: 1171-5.

Schneeweiss S, Sturmer T, Maclure M. **Case-crossover and Case-time-control designs as alternatives in pharmacoepidemiologic research.** *Pharmacoepidemiology and Drug Safety* 1997; 6 Suppl3: S51-S59.

Schneider MF, Gange SJ, *et al*. **Application of case-crossover and case-time-control study designs in analyses of time-varying predictors of T-cell homeostasis failure.** *Ann Epidemiol* 2005; 15: 137-44.

Sorock GS, Lombardi DA, Gabel CL, *et al*. **Cass-crossover studies of occupational trauma: methodological caveats.** *Injury Prevention* 2001; 7: i38-i42.

Suissa S. **The case-time-control design.** *Epidemiol* 1995; 6: 248-253.

Suissa S. **The case-time-control design: further assumptions and conditions.** *Epidemiol* 1998; 9: 441-5.

A comparison of estimators from self-controlled case series, case-crossover design, and sequence symmetry analysis for pharmacoepidemiological studies. BMC Medical Research Methodology 2018; 18(4): DOI 10.1186/s12874-017-0457-7.

Zaffanella LE, Savitz DA, Greenland S, *et al*. **The residential case-specular method to study wire codes, magnetic field, and disease.** *Epidemiol* 1998; 9: 6-7.

TRAINING SESSION 4:
IMPLEMENTING CASE DISTRIBUTION DESIGNS

IMPORTANT NOTICE

The practice data sets provided for training purposes in this course are fictitious. All results based on analysis of these fictitious data are scientifically useless. The course organisers and instructor take no responsibility for any misuse of the spurious results arising from the analysis of these fictitious data sets.

Background to Training Session 4

The aim of this training session is to give you hands-on experience in the use of linked administrative data to implement a case-crossover study and a case-time-control study. These designs are important in the field of pharmaco-epidemiology and, accordingly, the exercise centres around a classic issue in medication safety, that of the association of 'upper gastrointestinal complications' (UGC covering gastroduodenal inflammation, ulcer and haemorrhage) with exposure to a non-steroidal anti-inflammatory drug (NSAID). The exercise illustrates the potential value of linkage of prescribing information from the Commonwealth jurisdiction with adverse event information, in this case hospitalisation for UGC, from the State jurisdiction. The exercise also provides you with an opportunity to become familiar with some of the basic field methods of pharmaco-epidemiology, such as the defined daily dose methodology.

Training Session Resources

You will be making use of your updated platform file as well as your tagged and sequenced HMDSdata file and PBSdata file.

If you are using this statistical package …	*… you should use this save file.*
SPSS	Diabetes8.sav HMDSdata3.sav PBSdata3.sav
SAS	Diabetes8.sas7bdat HMDSdata3.sas7bdat PBSdata3.sas7bdat
Stata	Diabetes8.dta HMDSdata3.dta PBSdata3.dta
R	Diabetes8.rda HMDSdata3.rda PBSdata3.rda

Note that the following additional HMDS diagnostic codes are relevant to this exercise:

Data Set	Coding System	Codes
HMDS	*ICD-9* and *ICD-9-CM* *ICD-10-AM*	**Upper Gastrointestinal Complications (UGC)** 531.X, 532.X, 533.X, 534.X. 535.X, 578.X K25.X, K26.X, K27.X, K28.X, K29.X, K92.0, K92.1, K92.2

WARNING: You should not rely on these codes in any analysis separate from the course as they may contain errors.

You will be making use of additional PBS drug codes, but these are explained in more detail below.

Resources for Drug Utilisation and Outcomes Research

Anatomical Therapeutic Chemical (ATC) Drug Classification

The ATC drug classification system has been maintained by the *WHO Collaborating Centre for Drug Statistics Methodology* in Oslo, Norway, since 1982 (http://www.whocc.no/atcddd/). It provides a global standard for classifying medical substances in drug utilisation and outcomes research. The WHO recommends the ATC system for international comparisons and for reporting of adverse drug reactions. ATC codes are included both in international and national drug catalogues (including the PBS in Australia) and represent a common language. Courses in the ATC drug classification and the DDD (see below) methodology are arranged annually by the WHO Centre in Oslo.

ATC structure: Drugs are classified by ATC into groups at five levels. They are divided into 14 main groups (1st level) with pharmacological/therapeutic subgroups (2nd level). The 3rd and 4th levels are chemical/pharmacological/therapeutic subgroups and the 5th level is the chemical substance. The 2nd, 3rd and 4th levels are often used to identify pharmacological subgroups when considered more appropriate than therapeutic or chemical subgroups.

Example: Classification of the oral hypoglycaemic, metformin, illustrates the structure of the code:

A	*Alimentary tract and metabolism*	*(1st level, anatomical main group)*
A10	*Drugs used in diabetes*	*(2nd level, therapeutic subgroup)*
A10B	*Oral blood glucose lowering drugs*	*(3rd level, pharmacological subgroup)*
A10BA	*Biguanides*	*(4th level, chemical subgroup)*
A10BA02	*Metformin*	*(5th level, chemical substance)*

Principles of classification: Medicinal products are classified according to the main therapeutic use of the main active ingredient based on available literature. There is only one ATC code for each pharmaceutical formulation (ie, similar ingredients, strength and pharmaceutical form). More than one ATC code is used for two or more formulations with clearly different therapeutic uses.

The Defined Daily Dose (DDD) Methodology

The DDD methodology relates closely with the ATC classification and is also overseen by the *WHO Centre*. DDDs provide a fixed unit of measurement independent of price and formulation enabling researchers to assess drug consumption trends and to compare population groups.

Definition: The DDD is the assumed average maintenance dose per day for a drug used for its main indication in adults. A DDD is only assigned to drugs that already have an ATC code and the main indication is that reflected in the ATC. When the recommended dose refers to body weight, an adult is considered to be a person of 70 kg. The DDDs assigned to combination products count the entire combination as one daily dose, regardless of the number of active ingredients.

Limitations: Researchers using the DDD system should keep the following in mind:

- Drug consumption data presented in DDDs only give a rough estimate of consumption.
- The DDD does not necessarily reflect the recommended or prescribed daily dose. Actual doses will often differ from the DDD due to individual characteristics such as age, weight and pharmacokinetic considerations. The DDD is sometimes never prescribed at all, because it is an average of two or more commonly used doses.
- Pharmaceutical products mainly intended for children (eg, mixtures, suppositories) are still assigned the DDD used for adults, but there are exceptions.
- The maintenance rather than initial dose (where different) is preferred; as is the treatment rather than prophylaxis dose, unless prophylaxis is the main indication.
- DDDs are not established for topical preparations, sera, vaccines, antineoplastic agents, allergen extracts, general and local anaesthetics and contrast media.
- DDDs are normally based on monotherapy. Exceptions to this rule are given in the guidelines.
- Various salts of a substance are usually not given different DDDs, but there are exceptions.

Uses of DDDs in Pharmaco-epidemiology

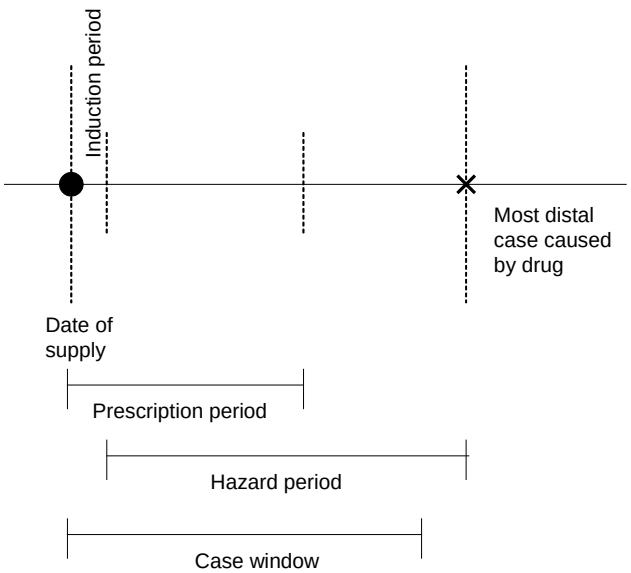
Therapeutic intensity: This is a commonly used measure of drug utilisation expressed as DDDs/1,000 people/day in national drug use statistics. Because 'day' cancels out, it may be reinterpreted as the average prevalence per 1,000 people taking the drug at the defined dose level.

Prescription period and point prevalence: Pharmaceutical administrative databases with individual level data may be used to estimate a point prevalence of drug use on a specified date based on the number of persons who have a ***prescription period*** of usage that crosses the date. Prescription period may be determined by (i) assigning a constant period to each filled prescription; or (ii) dividing the prescribed amount of drug by the DDD, thus assuming a constant level of use; ie,

$$PP = \frac{S \times Q}{DDD}$$

where PP is the prescription period, S is the strength of the product and Q is the quantity dispensed. Adjustments to point prevalence estimates can be made for non-compliance.

Prescription period in case-distribution designs: Individual prescription periods obtained from a pharmaceutical database and based on either DDDs or constant periods may be used to assess exposure status in case and control windows in case-crossover and case-time-control studies. As illustrated below, the duration and timing of the window is determined by estimates of the induction period, prescription period and hazard period. The induction and hazard periods will depend on an understanding of pharmacokinetics of the drug and the causal biology of the outcome. For example, acute hypoglycaemia following insulin administration may have an induction period of minutes and a hazard period of a few hours. Candidiasis as a complication of antibiotic therapy may have an induction period of a few days and a hazard period of some weeks. Breast cancer caused by HRT (which would not be a suitable research topic for a case distribution design) may have an induction period of several years and a hazard period of decades.



As a practical guide to defining an appropriate case/control window in pharmaco-epidemiology, a good starting point is to estimate the prescription period, extend it by a continuation to an observable upper limit of the hazard period based on an assessment of pharmacokinetics and causal biology of the outcome and, uncommonly, trim the minimum induction period from the end of the window proximal to the case or equivalent control date.

Example of Use of ATC, DDD and PBS Items Applicable to an NSAID

In this exercise we will follow the processes outlined above with respect to indomethacin, one of the NSAIDs identified by Henry *et al* as having a medium risk of UGC. The following table sets out details concerning indomethacin obtained from the ATC classification system and the PBS Schedule.

NSAID	ATC code	DDD	PBS items	Strength	Quantity
Indomethacin	M01AB01	100mg	2454	25mg Oral	100
			2459	25mg Oral	100
			2757	100mg Supp	20
			5126	25mg Oral	100
			5127	25mg Oral	100
			5128	100mg Supp	20

WARNING: You should not rely on these codes in any analysis separate from the course as they may contain errors.

The current PBS item numbers for a given ATC code may be found by consulting the current *Schedule of Pharmaceutical Benefits for Approved Pharmacists & Medical Practitioners*, which is laid out in order of ATC classification. However, typically a researcher will need to be aware not only of current PBS item numbers, but also those applicable in the past that may have been withdrawn after the commencement of the study period. A ***PBS Schedule*** is available from the Australian Department of Health & Ageing, which provides a mapping facility that generates lists of PBS items for given ATC codes or a generic name.

Multiplying the strength by quantity of the medication dispensed and dividing by the DDD results in estimation of a prescription period of 25 days (25×100/100) for oral indomethacin and 20 days (100×20/100) for suppositories. The half-life of indomethacin is about 4.5 hours, thus being largely excreted from the body within one day. UGCs are largely of an acute nature and might reasonably be expected to become clinically manifest any time from the first day of treatment (ie, same-day induction) to as much as 5 days after treatment ceases. Thus an appropriate case/control window for date of supply of oral indomethacin is 25 + 1 + 5 = 31 days, including the day of the case event. For the sake of simplicity, we will apply the same case/control window period to indomethacin suppositories, which may also cause upper gastrointestinal complications.

Overview of Exercise 4

Steps 1-2. Open HMDSdata, tag and sequence UGC records, cutdown to first-time UGC records and save as index. These are the index subjects.

Steps 3-4. Open PBSdata, tag indomethacin records, delete remainder, save as indexposure, load UGC admdate from index and delete non-UGC records.

Step 5. Assign exposure = 1 for dispdate within 31d of admdate and exposure = 2 for dispdate with 395-365d of admdate; select exposed records and delete remainder.

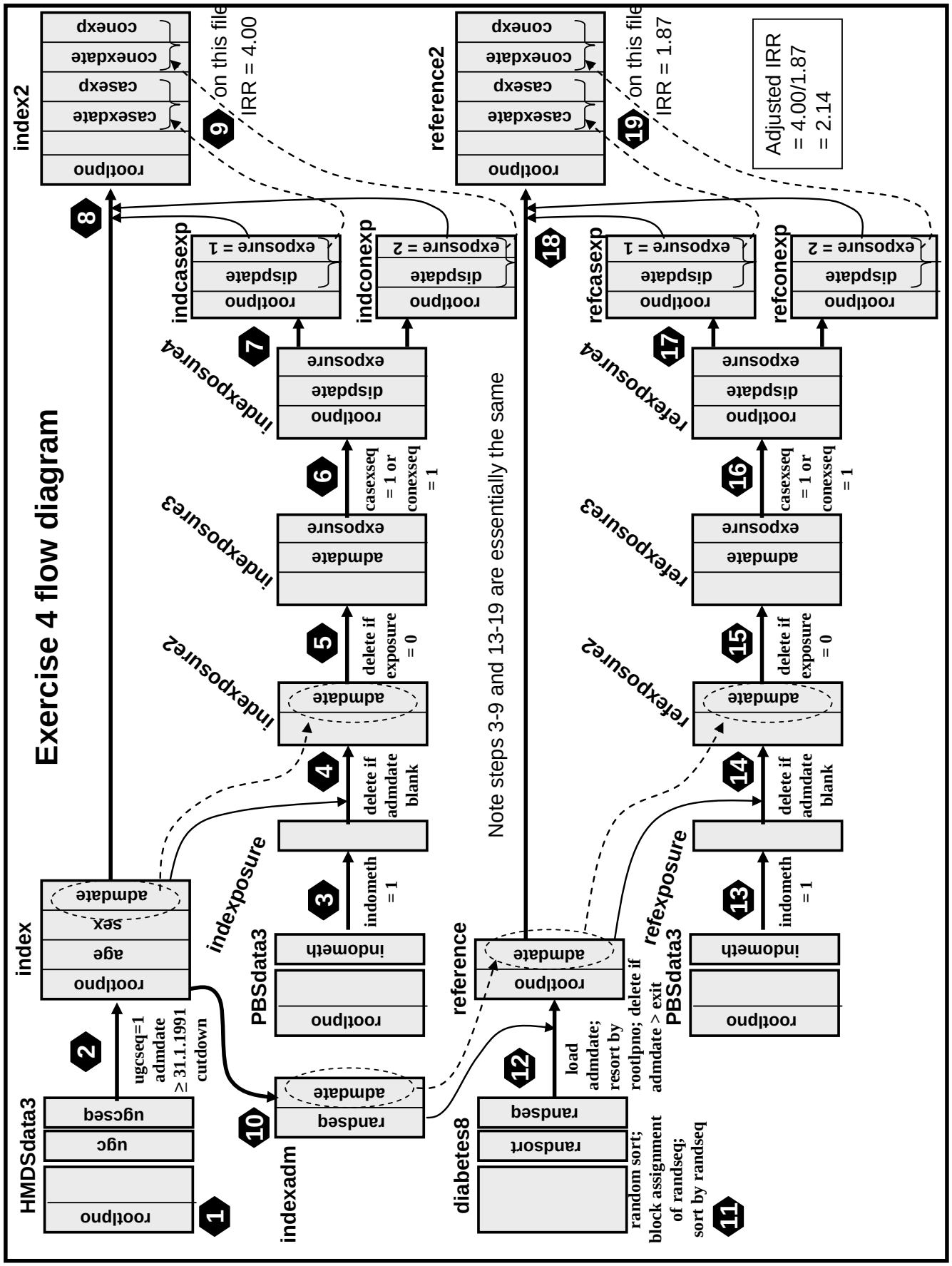
Steps 6-7. Apply 'upside-down' sequences, casexseq and conexseq to select last indomethacin exposures in case & control windows, delete remainder and resave as indexposure. Separately save the case window exposure records as indcasexp and the control window exposure records as indconexp.

Steps 8-9. Load exposure data from indcasexp and indconexp onto index file and perform matched pair analysis of case-crossover design.

Steps 10-12. Randomly assign index admdates to platform and keep diabetics who had not exited before their randomly assigned admdate (=reference).

Steps 13-17. Repeat steps 3-7 for reference subjects to ascertain exposures in case and control windows in reference subjects.

Steps 18-19. Repeat steps 9-10 to perform matched pair analysis on reference subjects, then adjust index effect measure by dividing by reference effect measure.


Exercise 4 flow diagram
HMDSdata3
rootlpno
ugc
ugcseq
1
2
ugcseq=1
admdate
≥ 31.1.1991
cutdown
index
rootlpno
age
sex
admdate
indexposure
PBSdata3
rootlpno
indometh
3
indometh
= 1
4
delete if
admdate
blank
indexposure2
admdate
5
delete if
exposure
= 0
indexposure3
admdate
exposure
6
casexseq
= 1 or
conexseq
= 1
indexposure4
rootlpno
disdate
exposure
7
indcasexp
rootlpno
disdate
exposure = 1
indconexp
rootlpno
disdate
exposure = 2
8
index2
rootlpno
casexdate
casexp
conexdate
conexp
9
on this file
IRR = 4.00
indexadm
randseq
admdate
10
diabetes8
randseq
randsort
11
random sort;
block assignment
of randseq;
sort by randseq
12
load
admdate;
resort by
rootlpno; delete if
admdate > exit
reference
rootlpno
admdate
Note steps 3-9 and 13-19 are essentially the same
refexposure
PBSdata3
rootlpno
indometh
13
indometh
= 1
14
delete if
admdate
blank
refexposure2
admdate
15
delete if
exposure
= 0
refexposure3
admdate
exposure
16
casexseq
= 1 or
conexseq
= 1
refexposure4
rootlpno
disdate
exposure
17
refcasexp
rootlpno
disdate
exposure = 1
refconexp
rootlpno
disdate
exposure = 2
18
reference2
rootlpno
casexdate
casexp
conexdate
conexp
19
on this file
IRR = 1.87
Adjusted IRR
= 4.00/1.87
= 2.14

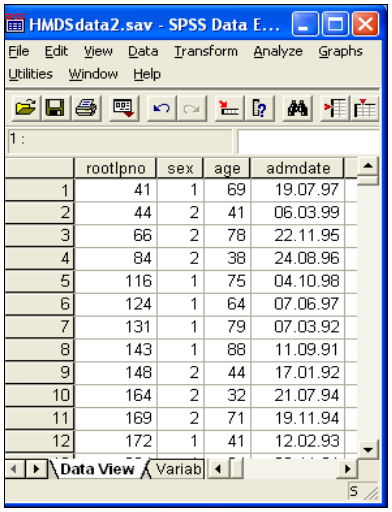
Exercise Instructions

Identification of Index Subjects Hospitalised for Upper Gastrointestinal Complications

1. Open your updated HMDSdata3 file and tag all records that have a UGC diagnosis code as the principal condition treated during hospital stay (ie, diag1), using a new variable, ugc. Set ugc to '1' for a tagged record and '0' otherwise.

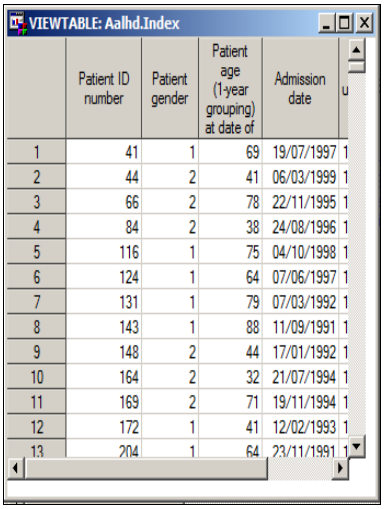
2. Create an ugcseq variable that numbers only the tagged UGC records within each block. Select all first-tagged records admitted on or after 31 January 1991 (ugcseq = 1 and admdate ≥ 31.1.1991) and delete the remainder. Note that transfers are ignored here on the basis that the admission date of the record in a transfer set bearing a UGC principal diagnosis is best regarded at the case incidence date. Delete all variables in this file except rootlpno, age, sex and admdate. Your file of 970 index subjects should appear like the screen shown below. Save this file with the name index.

SPSS



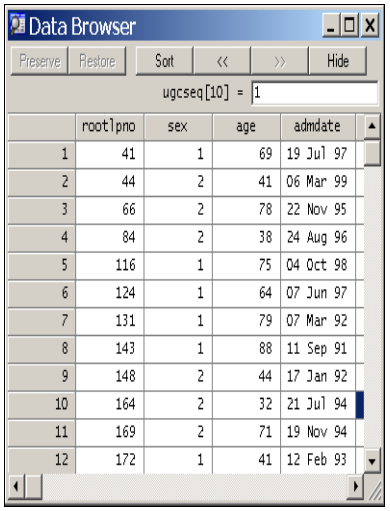
	rootlpno	sex	age	admdate
1	41	1	69	19.07.97
2	44	2	41	06.03.99
3	66	2	78	22.11.95
4	84	2	38	24.08.96
5	116	1	75	04.10.98
6	124	1	64	07.06.97
7	131	1	79	07.03.92
8	143	1	88	11.09.91
9	148	2	44	17.01.92
10	164	2	32	21.07.94
11	169	2	71	19.11.94
12	172	1	41	12.02.93

SAS



	Patient ID number	Patient gender	Patient age (1-year grouping) at date of	Admission date	u
1	41	1	69	19/07/1997	1
2	44	2	41	06/03/1999	1
3	66	2	78	22/11/1995	1
4	84	2	38	24/08/1996	1
5	116	1	75	04/10/1998	1
6	124	1	64	07/06/1997	1
7	131	1	79	07/03/1992	1
8	143	1	88	11/09/1991	1
9	148	2	44	17/01/1992	1
10	164	2	32	21/07/1994	1
11	169	2	71	19/11/1994	1
12	172	1	41	12/02/1993	1
13	204	1	64	23/11/1991	1

Stata



	rootlpno	sex	age	admdate
1	41	1	69	19 Jul 97
2	44	2	41	06 Mar 99
3	66	2	78	22 Nov 95
4	84	2	38	24 Aug 96
5	116	1	75	04 Oct 98
6	124	1	64	07 Jun 97
7	131	1	79	07 Mar 92
8	143	1	88	11 Sep 91
9	148	2	44	17 Jan 92
10	164	2	32	21 Jul 94
11	169	2	71	19 Nov 94
12	172	1	41	12 Feb 93

Ascertainment of Indomethacin Exposure

3. Open your updated PBSdata3 file and tag all records with an indomethacin PBS item number using a new variable, indometh. Set indometh to '1' for a tagged record and '0' otherwise. Select all of the tagged PBS records and delete the remaining records. You should have 5,351 records. Save this file as indexposure.

4. Externally load the index admission dates using the index file as a look-up table keyed by rootlpno. The admdate variable provides a reference point for identifying indomethacin scripts that fall within the case or control windows. You should delete all the PBS records where admdate remains blank as these records do not relate to an index subject with UGC. **HINT for Stata users:** You will have to open the index file and do a stable rootlpno sort first. Save as indexposure2.

5. Create a new variable, exposure, and set it to '0' if dispdate falls neither in the case nor control window; '1' if dispdate falls within the case window of 31 days up to and including admdate; and '2' if it falls within the control window from 395 to 365 days, inclusive, prior to admdate. Select all record with exposure equal to '1' or '2' and delete the remaining records. **HINT:** Initially assign the difference in days between admdate and disdate to exposure; then recode exposure into the three categories using the correct ranges of the difference. Save as indexposure3.

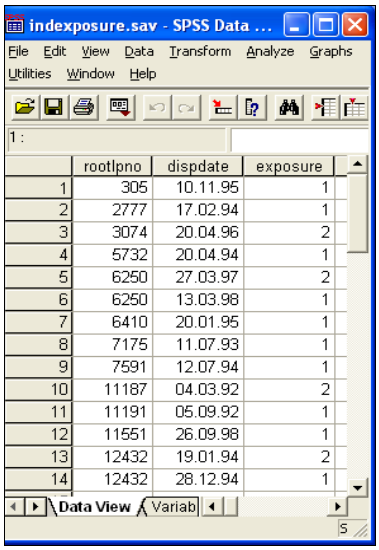
6. **SPSS and SAS users:** Turn the file up-side-down by sorting on fileseq in descending order and add two new sequence variables, casexseq and conexseq, to number separately the scripts falling within the case window and control window of each index subject. Turn the file right-way-up again. Select the records of last-time exposure in either a case or control window (ie, casexseq = 1 or conexseq = 1) and delete the remaining records. Delete all of the variables in the file except rootlpno, dispdate and exposure. Check that your file of 14 records looks like the one below and save it as indexposure4.

   Stata users: Complete a stable rootlpno sort and, using the by rootlpno: egen command, create two new sequence variables, casexseq and conexseq, to number separately the scripts falling within the case window and control window of each index subject. Create temporary variables to help select the records of last-time exposure in either a case or control window (ie, maximum value of or conexseq by rootlpno) and delete the remaining records. Delete all of the variables in the file except rootlpno, dispdate and exposure. Check that your file of 14 records looks like the one below and use the save , replace command to re-save it as indexposure4.

   R users: The syntax solutions manual illustrates how the sequence variables can be created by first creating temporary files, creating the sequences within those files, and merging them back on. However, this is just one way to do it.

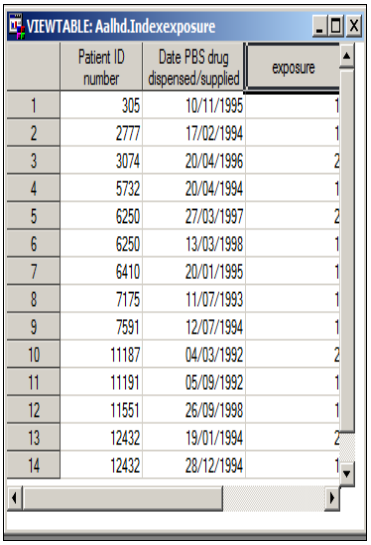
   Note that these steps could have been performed manually given the file is so small, but in other contexts a syntactical approach will be essential.

SPSS



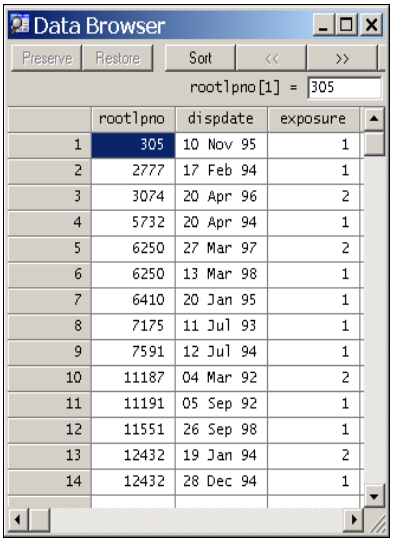
	rootlpno	dispdate	exposure
1	305	10.11.95	1
2	2777	17.02.94	1
3	3074	20.04.96	2
4	5732	20.04.94	1
5	6250	27.03.97	2
6	6250	13.03.98	1
7	6410	20.01.95	1
8	7175	11.07.93	1
9	7591	12.07.94	1
10	11187	04.03.92	2
11	11191	05.09.92	1
12	11551	26.09.98	1
13	12432	19.01.94	2
14	12432	28.12.94	1

SAS



	Patient ID number	Date PBS drug dispensed/supplied	exposure
1	305	10/11/1995	1
2	2777	17/02/1994	1
3	3074	20/04/1996	2
4	5732	20/04/1994	1
5	6250	27/03/1997	2
6	6250	13/03/1998	1
7	6410	20/01/1995	1
8	7175	11/07/1993	1
9	7591	12/07/1994	1
10	11187	04/03/1992	2
11	11191	05/09/1992	1
12	11551	26/09/1998	1
13	12432	19/01/1994	2
14	12432	28/12/1994	1

Stata



	rootlpno	dispdate	exposure
1	305	10 Nov 95	1
2	2777	17 Feb 94	1
3	3074	20 Apr 96	2
4	5732	20 Apr 94	1
5	6250	27 Mar 97	2
6	6250	13 Mar 98	1
7	6410	20 Jan 95	1
8	7175	11 Jul 93	1
9	7591	12 Jul 94	1
10	11187	04 Mar 92	2
11	11191	05 Sep 92	1
12	11551	26 Sep 98	1
13	12432	19 Jan 94	2
14	12432	28 Dec 94	1

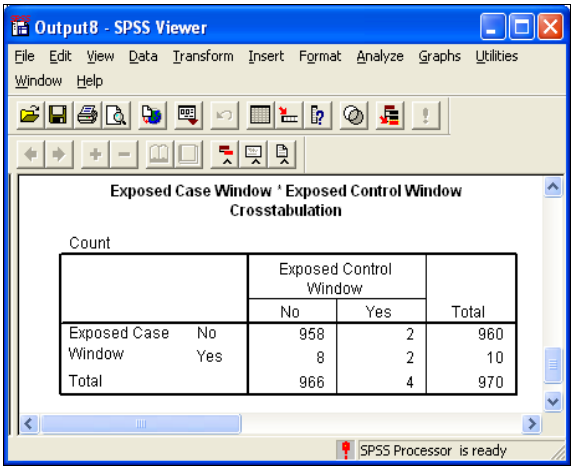
7. Select the exposed case records and save these as indcasexp. Reopen the indexposure4 file and select the exposed control records and save these as indconexp.

Analysis of the Case-Crossover Design

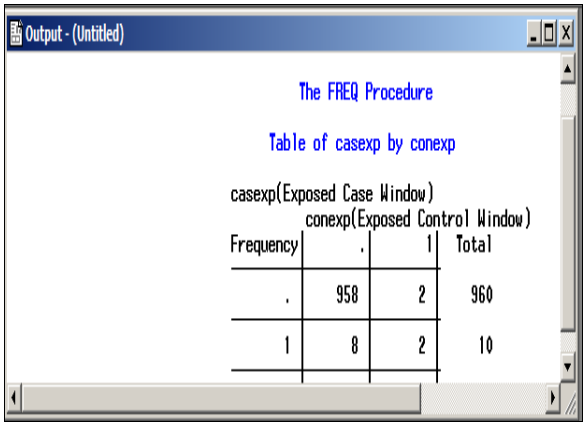
8. To create a file suitable for tabular analysis of case-crossover pairs, open the index file again, which contains all of your index subjects. Externally load the dispdate and exposure variables using the indcasexp file as a look-up table keyed by rootlpno. Rename dispdate to casexdat and rename exposure to casexp, also recoding the values of the latter so that '0' = no case window exposure and '1' = presence of case window exposure. Now externally load the dispdate and exposure variables (again, but from a different file) using the indconexp file as a look-up table keyed by rootlpno. This time rename dispdate to conexdat and rename exposure to conexp, also recoding the values of the latter so that '0' = no control window exposure and '1' = presence of control window exposure. Save your subject file as index2, which is now ready for tabular matched-pair analysis.

9. Perform a crosstabulation of casexp by conexp. Your results should appear like those opposite. What is your estimate of the incidence rate ratio of first-time UGC within 31 days from and including the dispensing date of an indomethacin script compared with at other times? Note that steps (8) and (9) in this exercise would be different if you wished to prepare an index subject file for conditional logistic regression analysis. Most such regression programs require a file structure in which case and control information are separate and usually adjacent records in the file; eg, case record followed by matching control records.

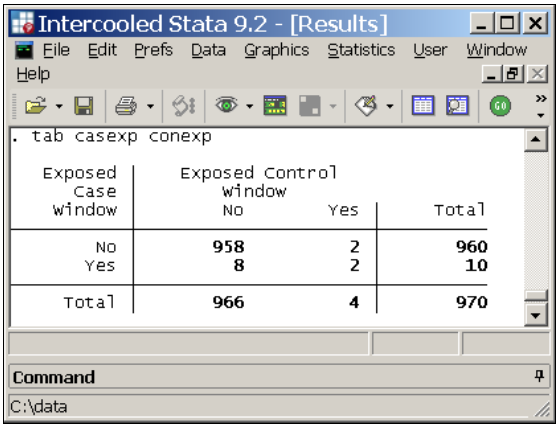
SPSS



SAS



Stata



Creation of a File of Reference Subjects

10. For reference subjects we will use the study base of person-time in the 10,675 diabetics. These must be randomly assigned the admission dates of the 970 index subjects in order to set up case and control windows. You learnt how to perform a random assignment of a variable's values from one file to another in training exercise 3, and the next steps closely follow that process. With the index2 file still open, set the seed in your random number generator to 2,500,000 and assign a random variable, randsort, using a uniform random distribution from 0 to 1. Sort the file in ascending order of randsort. Create a new file sequence variable in the usual manner to number the records from 1 to 970, but call it randseq. Delete all of the variables except randseq and admdate, then save this cut-down file of index admission dates as indexadm.
11. Open your updated diabetes8 platform file. Set the random number generator seed to 3,500,000; assign randsort as a uniform random number between 0 and 1; and sort the file in ascending order of randsort. A randseq variable is now created, but the method is different here than in the file of index subjects. The first 970 records are numbered sequentially from 1 to 970; the counter is then reset to 1 and numbering begins again such that the 1,940th record is again numbered 970; the last 5 records are numbered 1 to 5.
12. Sort the diabetes file by ascending order of randseq. Load admdate onto each record using indexadm as an external look-up table keyed by randseq. Resort the file by rootlpno. Delete the records of all diabetes where the exit date occurs before the randomly assigned admdate (in my file this excluded 977 leaving 9,698 reference subjects; in Stata the corresponding figures were 965 and 9,710). Delete all of the variables except rootlpno and admdate and save the file as reference.

The remaining steps repeat those undertaken on index subjects, so you should use your syntax from before with minor adjustments. Use of random numbers may mean my results differ from yours.

Ascertainment of Indomethacin Exposure in Reference Subjects

13. Open your PBSdata3 file and again tag all records with an indomethacin PBS item number using indometh. Select all of the tagged PBS records and delete the remaining records. **Stata users:** Perform a stable sort by rootlpno and save the file as refexposure.
14. Load the index admission dates from the reference file as a look-up table keyed by rootlpno. Delete all PBS records where admdate remains blank (4,728 records remained; 4,828 in Stata). Save as refexposure2.
15. Create a new variable, exposure, and set it to '0', '1' or '2' as before based on the difference between dispdate and admdate. Select all records with exposure equal to '1' or '2' and delete the remaining records (leaving 98 records in my file; 81 in Stata). Save as refexposure3.
16. **SPSS and SAS users:** Turn the file up-side-down by sorting on fileseq in descending order and add the sequence variables, casexseq and conexseq as before. Turn the file right-way-up. Select the records of last-time exposure in either a case or control window (ie, casexseq = 1 or conexseq = 1) and delete the remaining records. Delete all of the variables in the file except rootlpno, dispdate and exposure. My file contained 84 records at this point. Save it with the name refexposure4.

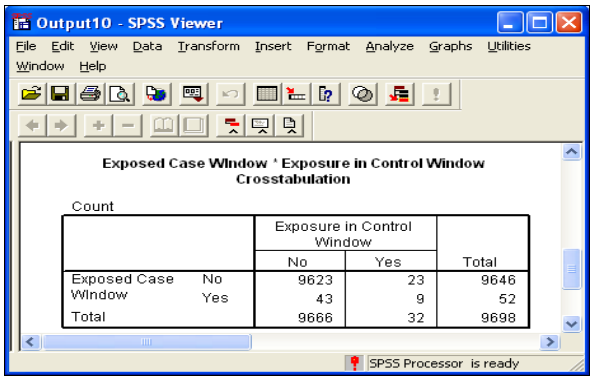
    Stata users: Add the sequence variables, casexseq and conexseq as before. Select the records of last-time exposure in either a case or control window (ie, casexseq = 1 or conexseq = 1) and delete the remaining records. Delete all of the variables in the file except rootlpno, dispdate and exposure. The Stata file contained 65 records at this point. Resave the file as refexposure4 using save, replace.

    R users: The syntax solutions manual illustrates the temporary file approach, as for Q6.
17. Select the exposed case records and save these as refcasexp. Reopen the refexposure4 file and select the exposed control records and save these as refconexp.

Analysis of the Case-Time-Control Design

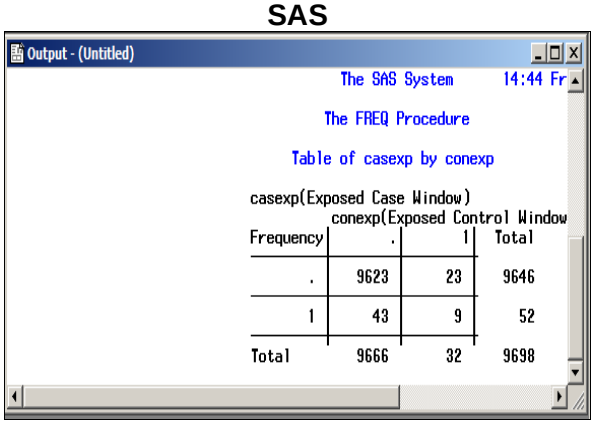
18. For the purpose of tabular analysis, we will analyse the reference subjects using a separate file to the index subjects. Note that in a regression analysis the steps would be different with the aim of combining all index and reference subjects in the one file with separate case and control records. Open the reference file again, which contains all of your reference subjects. Externally load the dispdate and exposure variables using the refcasexp file as a look-up table keyed by rootlpno. Rename dispdate to casexdat and rename exposure to casexp, also recoding the values of the latter so that '0' = no case window exposure and '1' = presence of case window exposure. Now externally load the dispdate and exposure variables using the refconexp file as a look-up table keyed by rootlpno. This time rename dispdate to conexdat and rename exposure to conexp, also recoding the values of the latter so that '0' = no control window exposure and '1' = presence of control window exposure. Save your subject file as reference2, which is now ready for tabular matched-pair analysis of the reference subjects.
19. Perform a crosstabulation of casexp by conexp. My results are shown opposite. Calculate the odds ratio in reference subjects and divide it into the odds ratio you obtained earlier in the index subjects (4.00/1.87 = 2.14). This is your estimate of the effect of first-time UGC within 31 days of the dispensing date of an indomethacin script compared with at other times, over and above any apparent effect introduced by time-trend bias.

SPSS

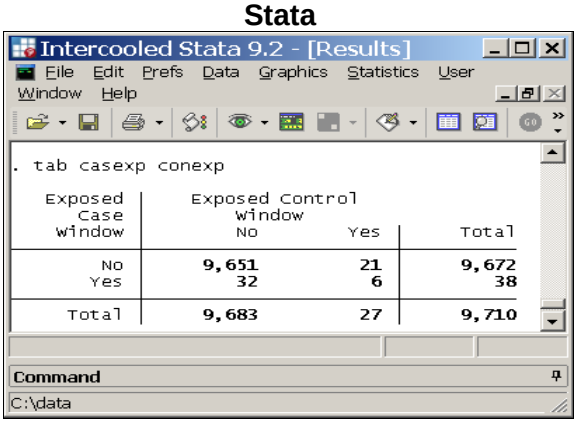


		Exposure in Control Window		
		No	Yes	Total
Exposed Case Window	No	9623	23	9646
	Yes	43	9	52
Total		9666	32	9698

SAS



Stata



References

Hallas J. **Pharmacoepidemiology – current opportunities and challenges.** *Norwegian J Epidemiol* 2001; 11: 7-12.

Hallas J, Stovring H. **Templates for analysis of individual-level prescription data.** *Basic & Clin Pharmacol & Toxicol* 2006; 98: 260-5.

Henry D, Lim LL-Y, Garcia Rodriguez LA, *et al*. **Variability in risk of gastrointestinal complications with individual nonsteroidal anti-inflammatory drugs: results of a collaborative meta-analysis.** *BMJ* 1996; 312: 1563-6.

REVIEW 4:
'CASE DISTRIBUTION DESIGNS USING LINKED DATA'

Please check that you are familiar with the following topics:

- Case-control sampling strategy and the *trohoc* fallacy ☐
- Matched case-control designs including tabular analysis ☐
- Rationale for case distribution designs ☐
- Case-crossover design ☐
- Case-time control design ☐
- Cohort-crossover and cohort Comparison-crossover designs ☐
- Self-controlled case-series design ☐
- Anatomical Therapeutic Chemical drug classification ☐
- Defined daily dose methodology ☐

Please check that you understand the meaning of the following terms:

- N-of-1 study, case-genotype study and case-specular study ☐
- Sequence symmetry analysis ☐
- Trigger and deterrent ☐
- Induction period and hazard period ☐
- Case window and control window ☐
- Index subjects and reference subjects ☐
- Time-trend bias ☐
- Exposure window and non-exposure window ☐
- Robust variance estimation ☐
- Prescription period ☐
- Uniform distribution ☐
- Sampling frame, sampling fraction, sampling distance and replacement ☐

Please check that you have demonstrated that you have acquired the following skills:

- Use of the ATC and DDD to identify PBS item codes ☐
- Estimation of prescription period and case/control window duration and timing ☐
- Implementation of a case-crossover design ☐
- Implementation of a case-time-control design ☐

5. ADVANCED TOPICS IN EXTERNAL AND INTERNAL VALIDITY OF EFFECT MEASURES

Particular emphasis has been given in this curriculum to the investigator's efforts, in order of priority, to (i) optimise internal validity through control of confounding and other biases; (ii) optimise external validity; and (iii) optimise precision, both absolutely and relative to study costs. In this final theory session we shall consider some advanced topics that revisit the first two of these considerations – the optimisation of internal and external validity. The topics selected are those that the researcher may find especially useful in health services research using linked administrative health data.

Advanced Theory Topic M:
5.1 Propensity Score Analysis

What is a Propensity Score?

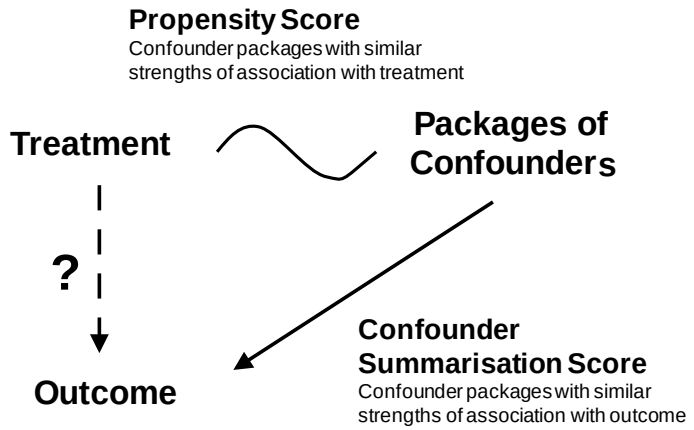
A ***propensity score*** is the conditional probability of assignment to a particular intervention or exposure given a vector of observed covariates. Mathematically, it is written as Pr[treatment | **x**] where **x** is the vector of covariates. Note that a propensity score should not be confused with the several variations of a ***confounder summarisation score*** proposed by Miettinen in the 1970s, which are akin to the conditional probability of the outcome, or a transformation of the outcome, given a vector of observed covariates; ie, Pr[outcome | x]. In practical terms a propensity score will be built from the regression coefficients of a model in which covariates predict exposure to an intervention; whereas a confounder summarisation score will be built from the regression coefficients of a model in which covariates predict the desired outcome from the intervention.

Significance of Propensity Scores

Use in the control of confounding: The advantage of propensity scores, in the way they were proposed by Rosenbaum and Rubin in the 1980s, is their ability to create a balanced distribution of measured risk factors between exposed and non-exposed subjects analogous to a matched cohort study. However, unlike the matched cohort study, control of confounding by stratification on propensity score is practicable even when the number of possible covariate combinations is very large. This is because instead of attempting to achieve a balanced design with respect to every confounder, the use of propensity score strata merely attempts to balance packages of confounders that have a similar propensity for association with treatment. Take the situation where there are 74 covariates. Even with just two value categories per covariate, matching would need to proceed with $2^{74} = 1.89 \times 10^{22}$ matching strata. Such an arrangement would be impracticable to implement due to the majority of strata containing only one exposed or one non-exposed subject. As shown by Rosenbaum and Rubin, a propensity score based on the same 74 covariates and then used as the basis for stratification or matching with only five strata is sufficient to remove over 90% of the confounding caused by the 74 covariates.

The concept of packages of confounders is an important one because this is the reality for individual subjects in research. Their confounders come as a package and it is the overall strengths of association of these packages with the propensity for treatment and the likelihood of the outcome that drives contributions to any overall confounding of treatment effect. Thus balanced designs do not require that every confounder is perfectly balanced across the treatment and non-treatment groups; but only that the confounder packages are balanced. Thus in a RCT, an apparent failure of randomisation to yield a perfectly balanced distribution of one measured confounder across the arms of the trial need not be a failure at all. The lop-sided distribution of one confounder may be offset by a lop-sided distribution in the opposite direction of one or more other measured or unmeasured confounders with the net result that the confounder packages are balanced and confounding is impossible. Propensity scores and confounder summarisation scores can only take account of packages composed of

measured confounders, but at least in that regard they can be used in various ways to impose a balance of measured confounder packages or to adjust for an imbalance through stratification or regression. The diagram below illustrates this point.



One may ask what advantage has the use of the propensity score over placement of the numerous covariates directly in a regression model in which the dependent variable is the study outcome. The answer is that the success of control of confounding by covariates directly in a regression model is dependent on the correct specification of the relationship between each covariate and the outcome. Simple linear model terms may grossly mis-specify covariates with complex non-linear effects and thereby fail to adjust adequately for their confounding; hence the need for use of non-linear terms such as fractional polynomials. Matching of covariates at baseline to balance their distribution across the exposed and non-exposed groups arrests confounding without any necessity to understand the shape of dose-responses relating the confounding covariates to the outcome. Thus a major source of residual confounding (known in the propensity score field as violation of the 'ignorability of treatment assignment' assumption) in routine regression models can be controlled.

It is important to re-emphasise that propensity score methods differ from a randomised controlled trial (RCT) in that the former can only balance the distribution of measured confounders between exposed and non-exposed, whereas the latter has the enviable property of balancing all confounders including those that are unmeasured or unknown.

Use in assessment of effect modification: Another use of propensity scores, and one you will experience in today's training exercise, is to facilitate the study of effect modification by propensity to receive treatment based on covariates influencing clinical decisions (as distinct from structural barriers to access). Patients with a high propensity score based on clinically relevant covariates will tend to have obvious indications for use of the intervention. Patients with a low propensity score will tend to have more marginal indications. Stratified analysis will therefore enable an assessment of whether the intervention is still effective when applied to patients in whom the justification for treatment is relatively less compelling.

Calculation of a Propensity Score

If the treatment or other intervention is specified as a binary outcome in a logistic regression model, then the propensity score as a probability of intervention between 0 and 1 is obtained from:

$$\Pr[\text{treatment} \mid \mathbf{x}] = [e^{\alpha + \beta 1x1 + \beta 2x2 + \ldots + \beta kxk}] \div [1 + e^{\alpha + \beta 1x1 + \beta 2x2 + \ldots + \beta kxk}]$$

Because this function monotonically increases as $[\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k]$ increases, the rank order of the latter sum-product of the regression coefficients and the values of their corresponding covariates in a subject is sufficient for ranking purposes and creation of quintiles.

Integrating Propensity Scores into Analysis

There are at least three methods:

1. ***Propensity regression***: The propensity score is included as a covariate (additional to treatment) in a multivariate regression model with or without additional important covariates.
2. ***Propensity stratification***: The study subjects are stratified into quintiles based on the propensity score and a stratified analysis is undertaken of the effect of treatment on the outcome.
3. ***Propensity matching***: Treatment and non-treatment subjects in each stratum are category matched or individually matched by propensity score. Matching has been suggested as the most statistically efficient and valid method of integrating propensity scores into an analysis.

Example: *Rosenbaum and Rubin used propensity scores in a three-year follow-up of 1,515 patients with coronary heart disease, 590 of whom underwent bypass surgery and the remainder were treated medically. The data were confounded by indications for surgery. Propensity scores for bypass surgery were derived from information on 74 covariates. The propensity scores were divided into quintiles to form five study strata. The authors used propensity stratification (see stratified patients below on the left) to obtain an adjusted three-year cumulative mortality ratio of 0.73 in favour of surgery. Had the authors performed propensity matching, they would have randomly sampled 26/277 of the medical patients in the lowest propensity stratum, 68 medical patients in the next stratum, and so on, ending with a sample of 69/234 of the surgical patients in the highest propensity stratum. This would have achieved a balanced study design as shown on the right.*

Prior to Matching

Propensity score quintiles	Medical	Surgical
Lowest	277	26
2nd lowest	235	68
Middle	205	98
2nd highest	139	164
HIghest	69	234

After Matching

Propensity score quintiles	Medical	Surgical
Lowest	26	26
2nd lowest	68	68
Middle	98	98
2nd highest	139	139
HIghest	69	69

Advanced Theory Topic L:
5.2 Instrumental Variable Analysis

Instrumental variables (IVs) offer another method designed to reduce the problems of confounding by indication and other biases in observation studies. Unlike standard multivariate regression methods and propensity scores, which cannot adjust for unknown or unmeasured risk factors, IVs are more akin to a RCT because under ideal circumstances they adjust for both known and unknown risk factors. The method is well known in the field of econometrics, but is still to receive more than cursory attention in health services research.

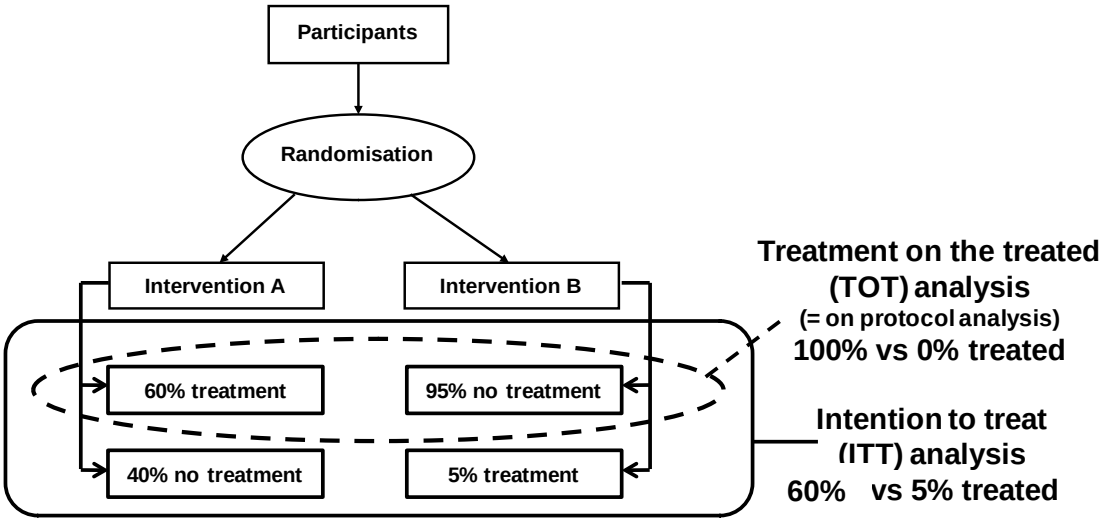
What is an Instrument Variable?

An IV (or set of variables) meets two criteria:

1. the variable is associated with the likelihood of receiving the intervention; and
2. the variable is not otherwise associated with the likelihood a good outcome through a causal effect itself or due to confounding with unmeasured risk factors.

Such a variable is called an ***instrument*** (for studying the outcome).

When valid, the instrument mimics some aspects of the way that randomisation in a RCT allocates patients to intention-to-treat groups. Random allocation does not necessarily guarantee treatment or non-treatment, but substantially influences the probability of treatment in a manner that bears no relationship to the outcome. If, for example, only 60% of patients randomly allocated to treatment actually adhere to it, whereas 5% randomly allocated to no treatment nevertheless obtain treatmen by other means, then an intention-to-treat analysis compares exposure to a 60% chance of receiving treatment to a 5% chance. While it would be preferable to measure the effect of the 'full strength' of treatment rather than merely a 60% chance of treatment, the intention-to-treat (ITT) analysis has the advantage (if the sample size is large) of high internal validity due to lack of bias. Compare this with a treatment on the treated (TOT or on-protocol) analysis of the same trial, comparing only the 60% who actually received treatment with the 95% of controls who did not. Here the contrast is stronger (100% treatment vs 0% treatment), but confounding factors are no longer balanced across the two groups with resultant loss of internal validity. Use of an IV resembles a non-experimental version of an ITT analysis in the presence of substantial variations from protocol in each arm of the trial.



Differences Between Use of Instrumental Variables and Intention-to-Treat Analysis

1. ***Assumption of balanced unmeasured risk factors across IV states***: In a RCT, the validity of this assumption arises from the statistical properties of randomisation, which render the assumption extremely robust as the sample size becomes large. Unfortunately, there is no guarantee nor any direct test of the validity of this assumption as applied to an IV. This assumption can be explored but not proven. One method of exploring the assumption is to study the effect of the IV on the distribution of measured risk factors. The nature of the IV may suggest that the assumption is likely to hold up also for unmeasured risk factors.
2. ***Effect of the intervention in marginal patients***: In general, it is uncommon in RCTs for many members, if any, of a control group to receive the treatment. Also, the reasons for non-adherence in the treatment group will generally relate to side effects of treatment or complications of illness that warrant other interventions. Many of the non-compliers may have received at least partial exposure to the treatment. The circumstances applicable to IVs are usually very different to this. Because the IV states differ usually only marginally in the likelihood of treatment, the method estimates the marginal effect of treatment over only that range of variation in treatment.

Examples of Instrumental Variable Analyses

The classic study using IVs in health services research is that reported by McClellan *et al*:

Example: McClellan et al used an IV to study the effects of cardiac catheterisation on case fatality from acute myocardial infarction in the elderly. They studied 218,427 US Medicare patients aged 65+ years who had an AMI in 1987, but not in 1986. The exposure was cardiac catheterisation within 90 days of AMI, often followed by a revascularisation procedure. The study outcome was case fatality within four years. Their IV was the difference in distance between a patient's nearest catheterisation hospital and their nearest hospital of any type (ie, how much further they had to travel to reach a catherisation hospital). The following results from the study showed that the marginal effect was -0.4/6.7 = a 6.0% reduction in four-year case fatality in marginal patients:

	Differential distance		
	<2.5 miles	>2.5 miles	Change
Catheterisation within 90 days	26.2%	19.5%	+6.7%
Mortality within 4 years	58.1%	58.5%	-0.4%

Other examples of IVs in the health services research literature have included use by Malkin *et al* of birth before and after noon as an IV to examine the effect of postpartum length of stay (LOS) on newborn readmission rates. Babies born before noon had a longer average LOS than those born after noon, whereas there was no reason to believe that birth before or after noon would itself affect the outcome or be associated with unmeasured risk factors. Another example was use by Howard of blood group as an IV to examine the effect of waiting time on liver transplant outcomes. Type O blood was harder to match and associated with a longer waiting time. There was no reason to suspect that blood group would affect transplant failure directly or through unmeasured risk factors.

The challenge with IV analysis remains the identification of suitable instruments and justification for the assumption of a balanced distribution of unmeasured risk factors across IV states. Some further possibilities for instruments include –

- Differential distances for high and low intervention facilities.
- Differential referrals to high and low intervention providers.
- Differential occurrence in high and low intervention time periods.

Rarely will the investigator be sufficiently confident of the performance of the IV to warrant reliance on the method as the mainstay of an analysis. Nevertheless, situations often arise, especially with respect to rapid changes in intervention levels over time with the introduction of new technology or pharmacotherapy, where the IV method is useful as part of an exploratory analysis.

Advanced Theory Topic M:
5.3 Some Aspects of the Assessment of Effect Modification

The optimisation of external validity or scientific generalisability is ultimately concerned with ***effect modification***, also known as ***heterogeneity of effect***. This occurs when the effect of a putative causal factor is modified according to the level of exposure to another factor. There are two ways in which effect modification impacts on external validity:

1. ***Restricted study domain***: If the study domain is restricted to only one level or a limited range of the effect modifier, scientific generalisation of the estimated effect to other levels of the modifier may be impaired, especially if there are considerations beyond the study to suggest that effect modification may be pronounced.
2. ***Unrestricted study domain***: When research is undertaken within a broad study domain having a typical distribution of the effect modifier, but without any specific assessment of it, the results may be generalised as the 'average' effect in any similar heterogenous population. But this belies the fact that the average effect is exactly as its name suggests – an average. It is a moot point as to whether this should cause concern. With the exception of extremes where the risk of outcome following exposure is either 0% or 100%, effect modification is at work in every research study result. The reason that, in the presence of identical measured exposures, some individuals develop an outcome and others do not is due to the influence of unknown effect modifiers. If perfect knowledge of all effect modifiers existed, it would be possible to identify in advance who would and would not sustain the outcome and the concept of 'risk' would become unnecessary.

It should be remembered that the presence or absence of effect modification in statistical terms, as distinct from aetiological or public health terms, is dependent on the choice of a relative or absolute measure of effect (see the diagrams below).

A well-established example of the effect modification in the literature is that of smoking on the association between alcohol intake and a range of cancers including those of the head and neck.

British Journal of Cancer (2000) 82(1), 204–207
© 2000 Cancer Research Campaign
Article no. bjoc.1999.0900

Short Communication

Risk factors for oral and pharyngeal cancer in women: a study from Italy and Switzerland

C Bosetti[1], E Negri[1], S Franceschi[2], E Conti[3], F Levi[4], F Tomei[5] and C La Vecchia[1,6]

Table 3 Combined effect of cigarette smoking and alcohol intake on the risk of oral and pharyngeal cancer in women. Italy and Switzerland, 1984–1997

Alcohol intake (drinks per day)	Odds ratio[a] for number of cigarettes per day [No. cases:No. controls]		
	Never smoker	1–14	≥15
0	1[b] [16:272]	2.25 [5:35]	3.14 [7:39]
< 2	1.92 [27:235]	2.54 [6:44]	2.04 [4:22]
≥ 2	1.83 [29:260]	11.55 [46:60]	15.41 [36:32]

[a]Estimates from conditional logistic regression, adjusted for education and body mass index. [b]Reference category.

constant relative effect across potential modifier categories. Multiple linear regression is scaled additively and thus the state of 'no interaction' equates with a constant absolute effect. For this reason and because the presence of product terms in regression models changes the interpretation of the regression coefficients for the main effect terms, product terms are of limited value in epidemiologic effect estimation. They are, however, the preferred approach when the object of regression modelling is risk prediction for individuals, given a specified vector of covariates, rather than effect measurement of risk factors.

Focus of Effect Estimation

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots$$

Focus of Risk Prediction

2. ***Mutually exclusive interaction categories***: Provided that the covariates involved in the putative effect modification can be collapsed into a small number of value categories, it is readily possible to assign mutually exclusive 0-1 binary indicators to each category above baseline. This enables the pattern of effect measures to be assessed moving along and between the axes of exposure to each covariate. It also allows for calculation of Rothman's synergy index (S) from relative effect measures:

$$S = (RR_{AB}-1)/(RR_A + RR_B - 2)$$

where RR_A is the relative effect measure comparing people exposed to A with those exposed neither to A nor B; RR_B is similarly defined; and RR_{AB} is the relative effect in those exposed both to A and B. A S value greater than one is empirically consistent with positive effect modification on an additive scale. This method is especially relevant when both exposures are amenable to adoption or intervention.

3. ***Stratum-specific analysis***: Separate estimates of effect are calculated in each stratum of the effect modifier. This method is especially relevant when one exposure is relatively fixed (eg, demographic characteristics) and the other is amenable to adoption or intervention.

	V1	V2	V3
-	0	0	0
A	1	0	0
B	0	1	0
AB	1	1	1

1. Product term in regression model

Ln[OR] = $\beta_A + \beta_B + \beta_{AB}$ = 0.405 + 0.693 - 0.105

Antagonism of a multiplicative scale

For A: Ln[OR] = 0.405; OR = 1.50
For B: Ln[OR] = 0.693; OR = 2.00
For AB: Ln[OR] = 0.405 + 0.693 - 0.105; OR = 2.70

	V1	V2	V3
-	0	0	0
A	1	0	0
B	0	1	0
AB	0	0	1

2. Mutually exclusive interaction categories

ORs	A-	A+
B-	1.00	1.50
B+	2.00	2.70

Synergism on An additive scale

Synergy index = (2.70 – 1)/(1.50 + 2.00 -2) = 1.13

3. Stratum-specific analysis

ORs	A-	A+
B-	1.00	1.50

ORs	A-	A+
B+	1.00	1.35

References

Bowden RJ, Turkinton DA. **Instrumental Variables.** Cambridge, Mass: Cambridge University Press, 1984.

Foster EM. **Propensity score matching. An illustrative analysis of dose response.** *Med Care* 2003; 41: 1183-92.

Greenland S. **An introduction to instrumental variables for epidemiologists.** *Int J Epidemiol* 2000; 29: 722-9.

Howard D. **The impact of waiting time on liver transplant outcomes.** *Hlth Serv Res* 2000; 35: 1117-34.

Iezzoni LI (ed). **Risk Adjustment for Measuring Health Care Outcomes.** Chicago, Illinois: Health Administration Press, 2003.

Imbens GW. **The role of the propensity score in estimating dose-response functions.** *Biometrika* 2000; 87: 706-10.

Kleinbaum DG, Kupper LL, Morgenstern H. **Epidemiologic Research: Principles and Quantitative Methods.** Belmont, California: Wadsworth, 1982

McClellan M, McNeil BJ, Newhouse JP. **Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality?** *JAMA* 1994; 272: 859-866.

McClellan MB, Newhouse JP. **Overview of the Special Supplement Issue.** *Hlth Serv Res* 2000; 35: 1061-9.

Malkin JD, Broder MS, Keeler E. **Do longer postpartum stays reduce newborn readmissions? Analysis using instrumental variables.** *Hlth Serv Res* 2000; 35: 1071-91.

Miettinen OS. **Stratification by a multivariate confounder score.** *Am J Epidemiol* 1976; 104: 609-20.

Miettinen OS. **Theoretical Epidemiology. Principles of Occurrence Research in Medicine.** New York: John Wiley & Sons, 1985.

Newgard CD, Hedges JR, Arthur M, *et al.* **The propensity score – a method for estimating treatment effect in observational research.** *Acad Emerg Med* 2004; 11: 953-61.

Newhouse JP, McClellan M. **Econometrics in outcome research: the use of instrumental variables.** *Ann Rev Public Health* 1998; 19: 17-34.

Rosenbaum PR, Rubin DR. **The central role of the propensity score in observational studies for causal effects.** *Biometrika* 1983; 70: 41-55.

Rosenbaum PR, Rubin DR. **Reducing bias in observational studies using subclassification on the propensity score.** *J Am Stat Assoc* 1984; 79: 516-24.

Rothman KJ. **The estimation of synergy or antagonism.** *Am J Epidemiol* 1976; 103: 506–11.

Rothman KJ, Greenland S, Lash TL. **Modern Epidemiology.** 2nd Edition. Raven, Philadelphia: Lippincott, 2008.

Bosetti C, Negri E, Franceschi S,Conti E,Levi F,TomeiF,La Vecchia C. **Risk factors for oral and pharyngeal cancer in women: a study from Italy and Switzerland.** Br J Cancer. 2000 January; 82(1): 204–207.

TRAINING SESSION 5:
OPTIMISING EFFECT ESTIMATION

IMPORTANT NOTICE

The practice data sets provided for training purposes in this course are fictitious. All results based on analysis of these fictitious data are scientifically useless. The course organisers and instructor take no responsibility for any misuse of the spurious results arising from the analysis of these fictitious data sets.

Background to Training Session 5

The aim of this training session is to give you hands-on experience in the use of linked administrative data to implement an exploratory analysis that will control the presence of confounding as well as uncover and assess the presence of effect modification. We will make use of a toolbox of analytic methods and perspectives, which draws on our understandings of advanced regression modelling, instrumental variables and propensity scores. The exercise centres on a classic issue in medication effectiveness, that of the chemopreventive value of use of statins (blood lipid-lowering drugs) on mortality in people with diabetes. You are reminded that the exercise is fictitious and that no real scientific conclusions should ever be drawn from your results based on our 'pretend' practice data sets. You should note also that from a peer-review perspective, your analysis will be flawed in several respects, especially because the administrative data lack baseline blood lipid measurements. Nevertheless, the exercise is important in illustrating the creative use of an armamentarium of analytic methods that will coax the linked data to yield up their secrets and, thereby, to improve internal and external validity in effect estimation.

Training Session Resources

You will be making use of your updated platform file, your tagged and sequenced PBSdata file and a file of supplementary covariate data, diabsupp.

If you are using this statistical package …	*… you should use this save file.*
SPSS	Diabetes8.sav PBSdata3.sav diabsupp.sav
SAS	Diabetes8.sas7bdat PBSdata3.sas7bdat diabsupp.sas7bdat
Stata	Diabetes8.dta PBSdata3.dta diabsupp.dta
R	Diabetes8.rda PBSdata3.rda diabsupp.rda

Note that the following additional PBS item codes are relevant to this exercise:

Data Set	Coding System	Codes
PBS	*Pharmaceutical Benefits Schedule*	**Statins**: 1224, 1453, 1687, 1942, 2011, 2012, 2013, 2831, 2833, 2834, 2892, 2967, 2978, 8023, 8024, 8173, 8197, 8213, 8214, 8215, 8303, 8304, 8313, 8419, 8521, 8721, 8722, 8757

WARNING: You should not rely on these codes in any analysis separate from the course as they may contain errors or be incomplete.

Preparing a Linked File for Running Regression Models

Much of this exercise is concerned with analytic strategy as applied to the use of regression models. Before modelling commences it is often necessary to 'prepare' the working file so that all of the necessary covariates are present in a suitable form. Often this involves attention to:

1. ***Missing covariate values***: These can either be (i) left blank, thereby excluding the record from models that make use of the covariate; (ii) assigned a binary ***missing value indicator*** in combination with assignment of a minimal value to the covariate; or (iii) given ***imputed values*** based on an algorithm, the simplest of which is imputation of the mean value.
2. ***Binary indicators for categorical covariates***: It is usually preferable to use 0-1 binary indicators and to make a conscious choice of a reference or non-exposed category for the covariate. For multi-category covariates, a series of *n*-1 binary indicators is needed, where *n* is the number of exposure categories. Some statistical packages offer options to form binary indicators automatically for categorical covariates.
3. ***Transformations and fractional polynomials for continuous covariates***: Continuous covariates thought to have non-linear effects are best transformed using fractional polynomials, or sometimes a simple function such as the natural log will suffice.
4. ***Interaction terms and indicators***: Product interaction terms or mutually exclusive interaction category indicators may be required if you plan to assess effect modification.

Fractional Polynominals

These were covered in the introductory course. They are similar to conventional polynomial functions, but offer the potential for a better fit by including non-integer and negative powers. The 'degree' of the fractional polynomial model determines the number and form of the possible terms. For example a second degree takes the form $\beta_1 x^{p1} + \beta_2 x^{p2}$ where p = {[-2,] -1, -½, 0, ½, 1, 2} subject to some complex rules. Some statistical packages (eg, Stata) advise the user on which fractional polynomial models offer the best fit.

Address Matching and Use of Geocodes

In address matching, a data file containing street addresses is matched to a geographic reference file to obtain latitude-longitude coordinates, known as ***geocodes***. Geographic reference files containing this information are readily available and typically geocodes are returned in the form of ***eastings*** and ***northings*** on the Australian Map Grid Zone system. For example, a geocode in the WA Data Linkage Project may look like this:

393838.066 6460404.141 50 HM03

where the first number is the easting, the second number is the northing, followed by the AMG Zone number and finally a four-character DLU code that defines the type of address and the level of resolution that was achieved through address matching.

Exact geocodes are not routinely released by the DLU in WA, because they are considered to be identifying information. However using ***mapping software*** such as MapInfo, it is possible to convert geocodes into ABS ***collectors' districts*** (CDs), which are clusters of about 200 households and the smallest areas used by ABS for statistical aggregation purposes. In WA these CDs have been assigned to hospital morbidity, mortality, midwife notification and electoral roll records from the mid-1980s onwards.

CDs are especially useful as a means to assign SEIFA indices of social disadvantage (Socio-Economic Indices For Areas) made available form the Australian Bureau of Statistics, and ARIA indices of remoteness (Accessibility/Remoteness Index of Australia) made available from the Australian Department of Health and Ageing.

Socio-Economic Indices For Areas (SEIFA)

SEIFA consists of four indeces developed by the ABS. Each index summarises a different aspect of the socio-economic conditions of the Australian population using a combination of variables from the Census. The most frequently used of the four indices in health services research is the ***Index of Relative Socio-Economic Disadvantage***, sometimes called just the ***IRD***, which is also the index preferred by the authors of the *Social Health Atlas of Australia*. The variables used to compose the IRD include low income, low educational attainment, high unemployment and people with low skilled occupations. IRD values enable areas to be ranked and compared against one another. A low index value reflects relative disadvantage and occurs where there are large proportions of low-income families, people with low skilled occupations and high proportions of the population without training. A high value reflects lack of disadvantage in an area.

A popular approach is to use the IRD for risk adjustment is through ***population-weighted quintiles***. The IRD scores for each CD or postcode in a State are ranked in descending order of the scores and are then divided into five groups that are of equal size with respect to population (not the number of postcodes or CDs). Group 1 then becomes the most advantaged and group 5 the most disadvantaged. An alternative approach is the use of population-weighted quartiles, but to further divide the most disadvantaged quartile into the extreme 10% of the total population (highly disadvantaged) and the remainder of that quartile. This approach is useful when one suspects that high risk is especially concentrated in the upper tail of the disadvantage distribution.

In this exercise, to save you time, you are given the population-weighted IRD quintiles based on postcode for each patient with diabetes in the diabsupp file. Note the steps that were followed to derive these for you: (i) postcode was ascertained at the time of the inaugural index date; and (ii) these postcodes were mapped to IRD population-weight quintiles using a mapping file as a look-up table. While it is always preferable to use CD-based rather than postcode-based IRDs, only postcode is available to researchers on Commonwealth data sets.

Accessibility/Remoteness Index of Australia (ARIA)

ARIA measures remoteness in terms of road access from populated localities to different categories of ***service centre***. Localities that are more remote have less access to service centres; those that are less remote have greater access. Service centres are categorised on the basis of their population size. Unlike SEIFA, ARIA scores are not generally grouped by researchers into equal sized percentile intervals. This is because a large proportion of the population of a typical Australian State resides in areas of very low remoteness; ie, the capital and other major cities. Rather, the scores are grouped as follows on the basis of the size of the score as a measure of remoteness:

- Highly Accessible (ARIA score 0 - 1.84) - relatively unrestricted accessibility to a wide range of goods and services and opportunities for social interaction.
- Accessible (ARIA score >1.84 - 3.51) - some restrictions to accessibility of some goods, services and opportunities for social interaction.
- Moderately Accessible (ARIA score >3.51 - 5.80) - significantly restricted accessibility of goods, services and opportunities for social interaction.
- Remote (ARIA score >5.80 - 9.08) - very restricted accessibility of goods, services and opportunities for social interaction.
- Very Remote (ARIA score >9.08 - 12) - locationally disadvantaged - very little accessibility of goods, services and opportunities for social interaction.

As with SEIFA, to save you time, you are given the ARIA group based on postcode for each patient with diabetes in the diabsupp file. Note the steps that were followed to derive these for you: (i) postcode was ascertained at the time of the inaugural index date; and (ii) these postcodes were mapped to ARIA groups using a mapping file as a look-up table. It is always preferable to use CD-based rather than postcode-based ARIAs, but this is not possible for researchers using Commonwealth data sets.

ARIA+ (ASGC Remoteness Areas)

While this course uses ARIA, you should note that it is no longer the preferred remoteness measures. Instead, the ABS has endorsed the ARIA+ accessibility/remoteness classification, which has been incorporated within the Australian Standard Geographical Classification (ASGC). ARIA+ includes an additional 'population size' category in the assessment of accessibility, its scores ranging between 0 and 15 rather than 0 and 12. More information is available at http://www.gisca.adelaide.edu.au/.

Multipurpose Australian Comorbidity Scoring System *(MACSS)*

You were introduced to this comorbidity scoring system in the introductory course. The MACSS score is based on the presence of the most frequent 102 comorbid conditions associated with a RR≥1.1 of death or readmission or a length of stay difference ≥0.5 days. Extensive testing has demonstrated that MACSS is a superior comorbidity adjustment tool than the Charlson Index.

Also to save you time, in today's exercise you are given a MACSS score for each diabetic patient in the diabsupp file. This score is based entirely on hospital morbidity data. MACSS conditions were identified from all hospital records with an admission date ranging from 365 days prior to the inaugural index date to 30 days after the inaugural index date. The later 30-day extension was designed to detect comorbidity on hospital records at around the time of an inaugural index MBS or PBS record. Research by Preen and others using MACSS examined the impact of different comorbidity lookback periods on modelling post-hospitalisation mortality and found that a one-year ***lookback period*** provides the best model-fit, and is therefore the best method of risk adjustment for comorbidity, when evaluating a mortality outcome.

Truncation of Time Series and Follow-up to Prevent Artefacts

Longitudinal research that examines outcomes in patients who commence their follow-up over a wide range of calendar times is subject to the introduction of a number of artefacts, which act by creating discrepancies in the comparability of information over time. Two of the important artefacts to avoid are as follows:

1. ***Comorbidity look-back***: To ensure comparability in information across all study subjects it is essential that they have data available covering the same duration of comorbidity lookback period. Thus, if linked data commence in 1990 and a one-year lookback is implemented, only index cases arising in 1991 onwards should be included in analysis, as the earlier patients have only incomplete ascertainment of comorbidity. Similarly, later patients should not have comorbidity ascertained over a period longer than one year.
2. ***Maximum length of follow-up***: Patients arising early in a series may have lengths of follow-up many years greater than the most recent patients. If the measures under study, eg incidence rates of adverse events in the exposed to non-exposed, are relatively constant during follow-up, varying lengths of time will make little difference to the result and can be ignored. However, more often than not, measures vary over time of follow-up, and this may become a source of error if the prevalence of exposure also varies over time. The most extreme example is where calendar year at time-zero is itself a covariate. This is quite common due to the desire to evaluate if outcomes have improved in patients treated more recently. However, the recently treated patients will have a shorter average length of follow-up than the earlier patients. If the risk of the outcome (the ***hazard period*** as covered in topic session 4) is concentrated in the first few year of follow-up, treatment of the recent patients will appear to perform poorly relative to earlier patients unless follow-up times are equalised. This is done by censoring the duration of follow-up after a fixed interval in all patients (eg, 'three-year mortality' or 'one-year readmission') and excluding from analysis the recent patients who are unable to at least substantially complete the designated follow-up time.

Overview of Exercise 5

Steps 1-2. Open PBSdata, tag and sequence statin scripts, cutdown to first-time statin scripts and save as PBSstatin.

Step 3. Open diabetes platform and load statin date from PBSstatin. Exclude patients first prescribed statin before aug date. Classify others as statin in first year or later/never.

Steps 4-6. Truncate time series by excluding 1990. Load SEIFA, ARIA and MACSS data from diabsupp file. Create stage at aug date variable.

Step 6. Create survival variables for follow-up truncated at 3 years. Select records < 1998.

Step 7. Bivariate Cox regression analysis relating 3-year mortality rate to early statin treatment.

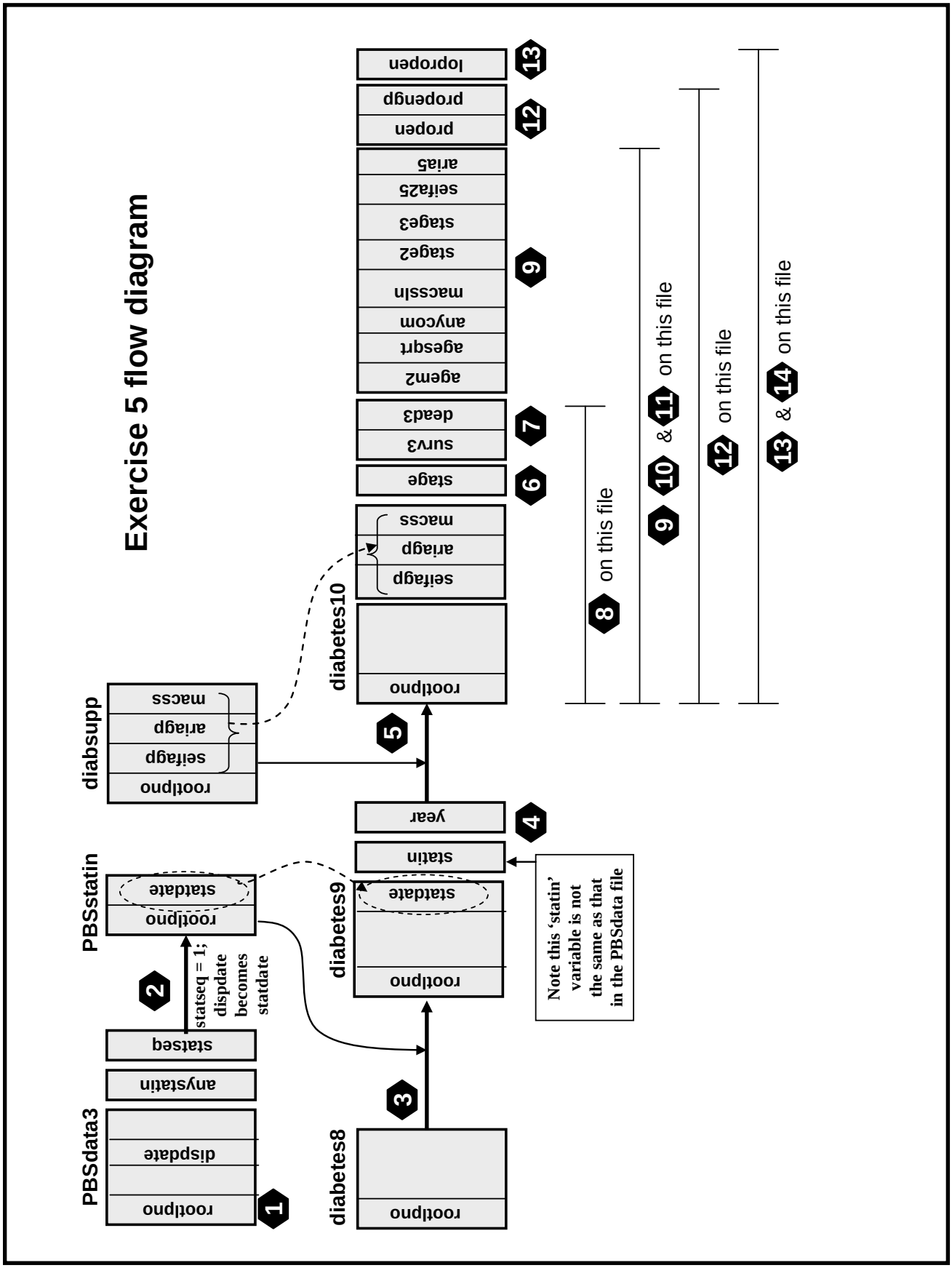
Step 8. Tidy and transform age, sex, MACSS, stage SEIFA and ARIA covariates and run multivariate Cox regression analysis.

Steps 10-11. Exploratory analysis with time as instrumental variable.

Step 12. Derive propensity scores relating covariates to early statin treatment.

Step 13. Assess effect modification by propensity score using an interaction term.

Step 14. Assess effect modification by propensity score using stratum-specific analysis.


Exercise 5 flow diagram
PBSdata3
rootlpno
dispdate
anystatin
statseq
statseq = 1;
dispdate
becomes
statdate
PBSstatin
rootlpno
statdate
diabsupp
rootlpno
seifagp
ariagp
macss
diabetes8
rootlpno
diabetes9
rootlpno
statdate
statin
year
diabetes10
rootlpno
seifagp
ariagp
macss
stage
surv3
dead3
agem2
agesqrt
anycom
macssin
stage2
stage3
seifa25
aria5
propen
propengp
lopropen
Note this 'statin' variable is not the same as that in the PBSdata file
on this file
on this file
on this file
on this file

Exercise Instructions

Ascertainment of Statin Exposure

1. Open your updated PBSdata3 file and tag all records with a statin PBC item number using a new variable, anystatin. Set anystatin to '1' for a tagged record and '0' otherwise.

2. Create a dispensing sequence variable, statseq, which numbers the statin scripts. Select the first statin script for each patient (statseq=1) and delete the remainder of records. Rename dispdate as statdate and delete all variables in the file except rootlpno and statdate. Now select the 3,624 first-time statin scripts and save them as PBSstatin after deleting the remaining records.

3. Open your updated diabetes8 file and load statdate onto the end of each applicable record in the platform using PBSstatin as an external look-up table keyed by rootlpno. The next step is designed to classify each diabetic as a patient who either filled a statin script prior to their inaugural index date; received one in the first year after that date; or either never received a statin script or received one later than the first year. Therefore, create a new variable, statin, and set a default value of '0'. Apply a series of tests to the number of days between statdate and date, such that when statdate precedes date, statin is assigned the value '9'; and when statdate falls on or within the next 364 days after date, statin is assigned the value '1'. Delete all the records from your file where statin = 9, leaving 9,713 records of diabetic patients not known to have already been prescribed a statin at the time of their inaugural index date. The reason for this exclusion is to restrict the study domain to diabetic patients prescribed a statin for the first time.

Preparation of Additional Covariates

4. Create a year variable by drawing the year off the inaugural index date variable. Now select the 8,263 records where year ≥ 1991 and delete the remaining records from 1990. The reasons for this exclusion are (i) to ensure that all patients in the analysis have at least one full year of hospital morbidity information prior to their diabetes diagnosis on which to base the assessment of comorbidity; and (ii) to strengthen the assumption that all patients prescribed statins prior to their diabetes diagnosis have been excluded. Save as diabetes9.

5. Load the variables, seifagp, ariagp and macss onto the end of your platform file, using diabsupp from your CD as an external file keyed by rootlpno. **Stata users:** You will need to open the diabsupp file and perform a stable sort by rootlpno before merging. Save as diabetes10.

6. Create a new stage variable that is assigned the value '1', '2' or '3' according to whether at the inaugural index date the patient was in clinical stage 1 (no diabetic pharmacotherapy), 2 (oral hypoglycaemic agent, but not insulin) or 3 (insulin). **Hint:** You can use the stgpt1, stgpt2 and stgpt3 variables that you created in exercise 2; eg, if stgpt1 = 0 and stgpt2 =0 and stgpt3 >0, then the patient must have commenced in clinical stage 3. You should have 5,489 stage1 patients, 2,291 stage 2 patients and 483 stage 3 patients.

Survival Variables

7. The analysis will be restricted to a maximum of three years of follow-up to prevent the calendar year in which follow-up commences being correlated with length of follow-up. Create surv3, to become survival time censored at three years, initially by assigning to it the difference in days between exit and date. **Stata users:** You will need to replace the 26 cases of survival time equal to zero with a small positive number (eg, 0.5) to ensure they are included in the analysis

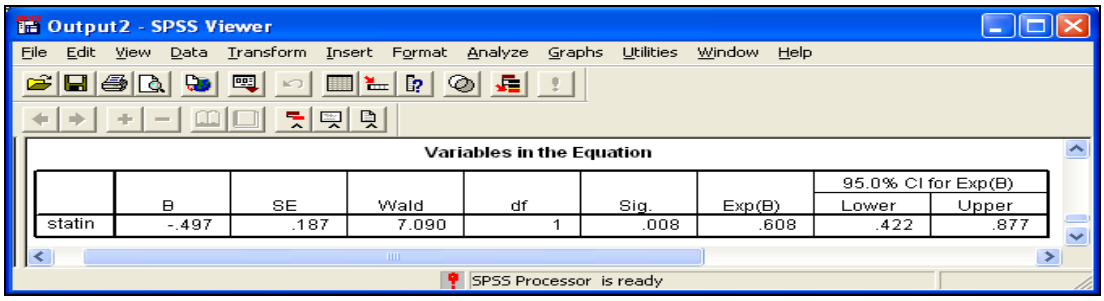
   Also create dead3, to become vital status at three years, initially by giving it the same value as dead. Two adjustments are then made such that if surv3 > 1,095, its value is trimmed back to 1,095 and dead3 is set to '0'. Select the 6,795 records with year < 1998 to ensure a minimum of two years of follow-up is available and delete the remaining records. Limiting the available follow-up to 2.5 years, on average, for new patients in 1997, is considered to be an acceptable

compromise. I suggest that you delete all variables except rootlpno, sex, age, statin, year, seifagp, ariagp, macss, stage, surv3 and dead3 and save the file with the name diabstat. Your file in now ready for analysis.

Bivariate Regression Model

8. Using Cox regression, run a simple bivariate model that estimates the mortality rate ratio over three years in diabetics newly prescribed a statin in the first year after diagnosis compared with other diabetics. Your results should resemble the following with MRR = 0.61, suggesting a crude protective effect of statins on premature mortality.

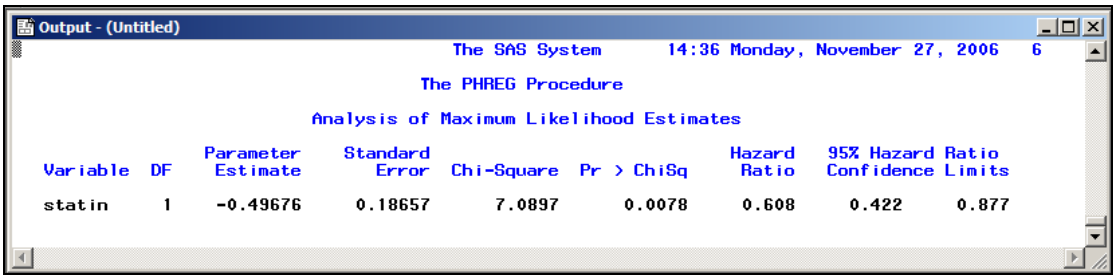
SPSS



Variables in the Equation

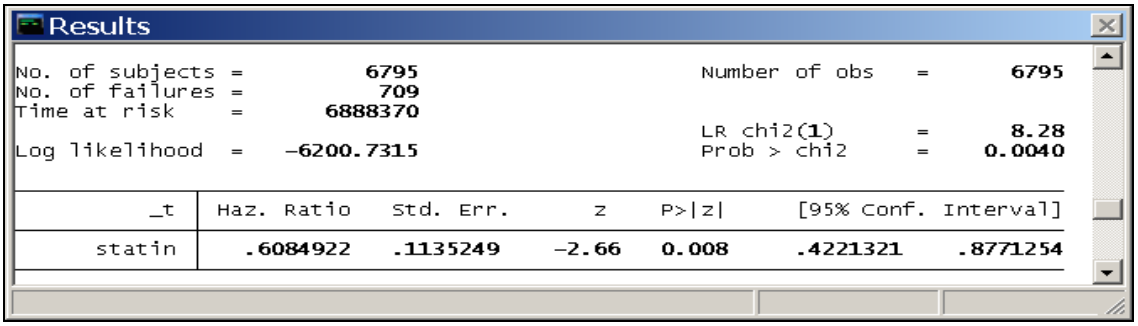
	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B) Lower	Upper
statin	-.497	.187	7.090	1	.008	.608	.422	.877

SAS




```
                                  The SAS System      14:36 Monday, November 27, 2006   6

                                 The PHREG Procedure

                           Analysis of Maximum Likelihood Estimates

                    Parameter    Standard                          Hazard   95% Hazard Ratio
Variable   DF        Estimate       Error   Chi-Square   Pr > ChiSq   Ratio   Confidence Limits

statin      1        -0.49676     0.18657       7.0897       0.0078   0.608     0.422     0.877
```

Stata




```
No. of subjects =        6795                    Number of obs    =      6795
No. of failures =         709
Time at risk    =     6888370
                                                 LR chi2(1)       =      8.28
Log likelihood  =   -6200.7315                   Prob > chi2      =    0.0040
```

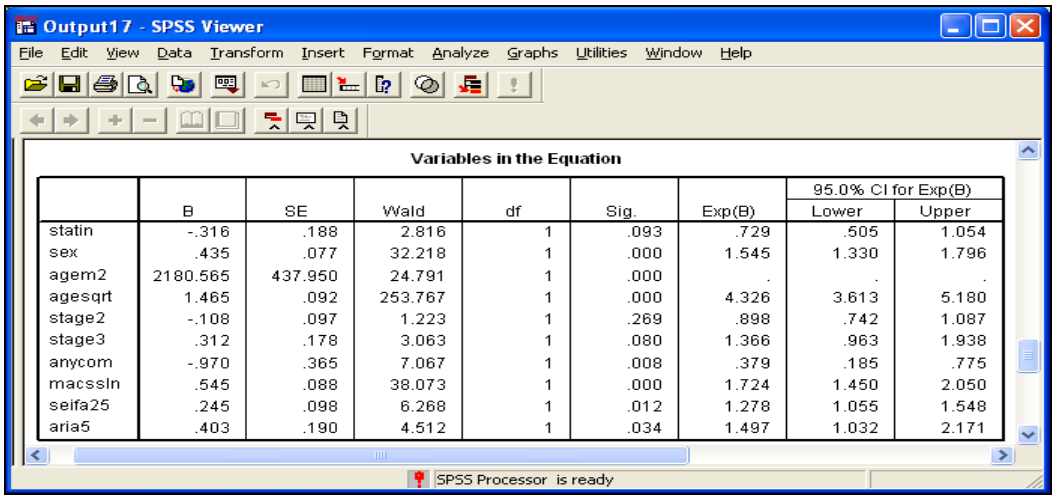
_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
statin	.6084922	.1135249	-2.66	0.008	.4221321	.8771254

Optimising Covariates for Multivariate Regression Model

9. Before running a multivariate model to adjust the crude effect, some transformations are necessary to optimise the fit of the covariates. This starts with imputing means for missing values for sex (90 instances) and age (9 instances). Recode these to 0.5 for sex (after also recoding '2' to '0' for females) and 59.2 for age. The following information will save you time: (i) optimal adjustment for age is thought to be achieved with a (-2, ½) fractional polynomial so replace a plain age covariate with two covariates, one of form $1/X^2$ (create as variable agem2) and the other $\sqrt{X}$ (create as variable agesqrt); (ii) optimal adjustment for comorbidity is achieved by a binary 0-1 indicator for any comorbidity, anycom, and by the natural log of the MACSS score <u>after</u> adding the constant 0.01 (create as variable macssln – but why did we add the 0.01?); and (iii) create a 0-1 binary indicator, stage2, for patient diagnosed in stage 2 diabetes and another, stage3, for those diagnosed in stage 3. In addition, (iii) equate a new variable seifa25 to

seifagp and recode seifa25 so that the value '1' becomes '0' and values '2-5' become '1'; and (iv) equate a new variable aria5 to ariagp and recode aria5 so that values '1-4' become '0' and value '5' becomes '1'. These two transformations will simplify your output given that most of the effect of socioeconomic status is expressed as protection in the least disadvantaged group and most of the effect of remoteness is expressed as a higher mortality in the most remote group. Save your files with these updates. Run a multivariate model of the effect of statin on three-year mortality with sex, agem2, agesqrt, stage2, stage3, anycom, macssln, seifa25 and aria5 as covariates. Your results should be like the following with the adjusted MRR = 0.73.

SPSS



Output17 - SPSS Viewer

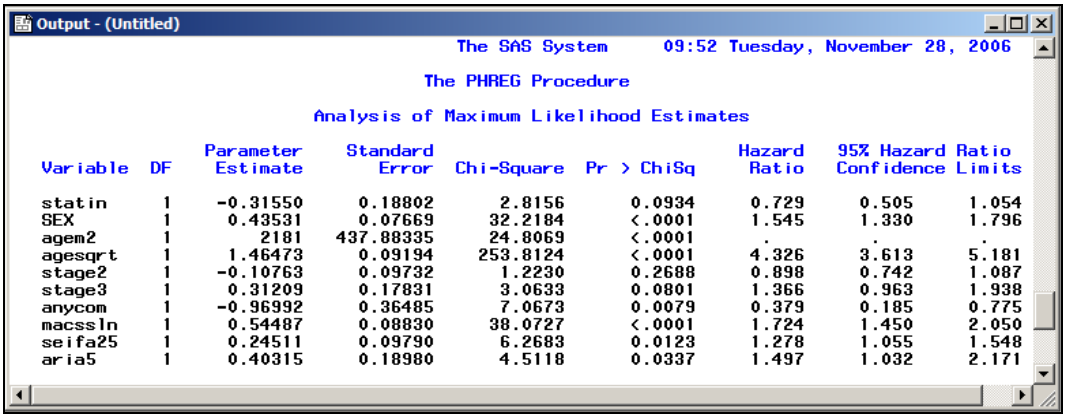
File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B) Lower	Upper
statin	-.316	.188	2.816	1	.093	.729	.505	1.054
sex	.435	.077	32.218	1	.000	1.545	1.330	1.796
agem2	2180.565	437.950	24.791	1	.000	.	.	.
agesqrt	1.465	.092	253.767	1	.000	4.326	3.613	5.180
stage2	-.108	.097	1.223	1	.269	.898	.742	1.087
stage3	.312	.178	3.063	1	.080	1.366	.963	1.938
anycom	-.970	.365	7.067	1	.008	.379	.185	.775
macssln	.545	.088	38.073	1	.000	1.724	1.450	2.050
seifa25	.245	.098	6.268	1	.012	1.278	1.055	1.548
aria5	.403	.190	4.512	1	.034	1.497	1.032	2.171

SPSS Processor is ready

SAS



Output - (Untitled)

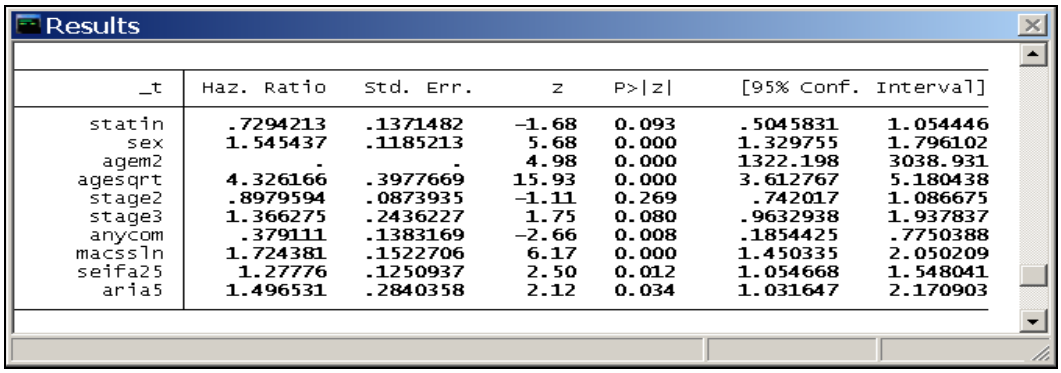
The SAS System 09:52 Tuesday, November 28, 2006

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
statin	1	-0.31550	0.18802	2.8156	0.0934	0.729	0.505	1.054
SEX	1	0.43531	0.07669	32.2184	<.0001	1.545	1.330	1.796
agem2	1	2181	437.88335	24.8069	<.0001	.	.	.
agesqrt	1	1.46473	0.09194	253.8124	<.0001	4.326	3.613	5.181
stage2	1	-0.10763	0.09732	1.2230	0.2688	0.898	0.742	1.087
stage3	1	0.31209	0.17831	3.0633	0.0801	1.366	0.963	1.938
anycom	1	-0.96992	0.36485	7.0673	0.0079	0.379	0.185	0.775
macssln	1	0.54487	0.08830	38.0727	<.0001	1.724	1.450	2.050
seifa25	1	0.24511	0.09790	6.2683	0.0123	1.278	1.055	1.548
aria5	1	0.40315	0.18980	4.5118	0.0337	1.497	1.032	2.171

Stata



Results

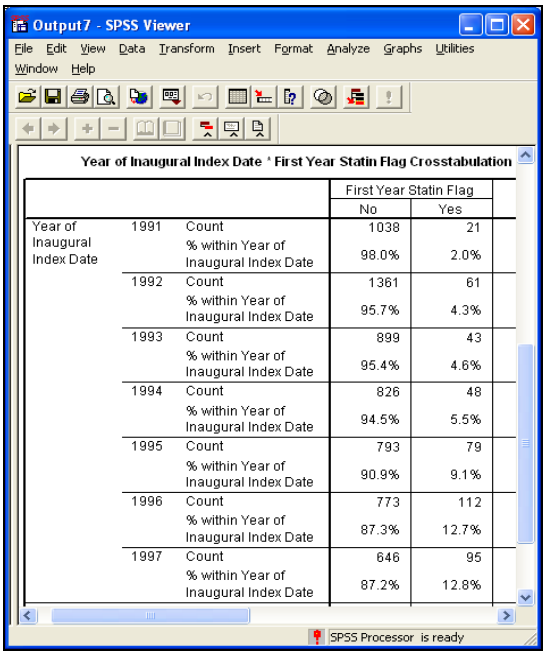
_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
statin	.7294213	.1371482	-1.68	0.093	.5045831	1.054446
sex	1.545437	.1185213	5.68	0.000	1.329755	1.796102
agem2	.	.	4.98	0.000	1322.198	3038.931
agesqrt	4.326166	.3977669	15.93	0.000	3.612767	5.180438
stage2	.8979594	.0873935	-1.11	0.269	.742017	1.086675
stage3	1.366275	.2436227	1.75	0.080	.9632938	1.937837
anycom	.379111	.1383169	-2.66	0.008	.1854425	.7750388
macssln	1.724381	.1522706	6.17	0.000	1.450335	2.050209
seifa25	1.27776	.1250937	2.50	0.012	1.054668	1.548041
aria5	1.496531	.2840358	2.12	0.034	1.031647	2.170903

On reviewing these results, do you have any reason to see any problem with the choice of a (-2, ½) fractional polynomial to model the effects of age may not have been idea? Why or why not?

Using Time as an Instrumental Variable

10. Although our results so far are consistent empirically (we ignore statistical precision for the purpose of this exercise) with a protective effect of early prescription of a statin on three-year mortality in diabetics, further exploratory analysis is warranted. The literature suggests that there has been a large increase in use of statins in Australia since the early 1990s and this can be confirmed in our data by crosstabulating statin by year (see opposite). The proportion of diabetics prescribed a statin in the first year increased from 2.0% in 1991 to 12.8% in 1997. Note how the marginal take-up was greater in the second half of the period than in the first half.

SPSS



Output7 - SPSS Viewer

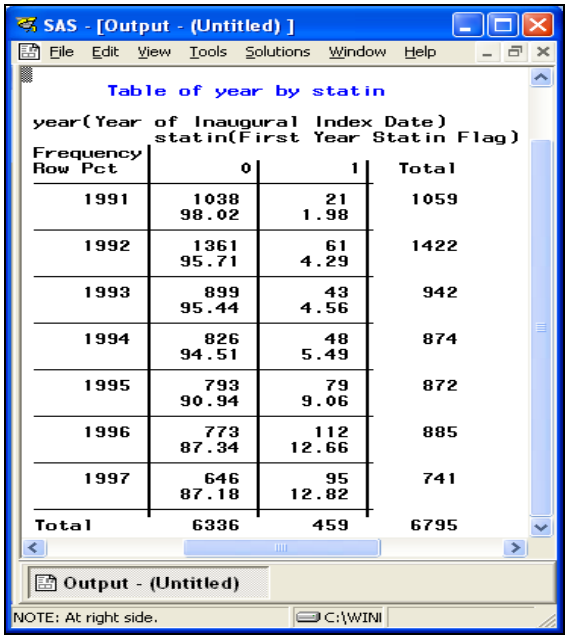
File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Year of Inaugural Index Date * First Year Statin Flag Crosstabulation

			First Year Statin Flag	
			No	Yes
Year of Inaugural Index Date	1991	Count	1038	21
		% within Year of Inaugural Index Date	98.0%	2.0%
	1992	Count	1361	61
		% within Year of Inaugural Index Date	95.7%	4.3%
	1993	Count	899	43
		% within Year of Inaugural Index Date	95.4%	4.6%
	1994	Count	826	48
		% within Year of Inaugural Index Date	94.5%	5.5%
	1995	Count	793	79
		% within Year of Inaugural Index Date	90.9%	9.1%
	1996	Count	773	112
		% within Year of Inaugural Index Date	87.3%	12.7%
	1997	Count	646	95
		% within Year of Inaugural Index Date	87.2%	12.8%

SPSS Processor is ready

SAS



SAS - [Output - (Untitled)]

File Edit View Tools Solutions Window Help

Table of year by statin

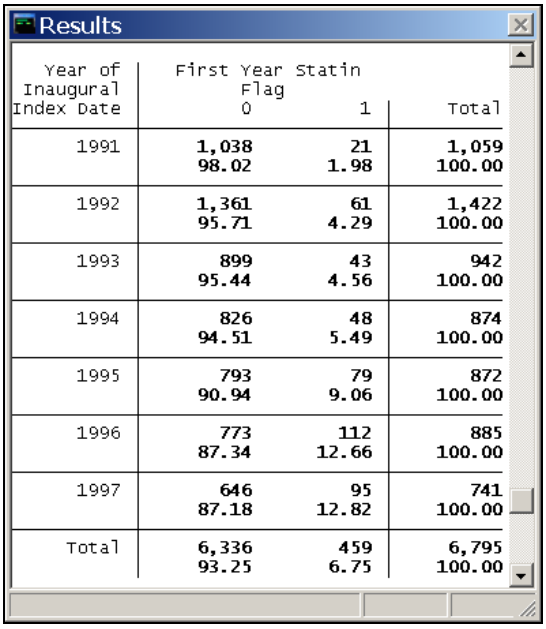
year(Year of Inaugural Index Date)
statin(First Year Statin Flag)

Frequency Row Pct	0	1	Total
1991	1038 98.02	21 1.98	1059
1992	1361 95.71	61 4.29	1422
1993	899 95.44	43 4.56	942
1994	826 94.51	48 5.49	874
1995	793 90.94	79 9.06	872
1996	773 87.34	112 12.66	885
1997	646 87.18	95 12.82	741
Total	6336	459	6795

Output - (Untitled)

NOTE: At right side. C:\WINI

Stata



Results

Year of Inaugural Index Date	First Year Statin Flag 0	1	Total
1991	1,038 98.02	21 1.98	1,059 100.00
1992	1,361 95.71	61 4.29	1,422 100.00
1993	899 95.44	43 4.56	942 100.00
1994	826 94.51	48 5.49	874 100.00
1995	793 90.94	79 9.06	872 100.00
1996	773 87.34	112 12.66	885 100.00
1997	646 87.18	95 12.82	741 100.00
Total	6,336 93.25	459 6.75	6,795 100.00

11. Run a bivariate Cox regression model of the effect of year on three-year mortality. Use the facility available in your statistical package to enter year as a categorical variable represented by a series of binary indicators, with the first year of 1991 as the reference category (**Stata users:** the xi command). **R users:** you may wish to create the binary indicator variables in your data file before fitting the model. Why is there no imperative to include all of the covariates in this regression model? Check that your results are similar to those below. Where is most of the marginal improvement in mortality concentrated: in the first half of the period or in the second half of the period? Is this observation consistent or inconsistent with the greatest marginal increase in uptake of statins being in the first half of the period? What are the limitations of using calendar year as an instrumental variable in these circumstances?

SPSS

Output9 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B) Lower	Upper
year			31.233	6	.000			
year(1)	.008	.112	.006	1	.940	1.008	.810	1.256
year(2)	-.177	.130	1.859	1	.173	.838	.650	1.080
year(3)	-.356	.140	6.503	1	.011	.700	.532	.921
year(4)	-.481	.145	10.967	1	.001	.618	.465	.822
year(5)	-.413	.141	8.537	1	.003	.661	.501	.873
year(6)	-.566	.169	11.180	1	.001	.568	.408	.791

SPSS Processor is ready

SAS

Output - (Untitled)

The SAS System 09:30 Tuesday, November 28, 2006

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
yr1992	1	0.00838	0.11187	0.0056	0.9403	1.008	0.810	1.256
yr1993	1	-0.17654	0.12965	1.8541	0.1733	0.838	0.650	1.081
yr1994	1	-0.35640	0.13979	6.5003	0.0108	0.700	0.532	0.921
yr1995	1	-0.48081	0.14521	10.9645	0.0009	0.618	0.465	0.822
yr1996	1	-0.41335	0.14148	8.5358	0.0035	0.661	0.501	0.873
yr1997	1	-0.56543	0.16922	11.1653	0.0008	0.568	0.408	0.792

Stata

Results

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Iyear_1992	1.008402	.1128032	0.07	0.940	.8098698	1.255602
_Iyear_1993	.8379763	.1086457	-1.36	0.173	.6499371	1.080419
_Iyear_1994	.7001353	.0978697	-2.55	0.011	.5323476	.9208072
_Iyear_1995	.6182563	.089772	-3.31	0.001	.4651281	.8217971
_Iyear_1996	.6614279	.0935756	-2.92	0.003	.501255	.8727832
_Iyear_1997	.5678579	.096104	-3.34	0.001	.4075517	.7912188

Derivation of Propensity Ranks

12. Given our suspicion from the instrumental variable analysis that the effect of early statin prescription may be modified as its use extends to patients with more marginal indications, we will now construct a propensity rank model to stratify patients into those with strong and more marginal indications for use of statins. To do this, run a logistic regression model with statin as the dependent (binary outcome) variable and sex, age, stage2, stage3, macsss and anycom as the independent variables. We exclude seifa2-5 and aria5 because the concern here is with clinical decision-making and not with structural barriers to access to treatment. Note that we do not use the transformations of age and comorbidity from before, as they were tailored to a Cox regression on mortality. See if you get the same logistic regression coefficients as shown below. **Stata users:** You will need to specify that you want the coefficients (rather than the odds ratios) by adding “, coef” at the end of the command line.

SPSS

Output24 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
sex	-.014	.098	.021	1	.884	.986
age	.002	.003	.382	1	.537	1.002
stage2	.746	.107	48.818	1	.000	2.108
stage3	.179	.216	.689	1	.406	1.196
macss	.389	.253	2.360	1	.124	1.475
anycom	-.109	.178	.376	1	.540	.896
Constant	-3.041	.208	213.754	1	.000	.048

SPSS Processor is ready

SAS

Output - (Untitled)

The SAS System 11:03 Tuesday, November 2

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.0413	0.2080	213.7547	<.0001
SEX	1	-0.0143	0.0980	0.0213	0.8839
AGE	1	0.00200	0.00324	0.3820	0.5365
stage2	1	0.7456	0.1067	48.8190	<.0001
stage3	1	0.1792	0.2158	0.6895	0.4063
MACSS	1	0.3885	0.2529	2.3603	0.1245
anycom	1	-0.1093	0.1784	0.3756	0.5400

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
SEX	0.986	0.813	1.195
AGE	1.002	0.996	1.008
stage2	2.108	1.710	2.598
stage3	1.196	0.784	1.826
MACSS	1.475	0.898	2.421
anycom	0.896	0.632	1.272

Stata

Results

```
Logistic regression                               Number of obs   =       6795
                                                  LR chi2(6)      =      53.21
                                                  Prob > chi2     =     0.0000
Log likelihood = -1653.4867                       Pseudo R2       =     0.0158
```

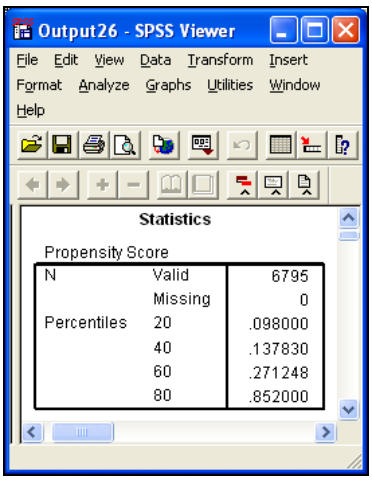
statin	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	-.0143113	.0980233	-0.15	0.884	-.2064335	.1778109
age	.0020023	.0032397	0.62	0.537	-.0043473	.008352
stage2	.7455493	.1067052	6.99	0.000	.5364109	.9546877
stage3	.1791637	.2157675	0.83	0.406	-.2437328	.6020603
macss	.3885197	.2528913	1.54	0.124	-.1071381	.8841775
anycom	-.1093198	.1783692	-0.61	0.540	-.458917	.2402774
_cons	-3.041324	.2080198	-14.62	0.000	-3.449035	-2.633612

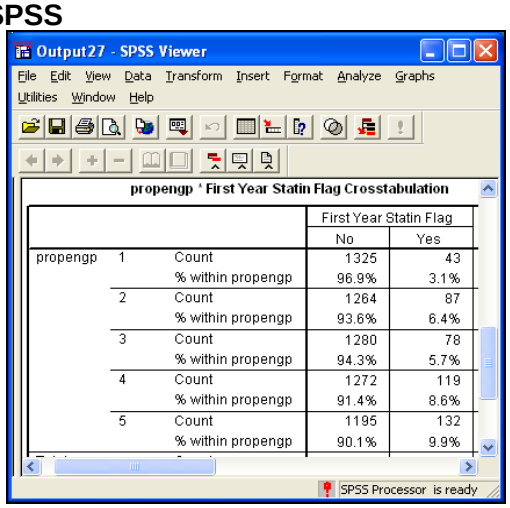
Use the logistic regression coefficients to construct a propensity rank, propen, for each individual in the file. This takes the form:

$$\text{propen} = \beta_1 \bullet \text{sex} + \beta_2 \bullet \text{age} + \beta_3 \bullet \text{stage2} + \beta_4 \bullet \text{stage3} + \beta_5 \bullet \text{macss} + \beta_6 \bullet \text{anycom}$$

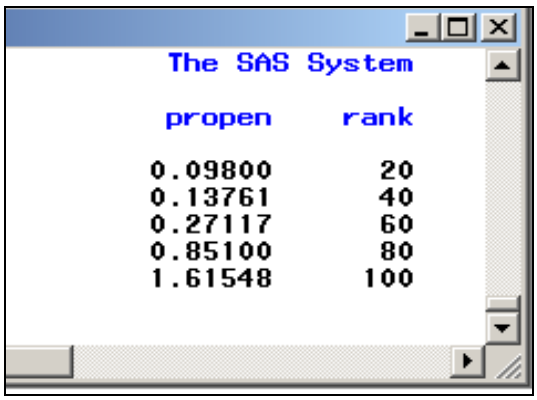
Assign these ranks and check that your mean score is 0.365979 with a standard deviation of 0.3338472. Run a frequency analysis on propen that includes the identification of cut-points for quintiles and compare your results with those shown opposite. Now create a propensity score group variable, propengp, which takes the values '1' to '5' with '1' representing patients in the lowest quintile of propensity to be prescribed a statin. Run a crosstabulation of statin by propengp to confirm the results shown above. Not surprisingly, the proportion of patients prescribed a statin early in their illness rises steadily across the propensity score quintiles. In a traditional propensity score analysis, this is the table from which random sampling would occur to achieve a balanced distribution of exposed and non-exposed subjects across each propensity score stratum. However, in this exercise we will use the propensity score to demonstrate effect modification.

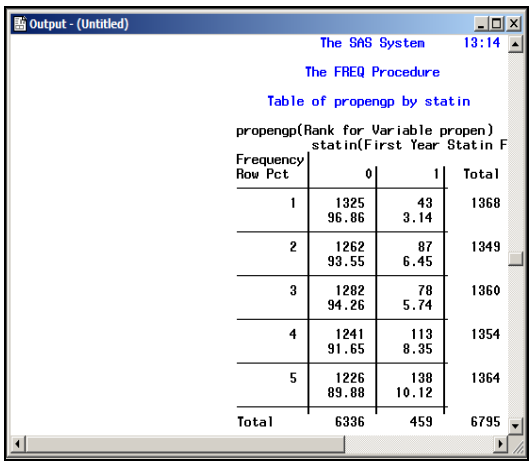
SPSS



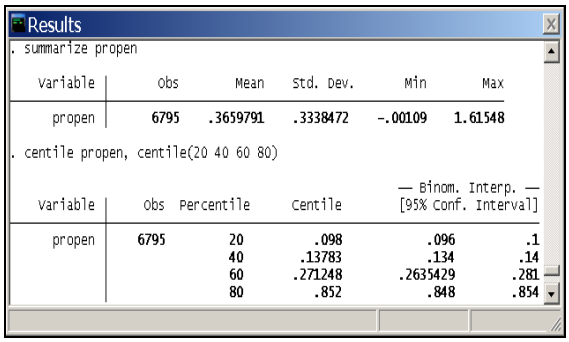


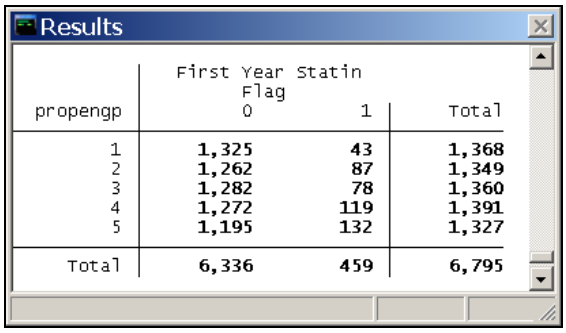
SAS





Stata





Effect Modification by Propensity Scores

13. To keep things simple we will compare only two collapsed strata of the propensity score; the lowest 40% (patients with lower propensity for statin treatment) with the remainder of patients (average to high propensity). Create a new variable, lopropen, by equating it to propengp. Recode lopropen so that values '3-5' become '0' and values '1-2' become '1' thus making it a binary indicator. Also create an interaction term, intstlo, equal to the product of statin and lopropen. **Stata users:** A quick way to create the interaction term is by using the xi command while running the regression which will automatically include the main effect and interaction terms in the modelling process. **R users:** As for Stata, the interaction term can be added directly into the model using the '*' syntax, or created separately first.

    Rerun your multivariate Cox regression model from step 9, but this time include lopropen and intstlo as covariates. Notice the relatively low p-value for the interaction term (p=0.084). It is uncommon for interactions in Cox and logistic regression models to be strong enough to yield small p-values, even with much larger sample sizes, due to the multiplicative rather than additive scaling of these models. This tells us something is up!

14. To better characterise the interaction for the purpose of scientific generalisation, we now remove the lopropen and intstlo covariates from the model and run two separate Cox regressions: one restricted to patients in the lowest 40% of propensity scores (lopropen=1) and the other restricted to patients in the higher 60% (lopropen=0). Confirm that the effect of statin in the low propensity patients is MRR = 1.54, whereas in the high propensity patients MRR = 0.63. These compare with our previous adjusted result of 0.73.

    You should now consider the following questions:

    A. On the basis of these fictitious results alone, what is our best estimate of the effect of early use of a statin in the average diabetic with appropriate clinical indications?

    B. How has the internal validity of this estimate been optimised in your analysis?

    C. Is it externally valid to generalise the estimate to all diabetics?

    D. How has the external validity of the estimate of effect been optimised in your analysis?

WARNING

Please remember that the practice data sets are 'pretend' and all of your results are fictitious. While the results of this practice exercise are not incredulous, you should draw no inferences from the exercise about the true nature of the occurrence relationship between mortality outcomes and use of statins in people with diabetes.

References

Glover J, Harris K, Tennant S. **A Social Health Atlas of Australia. Second Edition,** December 1999. Adelaide: Public Information Development Unit, The University of Adelaide, 1999.

Gotto AM, Kuller LH. **Eligibility for lipid-lowering drug therapy in primary prevention. How do the Adult Treatment Panel II and Adult Treatment Panel II guidelines compare?** *Circulation* 2002; 105: 136.

Mant A, Lowinger J, Hall W, *et al*. **Changing use of statins in Australia 1990-2004.** Abstract fro presentation at the National Medicines Symposium 2006. Available at: http://www.icms.com.au/nms2006/abstract/93.htm

REVIEW 5: 'RISK MODIFICATION AND ADJUSTMENT USING LINKED DATA'

Please check that you are familiar with the following topics:

- Propensity score analysis ☐
- Instrumental variable analysis ☐
- External validity and effect modification ☐
- Assessment of effect modification ☐
- Missing value representation ☐
- Address matching and use of geocodes, SEIFA and ARIA indices ☐

Please check that you understand the meaning of the following terms:

- Propensity score and confounder summarisation score ☐
- Propensity regression, propensity stratification and propensity matching ☐
- Instrumental variable ☐
- Restricted and unrestricted study domain ☐
- Interaction terms ☐
- Rothman's synergy index ☐
- Population-weighted quintile ☐
- Fractional polynomial ☐

Please check that you have demonstrated that you have acquired the following skills:

- Preparing a linked file for running regression models ☐
- Truncation of time series and follow-up to prevent artefacts ☐
- Use of time as an instrumental variable ☐
- Implementation of a propensity score analysis ☐