INTRODUCTORY ANALYSIS OF LINKED HEALTH DATA

Principles and Hands-On Applications

A Short-Course with

Professor David B. Preen

SHORT COURSE WORKBOOK

Version 3.9 November 2023

COURSE INFORMATION

Course Description

Introductory 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 introductory to intermediate level. Rapid growth in data linkage projects has led to a shortfall in analysts' skills. Some researchers understand epidemiological principles, but are unfamiliar with the specialised computing skills needed to analyse linked data files. Others have a strong grasp of computing concepts, but lack an adequate theoretical base to design high quality applications to answer research questions. This short-course endeavours to fill a gap in research training opportunities by combining the principles of health care epidemiology with hands-on practical exercises in the implementation of computing solutions.

The modular structure of the unit provides students with a theoretical grounding on each topic in morning sessions, followed by a training session on the corresponding 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 the hands-on exercises.

Course Content Overview

MODULE 1:	***INTRODUCTION TO DATA LINKAGE SYSTEMS***
Theory 1:	Case Studies of Data Linkage Systems
Training Session 1:	Use of Computing Syntax in Linked Data Analysis
Review 1:	Review of 'Introduction to Data Linkage Systems'
MODULE 2:	***CHARACTERISING POPULATIONS USING LINKED DATA***
Theory 2:	Theory of Data Linkage Methods and Applications
Training Session 2:	Preparing a Linked Data File and Characterising a Population
Review 2:	Review of 'Characterising Populations Using Linked Data'
MODULE 3:	***MEASURING HEALTH AND SERVICES USING LINKED DATA***
Theory 3:	Theory of Measurement of Health Events and States
Training Session 3:	Monitoring Health Care Utilisation with Linked Data
Review 3:	Review of 'Measuring Health and Services Using Linked Data'
MODULE 4:	***OUTCOMES RESEARCH USING LINKED DATA***
Theory 4:	Theory of Measurement of Health Outcomes
Training Session 4:	Health Outcomes Research with Linked Data
Review 4:	Review of 'Outcomes Research Using Linked Data'
MODULE 5:	***RISK ADJUSTMENT USING LINKED DATA***
Theory 5:	Theory of Risk Adjustment
Training Session 5:	Risk Adjustment with Linked Data
Review 5:	Review of 'Risk Adjustment Using Linked Data'

Learning Objectives

Introductory Analysis of Linked Health Data acquaints health researchers, clinical practitioners and managers with the theory and skills needed to analyse linked health data at the introductory to intermediate level. Upon completion the participant will:

- possess an overview of the theory of data linkage methods and features of comprehensive data linkage systems, sufficient to understand the sources and limitations of linked health data sets;
- understand the principles of epidemiologic measurement and research methods for the conceptualisation and construction of numerators and denominators used in the analysis of disease trends and health care utilisation and outcomes;
- understand sources of error in epidemiologic measurement, the difference between confounding and effect modification, and use of regression models in risk adjustment in health services research;
- be able to perform statistical analyses on linked longitudinal health data;
- be able to conceptualise and perform the manipulation of large linked data files; and
- be able to write syntax to prepare linked data files for analysis, derive exposure and outcome variables, relate numerators and denominators and produce results from statistical procedures.

Course Prerequisites

The computing component of the unit assumes a basic familiarity with computing syntax used in SPSS, SAS, Stata or R and methods of basic statistical analysis of fixed-format data files. There are no formal prerequisites in epidemiology or biostatistics for the course. However, it is recommended that participants who have not previously completed an introductory course in epidemiology familiarise themselves with the basic principles and terms used in that discipline. A working knowledge of statistical concepts, including regression models, used in data analysis in the medical and social sciences is assumed, however a brief video is available on LMS outlining the interpretation of statistical output for certain analyses that are performed during the course, as a refresher.

Helpful Syntax Writing Resources

If you are new to the world of using computer syntax for data analysis, rest assured you will not be the only one in the class. As this is an introductory-level course we provide instruction and resources to help you learn syntax throughout the course contact week. The course is also 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. Introductory videos on how to get started with each software package are available on LMS and complete step-by-step video walkthroughs for each day's exercise in SPSS, SAS, Stata and R, are also provided.

In addition, some useful resources are available through the weblinks below which provide introductory information on the use, rules, structure, main commands and implementation of statistical syntax. If you are a novice in code writing for data analysis, it is *highly recommended* that you review the information related to the statistical software package you plan to use on LMS and through the links below prior to the Monday of the course contact week as this will ensure you don't get bogged down in the first few days

SPSS
https://stats.idre.ucla.edu/spss/
https://libguides.library.kent.edu/SPSS/Syntax

Stata
https://stats.idre.ucla.edu/stata/
https://libguides.library.kent.edu/statconsulting/stata

SAS
https://stats.idre.ucla.edu/sas/
https://libguides.library.kent.edu/SAS/syntax

R
https://stats.idre.ucla.edu/r/
https://libguides.library.kent.edu/statconsulting/r

Debugging Syntax

Be aware that it is common for computing syntax not to always work correctly the first time it is run. Some more common errors that analysts experience (and which you should be familiar) include the following:

1. ***Grammatical or typographical errors*** in the way the syntax is written. These include omission or incorrect expression of a keyword or punctuation. These are usually the least dangerous errors, because the syntax crashes and generates an error message.
2. ***Minor logical errors***: These include omission or erroneous addition of a code to a list of target codes or incorrect expression of a relational operator such as > instead of ≥. These mistakes tend to cause marginal errors in outputs that may be inconsistent with other results.
3. ***Major logical errors***: These errors vary widely and require careful checking of file effects and outputs to uncover. They can be extremely dangerous to the integrity of the research.
4. ***Data file problems***: These include using the wrong file, using a file sorted in the wrong sequence and using a file that has been corrupted by a previous process. An example of the latter is when a file is resaved after some records have been mistakenly deleted. The consequences can range from marginal discrepancies through to very dangerous errors.

Class Participation and Adult Learning Philosophy

Introductory 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. Bibliographies are given for those who wish to read further.

Instruction Materials

Detailed course notes are provided. These are the subject of copyright and should not be reproduced without permission from the course instructor.

Course Instructor's Contact Details

Professor David Preen
Chair in Public Health
School of Population and Global Health, The University of Western Australia
35 Stirling Hwy, Crawley, Western Australia 6009.
Tel: +61 8 6488 1251; E-mail: david.preen@uwa.edu.au

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

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 80 competitive research grants as chief investigator (totalling >$58.2 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 330 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 multiple 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 12 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.

About the Course Founder

Winthrop Professor D'Arcy Holman
MBBS LLB(*Hons*) MPH(*Harv.*) PhD GAICD FACE FPHAA FAFPHM FAIM AM
Chair in Public Health 1994-2014
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. CASE STUDIES OF DATA LINKAGE SYSTEMS

1.1 Traditions of Health Data Integration

Historical Origins of Health Data Integration

With the explosions of diagnostic testing, medical specialisation and therapeutic diversity, and completion of the epidemiologic transition to chronic diseases following World War II, a team approach to medical care emerged and the old paradigm of the patient record as the personal musings of a lone physician became unsustainable. At the same time the cost spiral of health care began its upward climb and national health care access programs appeared in many countries with their attendant concerns over the efficient and equitable distribution of health care resources. Against this backdrop of accelerating complexity and its attendant challenges in medical informatics, two different species of health data integration luminary emerged, typified by Professor Lawrence Weed in the USA and Sir Don Acheson in the UK.

Weed tradition of health data integration: During the 1960s, as Professor of Medicine at the University of Vermont, Professor Lawrence Weed devised his system for the organisation of patient information known as the *Problem-Oriented Medical Record*. His major work on the topic appeared in 1969, and thereafter his ideas spread remarkably fast around the world. He believed that "medical care and education should be approached with scientific discipline, where intelligible records are produced, which not only prove adequate care, but which also contribute to it. Records shouldn't be ink splashes and private scribbling, but the concern of those who in the future will have to depend on those records for the continuing care of the patient." But Weed's vision went beyond tidying up the mess in the medical records department. He identified the potential for emerging computer technology to facilitate data integration and more effective communication between clinicians. When his mini-computer-based *PROMIS*, the *Problem-Oriented Medical Information System*, was implemented in 1970 on several wards of the Medical Centre of Vermont, it was said to be the first computerised medical record system in the world.

Acheson tradition of health data integration: The second species of post-War data integration luminary was exemplified by the eminent epidemiologist and public administrator, Sir Don Acheson, who was the inaugural Director of the *Oxford Record Linkage Study* in the 1960s and later a Chief Medical Officer of England. As early as 1961, with coauthors Truelove and Witt, he posited in the *BMJ* that the analysis of medical records linked nationally would "provide excellent morbidity statistics and create a science of prognosis. The study of human genetics would be transformed, provided the patients' identifying numbers were also inserted on birth certificates. These would be a base from which field studies of the epidemiology of the important diseases of the day would spring". Thus the *Oxford Record Linkage Study* was born and, while it was not the only project of its type at the time, and the concept of data linkage had been described in the *American Journal of Public Health* by Dr Hubert Dunn in 1946, the Oxford project was arguably the first to achieve a considerable degree of international notoriety.

Gertzog's Analysis of Differences between the Weed and Acheson Traditions: That the tradition of health data integration championed by Sir Don Acheson, conceived by a UK epidemiologist in the womb of nationalised medicine, was entirely distinct from the tradition championed by Professor Larry Weed, an eminent US clinician, practising at the forefront of individualistic medicine, was a point not lost on commentators of the day. Jack Gertzog, an operations research analyst, published a commentary in 1970, pointing out the differences and some similarities between "two of the more innovative students of medical records", namely Weed and Acheson.

Gertzog concluded that while Weed was concerned about logical fragmentation in clinical decision-making, Acheson was concerned about institutional fragmentation as a barrier to planning and evaluation. He noted that unlike Acheson, Weed's principal focus was not to bring together all available data persons and populations, but rather the relevant data about a patient's particular medical problem,

a feature that persisted when Weed later described the concept of *knowledge couplers*, whereby the patient's problem is further linked to decision-making algorithms, systematic reviews and clinical guidelines. Gertzog noted that both Weed and Acheson wished to improve communication, but for Weed this was communication between clinicians, whereas Acheson sought to improve communication between institutions relevant to national planning and their databases. These essential differences are summarised below.

	Traditions of Health Data Integration	
	Weed Tradition	**Acheson Tradition**
Motivating concern	Logical fragmentation in clinical decision-making	Institutional fragmentation as a barrier to planning and evaluation
Focus for integration	The patient's medical problem	The person as a whole and as a member of the population
Communication objective	Improved communication between health care providers	Improved communication between institutions and their databases
Contemporary manifestation	Electronic health records	Data linkage systems

Contemporary Manifestations of the Weed and Acheson Traditions

Discourse on health data integration continues to follow two lines of development. One, in the Acheson tradition, is the linkage of administrative and clinical databases for the purposes of research, planning and evaluation. The other, in the Weed traditions, is technical solutions for clinical data sharing and electronic health records to improve and audit the care of individual patients. The potential for these two worldwide trends to merge, synergise, exist in isolation or antagonise reflects a multiplicity of different histories, institutional arrangements and legal frameworks, as well as the levels in the health system where integration projects take place.

Australia has been a world leader in taking the two varieties of health data integration strategy down separate paths, albeit debatable whether the Australian framework has arisen by design or happenstance. There are many issues, including pros and cons, surrounding the question of whether or not the Weed and Acheson traditions of health data integration should be brought closer together.

Definition of 'e-Health'

The term e-health encompasses a set of disparate concepts, including health, technology, and commerce. There are well over 50 different published definitions of e-health and thus to assign any specific significance to the term may be unwise; certainly one needs to be aware that the terms is used very loosely and has potential to have different meanings to different people. The 58th World Health Assembly essentially defined e-health as the combined use of electronic communication and information technology in the health sector: "eHealth is the cost-effective and secure use of information and communications technologies in support of health and health-related fields, including health-care services, health surveillance, health literature, and health education, knowledge and research" (resolution WHA58.28).

1.2 Data Linkage Systems as Research Infrastructure

In keeping with the Acheson tradition, infrastructure to enable research has become a major planning focus in many countries, including Australia, New Zealand, Singapore, Canada, the UK, other members of the European Union and the USA. Several countries have undertaken national audits of their stock of research infrastructure, which in the health sector includes biomedical laboratories, imaging equipment, biobanks and supercomputers.

Concepts about enabling national infrastructure for research on population health and welfare have existed for over half a century. In 1946, Dr Halbert Dunn proposed:

> *"Each person in the world creates a **Book of Life**. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book into a volume".*

Thus Dunn developed the concept of collating the records into a personal file and named the process record linkage. He predicted that statistical analysis of linked records would be useful to health and welfare agencies in evaluating their service programs. He suggested also that linking files of data would establish the accuracy of the source records.

All of Dunn's predictions have been proven to be correct in theory, albeit in practice comprehensive national systems of linked health and welfare data still remain difficult to achieve. Successful implementation demands high levels of community capital and inter-agency cooperation before the many diverse stakeholders will relinquish their immediate interests in favour of a more futuristic and publicly-spirited common vision. It was not surprising, therefore, that an international review by Professor Leslie Roos and others in 2004 described only a handful of information-rich environments worldwide, which have been constructed from the linkage of multiple, large, population-based, administrative and clinical data sets. At around that time they included the Manitoba Population Health Information System, Oxford Record Linkage Study, Scottish Record Linkage System, Rochester Epidemiology Project, the British Columbia Linked Health Database, the WA Data Linkage System (WADLS) and the Centre for Health Record Linkage (CheReL) in New South Wales. However, since that time several new comprehensive data linkage systems have appeared, including additional systems in the Canadian provinces, the Secure Anonymised Information Linkage (SAIL databank) system in Wales, systems emerging in other areas of the UK and the commencement of construction of Australia's national data linkage system, the Population Health Research Network (PHRN).

International Population Data Linkage Network

A number of the data linkage centres around the world met in London in December 2008 to inaugurate the International Health Data Linkage Network (IHDLN). The aims of the Network were: (i) to establish and maintain an effective and useful network of health data linkage centres; (ii) to foster collaboration and exchange programs between data linkage centres; (iii) to produce a compendium of measurements based on linkage of health data across Australia, Canada and the United Kingdom; and (iv) to record the outputs from health data linkage activities and programs across the globe. In 2014 the IHDLN was rebranded to the International Population Data Linkage Network (IPDLN) which currently has well over 100 member organisations involved with data linkage. The website for the Network is found at: http://www.ipdln.org.

Ad Hoc vs Systematic Data Linkage

Data linkage activities may be undertaken to support just one or a small number of research projects and in this form it is widely practiced worldwide in clinical, health services and public health research, often to connect records of harmful exposures or beneficial health care interventions with records of health outcomes. Alternatively, linkage of health data may be undertaken on a proactive and systematic basis for health data pertaining to an entire population with a view to supporting an indefinite number of future (and as yet undefined) research projects. The contrasting characteristics of these approaches are shown below.

	***Ad hoc* Data Linkage**	**Data Linkage System**
Purpose	Supports one research project (or a small number) with known objectives.	Supports an indefinite number of mostly unknown future research projects.
Data sets	Limited to those needed for the known research objectives (often 2-3 data sets).	Unlimited – the more data sets the more versatile and effective is the system.
Data requirements	Usually partial identifiers and clinical data come together.	Only requires partial identifiers – clinical data can be sought later on a project by project basis.
Time of activity	Data linkage activity closes once links between the specified data are in place.	Open-ended and requires continuous updates of the links as new data arrive.
Storage of links	Usually links are stored as an integral part of the research project data.	Requires a dedicated storage mechanism for links = the **master linkage key**.
Funding	Usually draws on the research grant used to fund the project.	Requires dedicated infrastructure funding for a central, ongoing unit.

The principles covered in this course are sufficiently generic to be useful in the analysis of data arising from *ad hoc* applications of data linkage as they are for applications based on data linkage systems. However, the background assumption is always that the rich resources of a comprehensive data system are available as the enabling research infrastructure.

1.3 A Note about Data File Types

By way of explaining one of several important distinctions between the introductory and advanced courses in this curriculum on the analysis of linked health data, it is useful to distinguish between three types of data files:

- ***Type 1 files*** have just one valid record per individual; eg, death registrations.
- ***Type 2 files*** have mostly one record per individual; eg, cancer notifications.
- ***Type 3 files*** have mostly multiple and variable records per individual; eg, hospital separations.

A linked data analyst is said to be working with a ***simple file group*** if there is only one type 3 file in the group that contains a census of all individuals included in the analysis. An analyst is said to be working with a ***complex file group*** if it includes two or more type 3 files or none of the files contains a full census of the all individuals included in the analysis. This introductory course deals only with the analysis of simple file groups. The advanced course deals with the analysis of complex file groups.

1.4 Case Study of the WA Data Linkage System

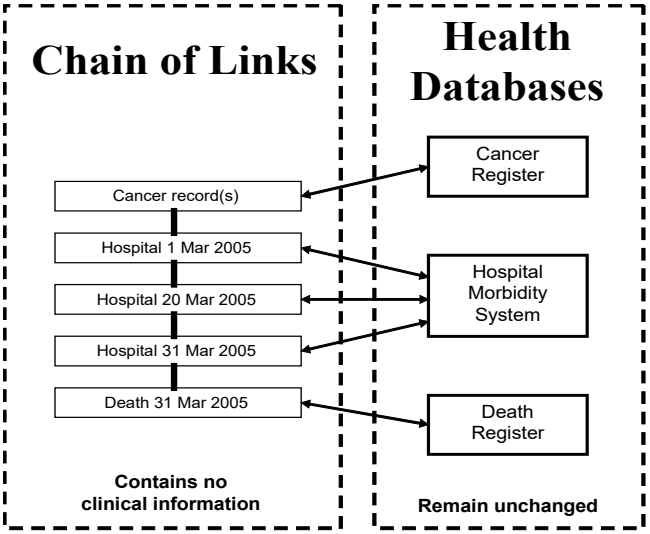
Scope of the Western Australian Data Linkage System

The WADLS is a complex, multi-set system for the creation, storage, update and retrieval of links between health and welfare-related data. It uses computerised probabilistic matching and clerical review to create a dynamic master linkage key between over 40 population-based administrative and research health data collections in WA. The linkages mean that the total historical population can be researched for all major diseases, their risk and protective factors and the utilisation and outcomes of health services. Data linkage within the WA Department of Health, consists of linkage systems development and client support groups with specific responsibilities for quality control and liaison with data custodians and researchers. A schematic diagram of the WADLS appears below. The website for the system is available at: http://www.datalinkage-wa.org/.

Source: WA Data Linkage Branch

Design of the Western Australian Data Linkage System

A key aspect of the design of the WADLS is the separation of linkage-related processes from those operating on sensitive clinical and service data. The WADLS is neither a centralised data repository nor a distributed virtual data warehouse. Rather, the information retained in the system consists of pointers to the source data elements. Examples of such data elements are a single record in the hospital morbidity data system; a set of records for a single patient in the cancer register; or a single record in the death register. The elements for a deceased individual might therefore include a cancer registry entry, three hospital separation records and a death registration, as depicted below.



This introduces the concept of a chain of links, in which pointers to health-related data elements are ordered chronologically for each person. Thus one chain represents all of the known events in the life of a single person, much like an index to Dunn's 'book of life'. The chain itself does not contain any clinical information.

Benefits of the Western Australian Data Linkage System

The total marginal cost of implementing a data linkage system in WA, estimated to have been Au$3.4 million over the first 10 years, has been inexpensive in comparison with biomedical research infrastructure. The advantages and social benefits arising from this investment are now apparent. Some benefits were anticipated from the outset. Others, such as the conservation of patient privacy in research, have only come to light with experience. The following are distinct areas where the benefits of data linkage have accrued in approximate ascending order of importance:

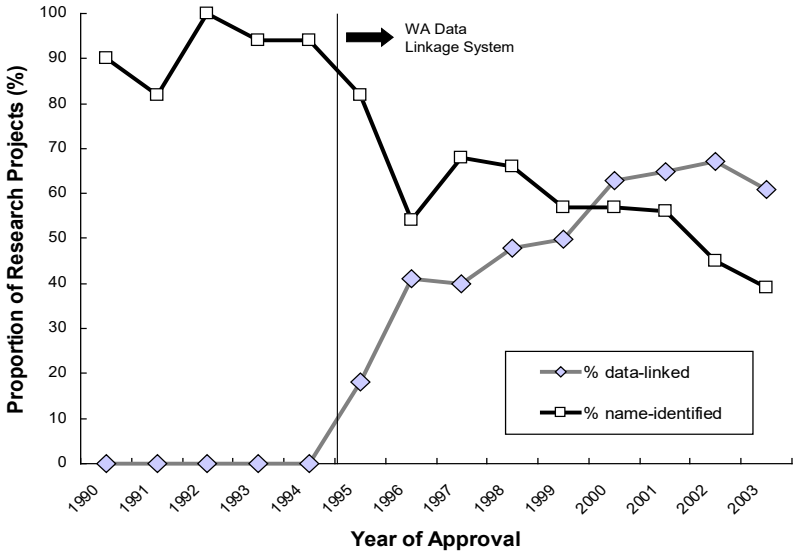
Commercial and competitive benefits: The WADLS has endowed its research community with a competitive advantage in attracting funds relative to research communities with no data linkage system. The resultant flows of research income from outside WA, including several NIH grants from the US and many large Australian NHMRC grants, summed to Au$58 million within the first decade (the figure had reached Au$129 by 15 years) and when discounted at 4% p.a. back to the years of infrastructure expenditure represented greater than a ten-fold return on investment. This revenue provided employment and stimulated the economy. We believe that WA has also been relatively more successful in attracting and retaining individual research leaders of high quality. This has enhanced the intellectual life of the State and has resulted in flow-on benefits in the local capacities for policy development and quality improvement in the health system.

Adding value to existing information assets: Data linkage has generated a greater return on the substantial existing investment in routine administrative data sets, which had often been perceived in the past to be unproductive record-keeping systems. It has added value to existing data sets through the improvement of data quality. Tens of thousands of duplication errors in records have been corrected

and other technical glitches with data have come to light, leading to greater accuracy of recording at an administrative level.

Cost-efficiency of research: Linking existing data has proven to be a cheap and effective alternative to performing *de novo* longitudinal studies and other more traditional approaches to epidemiologic and health services research, especially in mobile and vulnerable populations. This has conserved the limited resources available to support health and medical research, thus enabling more research to be performed with a given research budget.

Conservation of patient privacy: The privacy of individuals has been conserved by reducing the need to release names and other personal identifiers to researchers, because a major reason for release of identifiers in the past (to find and clerically collate information on patients) has been removed. Contrasted with consent-based responses to privacy concerns, data linkage systems are superior because they conserve the privacy of all patients, regardless of whether or not they would have given consent to the use of information. A consent-based approach conserves the privacy only of those who decline to participate, and this may be at the cost of irretrievable impracticability of the research. As shown below, epidemiological and health service research projects using named data provided from major WA health collections fell from >90% prior to the WADLS to 36% in 2003.



Community development: Although difficult to quantify, the WADLS has enhanced the level and quality of interactions between researchers, clinicians, administrators, consumer groups and the mass media. The Quality of Surgical Care Project and the *Health+Medicine* newspaper supplement are examples. It has provided a focal point for envisioning future possibilities, for improved cooperation and for rigorous debate about the uses of data and the results of subsequent research. This has strengthened the community by exercising the machinery necessary for teamwork and other organised responses to health and social problems.

Contributions to medical and scientific knowledge: The WADLS has provided data linkage and related services to researchers in academic, community, health industry and government organisations, whose work had yielded over 400 distinct studies by 15 years. The outputs included over 250 scientific journal publications and over 35 graduate research degrees, completed or in progress.

Improvements in population health: These have occurred at least at two levels. Contributions to knowledge published in peer-reviewed journals have contributed to worldwide medical and scientific knowledge, which will eventually translate into better programs of preventive services, treatment and care for human populations. Specific local improvements in population health include Australia's first 30-day surgical mortality audit program. Reforms in mental health legislation and service delivery can also be traced directly to WA data linkage-based research.

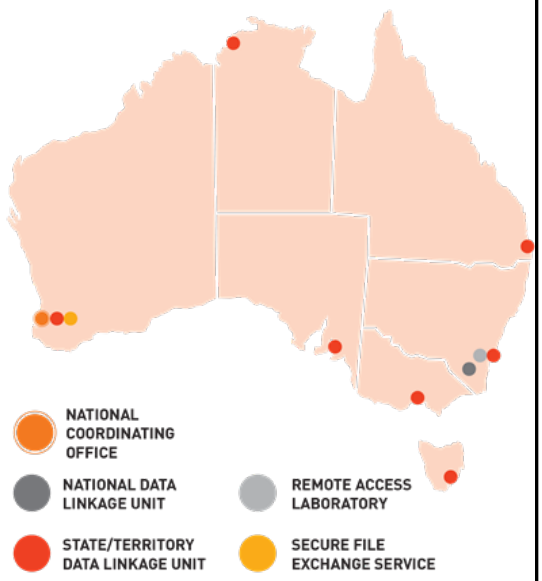
The creation of a data linkage system is principally a challenge in social organisation. It demands leadership, inter-agency and inter-sectoral cooperation and perseverance. It is the relative scarcity of

these community resources that explains why comprehensive, population-based data linkage systems are difficult to create and equally challenging to sustain. The requirements for technological solutions and resources are overshadowed by the importance of high-quality relationships between leading individuals and between institutional players. Overcoming privacy concerns is also largely a problem in underlying interests and inter-agency relations, especially when one considers that data linkage is the most effective intervention available to conserve patient privacy in a research-rich environment.

1.5 Australia's Population Health Research Network

Under the Australian PHRN, the existing data linkage systems in WA and NSW has been replicated across the other states. The current operating data linkage systems and capacity across the country are shown below.

Data Linkage Unit	Location
AIHW	AIHW, ACT
Centre for Data Linkage	Curtin University, WA
Data Linkage Branch	Department of Health, WA
SA-NT DataLink	University of South Australia
CHeReL	Ministry of Health, NSW
Centre for Victorian Data Linkage	Department of Health and Human Services, Vic
Tasmanian Data Linkage Unit	University of Tasmania
Queensland Research Linkage Group	Queensland Health



Source: PHRN (www.phrn.org.au)

The Australian approach to health data integration under the Acheson tradition use in the WADLS, the CHeReL in New South Wales and the national PHRN is typified by a number of distinct features, which are notable but by no means unique.

- Individual *consent* is not mandatory for any processes, regardless of whether or not the databases are statutory collections. Instead strong *community and consumer participation* is encouraged
- The use of *probabilistic matching*, with or without clerical checking, to identify valid links.
- A *chain of links* on each individual held centrally, created solely using partial identifiers and contains no clinical information.
- Absence of a *central repository*: rather, the majority of datasets remain exactly as they were, peripherally, under the control of their respective data custodians.
- Researcher requesting linked health data are provided with separate anonymised but linkable files from each of the data custodians through use of a *best practice protocol*.

In some other countries, methods akin to this are known as the *honest broker* approach and, indeed, data linkage units in Australia act as honest brokers between data custodians and researchers.

1.6 International Variation in Data Linkage Systems

Any student of the world's comprehensive data linkage systems will be struck by the diversity of approaches and applications that have emerged. Some systems rely principally on probabilistic matching whereas others rely principally on a unique and reliable national health number. Some have supported mostly applications in clinical epidemiology whereas others have been used with a stronger social science focus. The data linkage systems across the world vary in their design according to a number of key attributes, some of which and their common variations are set out below.

Attribute	Common Variations in System Design
Scope of sectoral coverage	Health only vs health + welfare
Scope of person-based linkage	Individual only vs individual + genealogical
Consent	No individual consent vs opt out vs opt in
Predominant linkage technique	Deterministic vs probabilistic vs probabilistic + clerical review
Storage of links	Chain of links vs look up tables
Storage of substantive data	Central repository vs distributed custodial databases
End user access	Restricted personnel vs safe haven vs unrestricted
Type of data release	Named vs anonymised
Types of uses	Clinical epidemiology vs social science uses vs both

1.7 Summary of Desirable Design Criteria for a Data Linkage System

Regardless of the approach to system design, there do appear to be some general conclusions that can be drawn about what makes for a data linkage system that performs well. The following are general guidelines for desirable design criteria for comprehensive data linkage systems:

Population: The population covered by the system is ideally geographically defined and relatively stable, of adequate size for most analyses (ideally >1 million) and with adequate longitudinal coverage (ideally >10yr). The advantage of a geographically defined population, as distinct from register-based populations such as members of a health insurance plan, is that research based on the former will enjoy greater external validity, being more representative of the general community.

Data resources: Health events and other data should be ascertainable for the whole population (not merely a sample or population subgroup) on a continuous basis (not an *ad hoc* or intermittent basis). The range of data resources should be sufficient to address the research agendas of primary interest. Linked hospital morbidity data and deaths are arguably the minimal system for health data linkage. Inclusion of a broader range of human services databases, covering education, community services, housing and justice is desirable. Ideally the system should include a population register such as a national health insurance enrolment file or an electoral roll.

Technical facilities: Linkage should be achieved by either high performance probabilistic matching or a reliable unique person identification number or a hybrid of these methods. A program of data validation and assessment of linkage performance is highly desirable. A tracing system is desirable to censor patients who emigrate from the geographic area and are therefore lost to longitudinal follow-up. Geocoding of addresses will enhance the capacity for spatial analysis and assignment of socio-economic status and, where relevant, remoteness indices. Genealogical linkage will provide capacity for genetic epidemiology and family studies.

Organisational supports: These include a governance structure for leadership and management of the data linkage enterprise; instruments for co-operation between agencies providing data (eg, MOUs); consumer participation; a facilitative or at least non-obstructive legal framework; and multidisciplinary research teams including collaboration with clinicians and health service providers.

References:

Acheson ED. **Oxford record linkage study: a central file of morbidity and mortality records for a pilot population.** *Br J Prev Soc Med* 1964; 18: 8-13.

Acheson ED. **Medical Record Linkage.** London: Oxford University Publications, 1967.

Acheson ED (Ed). **Record Linkage in Medicine. Proceedings of the International Symposium, Edinburgh, July 1967.** London: ES Livingstone, 1968.

Acheson ED, Truelove SC, Witts LJ. **National epidemiology.** *BMJ* 1961; 1: 668.

Armstrong BK, Kricker A. **Record linkage – vision renewed.** *ANZ J Public Health* 1999; 23: 451-452.

Arnardottir OM, Bjorgvinsson DT, Matthiasson VM. **The Icelandic Health Sector Database.** *Eur J Health Law* 1999; 6: 307-362.

Australian Health Ministers' Conference. **National e-Health Strategy.** Melbourne: Victorian Department of Human Services, 2008.

Baldwin JA. **The Oxford Record Linkage Study as a medical information system.** *Epidemiol Prev Med* 1972; 65: 237-239.

Blakely T, Woodward A, Salmond C. **Anonymous linkage of New Zealand mortality and Census data.** *ANZ J Public Health*, 1999; 24: 92-95.

Boyd J. **Linking Health Information & Access to National Data.** Edinburgh: Information & Statistics Division, National Health System in Scotland, 2004.

Brook EL, Rosman DL, Holman CDJ. **Public good through data linkage: measuring research outputs from the Western Australia Data Linkage System.** *ANZ J Public Health* 2008; 32: 19-23.

Chamberlayne R, Green B, Barer ML, *et al*. **Creating a population-based linked health database: a new resource for health services research.** *Can J Public Health* 1998; 89: 270-273.

Clark A, Preen DB, Ng JQ, Semmens J, Holman CDJ. **Is the Western Australian population representative of other Australian States and Territories?** *Australian Health Review* 2010; 34(10): 210-215.

Dunn HL. **Record linkage.** *Am J Public Health* 1946; 36: 1412-1416.

Ford DV, Jones KH, Verplancke J-P, *et al*. **The SAIL databank: building a national architecture for e-health research and evaluation.** *BMC Health Serv Res* 2009; 9: 157 [doi: 10.1186/1472-6963-9-157].

Gertzog J. **Changing concept of medical records.** *Public Health Rep* 1970; 85: 673-679.

Gill L, Goldacre M, Simmons H, *et al*. **Computerised linking of medical records: methodological guidelines.** *Epidemiol Community Health* 1993; 47: 316-319.

Glasson EJ, de Klerk NH, Bass AJ, *et al*. **Cohort profile: The Western Australian genealogical project.** *Int J Epidemiol* 2008; 37: 30-35.

Haggar FA, Preen DB, Pereira G, Holman CDJ, Einarsdottir K. **Cancer incidence and mortality trends in Australian adolescents and young adults, 1982-2007**. *BioMed Central Cancer.* 2012; 12: 151 – 161

Havard A, Jorm L, Preen D, Daube M, Kemp A, Einarsdottir K, Randall D, Tran D. **The Smoking MUMS (Maternal Use of Medications and Safety) Study: protocol for a population-based cohort study using linked administrative data.** *British Medical Journal Open* 2013; 3: e003692. doi:10.1136/bmjopen-2013-003692.

Hobbs MST, McCall MG. **Health statistics and record linkage in Australia.** *J Chron Dis* 1970; 23: 375-381.

Holman CD, Bass AJ, Rosman D, *et al*. **A decade of data linkage in Western Australia: Strategic design, applications and benefits of the WA Data Linkage System.** *Aust Health Rev* 2008; 32: 766-777.

Holman CDJ, Bass AJ, Rouse IR, *et al.* **Population-based linkage of health records in Western Australia: development of a health services research linked database.** *ANZ J Public Health* 1999; 23: 453-459.

Holman D, Bass J, Rosman D, *et al*. **Workshop on data linkage in epidemiologic research.** *Aust Epidemiol* 2003; 10: 14.

Kelman CW, Bass AJ, Holman CDJ. **Research use of linked health data - a 'best practice' protocol.** ANZ J Public Health 2002; 26: 251-255.

Kendrick S, Clarke J. **The Scottish Record Linkage System.** *Health Bulletin*, 1993; 51: 72-79.

Kurland LT, Molgaard CA. **The patient record in epidemiology.** *Scientific American* 1981; 245: 54-63.

Lawrence G, Dinh I, Taylor L. **The Centre for Health Record Linkage: a new resource for health services research and evaluation.** *Health Inf Man J* 2008; 37: 60-62.

Lyons RA, Jones KE, John G, *et al*. **The SAIL databank: linking multiple health and social care datasets.** *BMC Med Informatics and Decision Making* 2009; 9: 3 [doi:10.1186/1472-6947-9-3].

Malacova E, Kemp A, Hart R, Jama-Alol K, Preen DB. **Effectiveness of in vitro fertilization in women with previous tubal sterilization.** *Contraception* 2015; 91(3): 240-244.

National Collaborative Research Infrastructure Strategy [Australian Government website]. **Population Health Research Network.** Available at: http://ncris.innovation.gov.au/Capabilities/Pages/PopHealth.aspx.

Oh H, Rizo C, Enkin M, *et al*. **What is e-health: a systematic review of published definitions.** *Med Internet Res* 2005; 7: e1 [doi:10.2196/jmir.7.1.e1].

Preen DB, Moorin RE, Einarsdottir K, Kemp A, Gibson D. **Approaches to health services research and health policy evaluation using longitudinal, whole-population medical record linkage.** *International Journal of Longitudinal and Life Course Studies* 2010; 1(3): 59.

Roos LL, Menec V, Currie RJ. **Policy analysis in an information-rich environment.** *Soc Sci Med* 2004; 58: 2231-2241.

Roos LL, Nicol P. **A research registry: uses, development, and accuracy.** *J Clin Epidemiol* 1999; 52: 39-47.

Roos NP, Black CD, Frohlich N, *et al.* **A population-based health information system.** *Med Care* 1995; 33: DS13-20.

Sibthorpe B, Kliewer E, Smith L. **Record linkage in Australian epidemiological research: health benefits, privacy safeguards and future potential.** *Aust J Public Health* 1995; 19: 250-256.

Stanley FJ, Croft ML, Gibbons J, *et al*. **A population database for maternal and child health research in Western Australia using record linkage.** *Paed Perinatal Epidemiol* 1994; 8: 433-447.

Stanley FJ, Read AW, Kurinczuk JJ, *et al*. **A population maternal and child health research database for research and policy evaluation in Western Australia.** *Semin Neonatol* 1997; 2: 195-201.

Trutwein B, Holman CDJ, Rosman DL. **Health data linkage conserves privacy in a research-rich environment.** *Ann Epidemiol* 2006; 16; 279-280.

Weed LL. **Medical Records, Medical Education, and Patient Care.** Cleveland: Case Western Reserve University Press, 1969.

TRAINING SESSION 1:
USE OF COMPUTING SYNTAX IN LINKED DATA ANALYSIS

IMPORTANT NOTICE

The practice data sets provided for training purposes in this course are synthetic. All results based on analysis of these synthetic/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.

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 in the files have been modelled to mimic qualitatively, but not quantitatively, the type of results observed in the studies using real data. Thus the exercises endeavour to simulate the analysis of real linked health 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 known as ***end of record loading***. Again, it is emphasised that EOR loading is not the only possible technique, nor necessarily the best 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. Although syntax solutions are provided, you should attempt to achieve the desired result by writing your own syntax first.
5. Seek assistance from your computing 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. 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 training session, you are going to become familiar with the use of computing syntax to work with single files of internally linked data. Although fictitious, they are modelled on data used in WA to study the risk of cancer in men who had undergone an in-hospital vasectomy. While some events, such as death or a first-time hospital admission, can occur only once to a person, other events may occur multiple times. The occurrence of hospital admission or a condition like cancer falls into the latter category although in different ways. It is common for people to have multiple (sometime many) hospital admissions over the lifecourse and therefore a hospital registry will usually have many rows per patient providing an example of a ***type 3 data file*** structure. It is also possible for one person to develop several different primary cancers during their lifetime, although most cancer patients develop only one cancer (of organs other than skin in Australia) and so the majority of people in cancer registers contribute one record per person with a much smaller number having two or more records. Thus, cancer registry data provide an example of a ***type 2 file***.

Training Session Resources

The following data files are used in this training session:

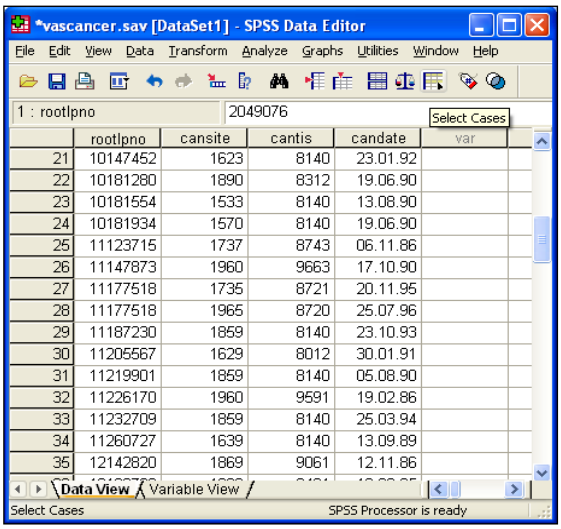
If you are using this statistical package …	*… you should use this save file.*
SPSS	vashmds.sav vascancer.sav
SAS	vashmds.sas7bdat vascancer.sas7bdat
Stata	vashmds.dta vascancer.dta
R	vashmds.rda vascancer.rda

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.

General Principles for Working with Linked Health Data

Open the vascancer file (the SPSS version is below as an example of the data file structure) and see it contains four variables: *rootlpno* (as above), cansite (an ICD-9 diagnosis code), cantis (an ICD-O tissue code) and candate (the date of diagnosis). Note that the structure of linked data sets is influenced by the structure of the initial data sets. In this case we have a rectangular file structure and it is possible to have multiple lines of data per individual, as shown by the two cancer records for rootlpno 11177518. This may become apparent visually only when the records are sorted in ascending order of rootlpno, so records for the same individual are blocked together. Usually, the ***preferred sequence*** of records in an internally linked file is sorted chronologically by individual; ie, in this instance, by ascending candate within ascending rootlpno. The presence of the rootlpno variable combined with the preferred sequence of records in any form of linked morbidity or service encounter file allows specific types of analyses to be performed that would not otherwise be possible.

SPSS



	rootlpno	cansite	cantis	candate	var
21	10147452	1623	8140	23.01.92	
22	10181280	1890	8312	19.06.90	
23	10181554	1533	8140	13.08.90	
24	10181934	1570	8140	19.06.90	
25	11123715	1737	8743	06.11.86	
26	11147873	1960	9663	17.10.90	
27	11177518	1735	8721	20.11.95	
28	11177518	1965	8720	25.07.96	
29	11187230	1859	8140	23.10.93	
30	11205567	1629	8012	30.01.91	
31	11219901	1859	8140	05.08.90	
32	11226170	1960	9591	19.02.86	
33	11232709	1859	8140	25.03.94	
34	11260727	1639	8140	13.09.89	
35	12142820	1869	9061	12.11.86	

As we will discover in this course and as you work with linked data, some of the possibilities include:

- Identification of the first event of a particular kind (often called the ***index event***); eg, the first diagnosed cancer or the first hospital admission for ischaemic heart disease.
- Identification of past or co-existing events of a particular kind (often called ***comorbid conditions***); eg, a previous history of a specific disease such as pre-existing diabetes or previous myocardial infarction.
- Calculation of time (often called ***survival time***) from one event to another; eg, the time between first and second diagnosed cancers or from initial cardiac revascularisation procedure to the next procedure.
- Calculation of ***cumulative scores*** summed over any number of relevant records pertaining to the same individual; eg, cumulative length of stay in the first year following an index admission for renal disease.

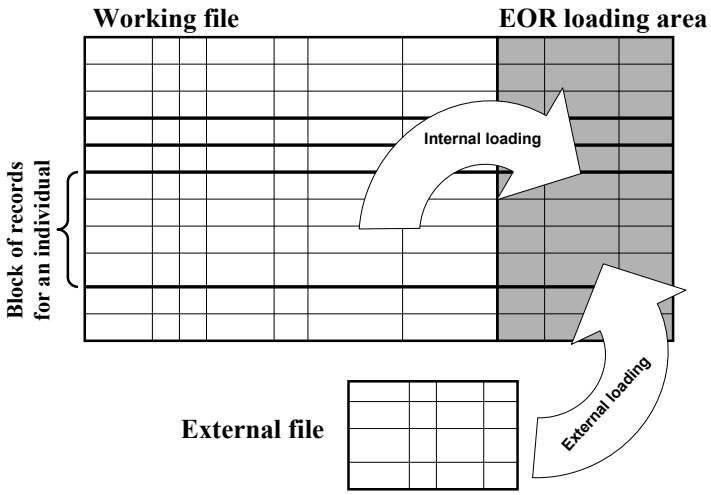
End of Record Loading

The EOR loading method refers to three types of procedures used to prepare and develop a linked file for analysis:

		Destination of information	
		All records	Selected records
Source of information	Within the file	***Generic internal loading***	***Selective internal loading***
	External file	***Generic external loading***	

1. ***Generic internal loading***: The working file receives data appended as one or more new variables at the end of each record, the values of which are derived directly or indirectly from other records in the same file pertaining to the same individual. This procedure may be used to establish record sequences for individuals or to place the date of the previous event on the next record for the same individual.
2. ***Selective internal loading***: If a specific type of event, which is usually the object of the research, defines one record per individual in the working file as the index record, it may be desired to selectively load additional information on the end of that record. One or more new variables are added, the values of which are derived from other records in the same file pertaining to the same individual. This procedure is used for a wide range of applications, including outcome events based on records that are 'remote' from (as distinct from 'adjacent' to) the index record, accumulation of person-time and incident events and comorbidity indices.
3. ***Generic external loading***: Data contained in an external file are merged onto the end of a working file, with the rootlpno as the matching key, and the external file as a look-up table. This procedure could be used to load cancer registration data or death registration data onto the end of every hospital morbidity data record pertaining to the same individual.

The following diagram illustrates the construction of an EOR loading area.



Some Important Computing Syntax Concepts

Regardless of the software used, syntax will need to be written to perform tasks such as:

- Searching across variables for specific codes.
- Searching down records belonging to the same individual.
- Identifying first and subsequent events.
- Aggregating totals of events for each individual.
- Identifying duplicate records and deleting appropriately.
- Identifying between hospital transfers for the same episode of care

Each statistical software package has a number of functions and commands that are especially relevant to these tasks. You should ensure that you are familiar with the following:

1. How to write, edit, save and run syntax using the ***syntax editor***. You should become familiar with use of the syntax window in 'batch mode' as an alternative to the use of pull-down menus and dialogue boxes. The demands of analysis of linked data exceed what is possible using the latter. An efficient strategy is to use pull-down menus and dialogue boxes for simple file manipulation and statistical procedures and the syntax editor for most other tasks.
2. The different ***date functions***, and especially the function used to calculate the number of days between two dates.
3. The use of ***lag functions***, which in SPSS returns the value of a variable taken from the record immediately before it in the file. Note that in SAS the lag function works on the last value of the variable held in the stack and does not always produce the same result as in SPSS if imbedded within a conditional statement (eg, if). In Stata and R the lag function can look forward as well as backward.
4. The use of ***assignments*** to set a variable to a particular constant or derived value.
5. ***Loops*** using 'repeat' commands and ***conditionals*** using 'if' commands.
6. The use of ***scratch variables*** in SPSS such as #counter, #flag, etc, which are memory variables separate from the data file and which hold their values during the processing of records, except where specifically modified by the user's syntax.
7. Some users may wish to familiarise themselves also with combined use of the AGGREGATE and MATCH FILES commands in SPSS, which although not demonstrated in this course, are used extensively by some linked file analysts.

The following table provides a comparison of some of the most important functions and commands in SPSS, SAS, Stata and R relevant to the analysis of linked health data:

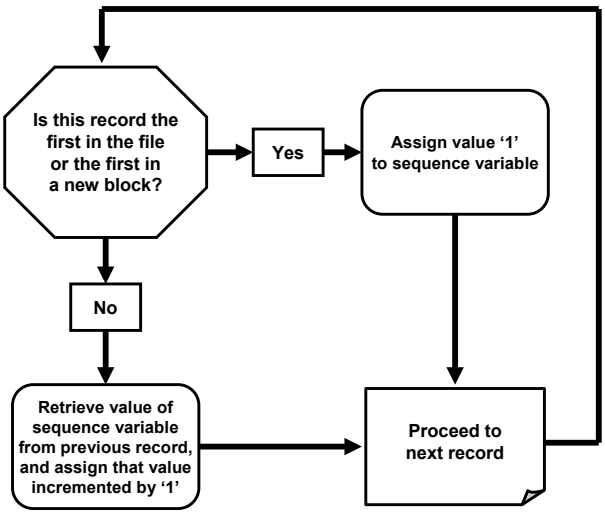
Package:	SPSS	SAS	Stata	R
		Functions		
Create a specified date	datevar1 = DATE.DMY(dd,mm,yyyy)	Format datevar1 = ddmmyy10.; datevar1 = dd.mm.yyyy;	generate datevar1=d(ddmmy) *[or]* =mdy(mm,dd,yy)	datevar1 <- as.Date("yyyy-mm-dd")
Number of days between two dates	XDATE.TDAY(datevar1) – XDATE.TDAY(datevar2)	Datevar1 – datevar2	datevar1-datevar2	difftime(datevar1,datevar2,units="days")
Lag function: returns variable's value from the previous record	LAG(var1)	lag(var1) *[on every pass]* var1 = var1 first.var1	var1[_n-1]	Lag(var1)
		Commands		
Assignments	COMPUTE var1 = y	Var1 = y	generate or egen	var1 <- y
Conditionals	IF *[single statement]* DO IF *[multiple statements]* and END IF	If *[single statement]* if …then do; *[multiple statements]* end;	If *[single statement]*	1) if(*single statement*) 2) *Multiple statements e.g.:* if(*statement 1*) else if (*statement 2*) 3) ifelse(*statement for "vector" operation*, *value if TRUE*, *value if FALSE*)
Loops	DO REPEAT var1 to varx and END REPEAT	array vars {no. of vars} var1 - varx; do i = 1 to {no. of vars}; … end;	foreach x in var1 to varx { }	for(i in *vector*){ }
		Miscellaneous		
Internal file sequence variable	$casenum	*_n_*	_n	*Not applicable for R. User assigns a row variable for each data frame object as required.*
Scratch variables: memory variables not a part of the data file	#var	Generate new variable and delete later	tempvar or generate new variable and delete later	*Generate a temporary named scalar variable then remove from workspace afterwards, e.g.,* flag <- numeric(1) rm(flag)

Each Line of Syntax is Not Like a Drop-down Menu Command

A common conceptual mistake made by newcomers to syntax is to think of each line of the syntax of a program as if it were the same as running equivalent tasks, one by one, using drop down menus. There are at least two reasons why running a program of syntax in batch mode is not the same as performing drop down menu tasks one by one. First, some syntax statements such as loops do not have equivalent drop down menu processes. Second, and very importantly, the lines of syntax are not necessarily independent one from the other in the same way as are drop down menu tasks. This is because the ***file processing order*** when running a batch of syntax is completely different from the file processing order involved in performing a series of drop down menu tasks. We will return to the topic of file processing order later in the course in order to understand it in more detail. Suffice to say here that you should be aware that there are important differences. In fact, running syntax just one line at a time can be similar to the effects of equivalent drop down menu tasks; but again this is not the same as running a whole program of syntax in batch mode.

Your First Syntax for a Linked Data File

In the exercise you will be asked to write syntax that performs a generic internal loading task. The aim is to load a new variable onto the file, which numbers the cancer records for each person in order starting with the number '1' for the first cancer record, '2' for the second record, and so on. A conceptual flow diagram of the required process is shown opposite. The corresponding syntax relies on the use of two conditional statements containing lag functions. The first statement identifies the first record in each ***block of records*** pertaining to a new individual and assigns the value '1' to the new ***sequence variable*** in the EOR loading area. The second statement identifies second and subsequent records in each block and assigns progressively higher values to the sequence variable, each incremented by one unit. This approach is adopted in the syntax below to achieve the desired result in SPSS and SAS. The first record in each block for a new individual has the unique characteristic (ie, unique within the block) that the value of rootlpno is different from that of the previous record. Of course, this presupposes that the file is sorted in the 'preferred sequence' (see above). The SPSS and SAS versions of the syntax illustrate how to make special allowance for the first record in the file given that the value returned from a lag function on the first record will be undefined, as there is no previous record.



The syntax shown below for Stata illustrates an alternative approach, where a default value of 1 is firstly assigned to the new sequence variable for every record in the file. This default value is then replaced by an incremented higher value for records that do not relate to the same individual as the preceding record. This alternative approach could have just as easily been implemented in SPSS, but not in SAS. The reasons are explained further on.

SPSS (lag function with two conditionals)

```
IF ($casenum EQ 1 OR rootlpno NE LAG(rootlpno)) morbseq =1.
IF ($casenum  GT 1 AND rootlpno EQ LAG(rootlpno)) morbseq = LAG(morbseq) + 1.
EXECUTE.
```

SAS (lag function with two conditionals)

```
data dlcourse.vascancer;
set dlcourse.vascancer;
if _n_ = 1 or rootlpno ~= lag(rootlpno) then morbseq = 1;
if _n_ > 1 and rootlpno = lag(rootlpno) then morbseq + 1;
run;
```

Stata (lag function with replacement of default value)

```
sort rootlpno, stable
generate morbseq=1
replace morbseq=morbseq[_n-1]+1 if rootlpno==rootlpno[_n-1]
```

R

```
morbseq.fn <- function(x){return(c(1:length(x)))}
vashmds <- ddply(vashmds, "rootlpno", transform,
                 morbseq = morbseq.fn(rootlpno))
```

Whilst a simple inspection of this small data file will reveal that no individual has more than two cancer records, you should imagine that it is a much larger file and that the maximum number of cancer records

per person is unknown. And here is a **useful tip**: in general, it is best to write syntax that allows for a full range of theoretical possibilities even when you believe that only a limited range of possibilities exists in the data file. Why? There are two reasons: (i) you could be wrong; and (ii) if you construct syntax that copes with all theoretical possibilities, it can be stored and used later on other data files regardless of their characteristics. The aim of the game is to write syntax for the most commonly used routines only once in your career.

By-Group Processing

SAS users and the LAG function: Although the lag function in SAS works in the first syntax given above, more generally, it is unsuitable for use in the same way as in SPSS and Stata as often it cannot be used reliably within a conditional statement. This is due to the different implementation of the lag function in SAS, in which it returns the previous value of the variable held in the processing stack, which is not necessarily the value of the variable in the previous record. However, SAS has alternatives. A useful feature for analysing linked data sets is ***by-group processing***. SAS recognises the first and last members of a group and creates automatic variables, which can be used in programming. These are first.group and last.group (eg, first.rootlpno and last.rootlpno). They require the file to be sorted by the group variable (eg, rootlpno) and that a program line 'by group' is inserted after the data step. The following is a by-group processing alternative in SAS for the creation of a morbidity sequence variable.

SAS (by-group processing)

```
data dlcourse.vascancer;
set dlcourse.vascancer;
by rootlpno;
if first.rootlpno then morbseq =0;
morbseq + 1;
run;
```

By-group processing is supported also by Stata. An advanced Stata and R user could make use of this facility to create a morbidity sequence variable as follows:

Stata (by-group processing)

```
sort rootlpno
by rootlpno: generate morbseq = _n
```

R (by-group processing)

```
morbseq.fn <- function(x){return(c(1:length(x)))}
ddply(vascancer, "rootlpno", transform, morbseq = morbseq.fn(rootlpno))
```

With R we insert, for each rootlpno (records previously sorted ascending by date), the morbidity sequence as calculated using our user-defined function, as an argument within the ddply function, to facilitate the by-group processing.

Aggregating

One approach that is often used by analysts is to aggregate selected and derived information from blocks of records pertaining to each individual into certain basic descriptive metrics. For example, it is possible to aggregate the number of admissions or the length of stay in hospital for each patient over many years to determine the minimum, maximum, mean, median, total etc of hospital stays and bed-days per person in the dataset across multiple records. This can then be further summarised at the population level. The syntax for aggregating data in a dataset is shown below for each software package, using the example of determining the total number of hospital records for each person in the data set.

SPSS (can also use pull down menus)

```
AGGREGATE
  /OUTFILE=* MODE=ADDVARIABLES
  /BREAK=rootlpno
  /totrec = LAST(morbseq)
```

SAS

```
proc means data = max;
        where first_hospitalisation = 1;
        var totrec;
run;
```

Stata

```
by rootlpno: egen totrec = max (morbseq)
```

R

```
vashmds <- ddply(vashmds, "rootlpno", transform,
            totrec = max(morbseq))
```

Tailored Exercise Instructions for SPSS, SAS, Stata and R Users

Professor Preen's native statistical software package is SPSS. Thus, the linguistics of SPSS and the functions that it supports colour the approaches taken to solving the syntax problems in this course. Nevertheless, the course is multilingual and all of the syntax problems can be solved in each of the three statistical packages.

General Implementation Notes for SAS Users

SAS data files are organised into libraries. A SAS data file has the structure 'library.datafile'. Libraries can be permanently set up using 'File – new – library' from the drop down menus or can be set up at the beginning of each session using syntax. Libraries correspond to Windows file directories. The available libraries can be viewed in the SAS explorer window. Files not saved into a particular library will go into the 'work' library which is temporary and only exists for the length of the SAS session.

SAS programming is usually either in the form of a data step or a procedure. Data steps have the following form:

```
DATA "new datafile";
SET "old datafile";
Programming statements;
Run;
```

If the new datafile has the same name as the old datafile it will overwrite the old datafile. The datafiles can either be saved into a permanent library (in which case the dataset name has the form 'library.datafile') or a temporary library (named in format 'datafile'). Temporary files are useful while testing your programs and if you name them incrementally it is easy to go back several steps if necessary.

SAS procedures have the form:

```
PROC SOMETHING data = "datafile name";
Procedure specific program (consult manual for this);
Run;
```

If you do not specify the data set name, the procedures will run on whatever data set is in the memory, which will be the last one which was used. There are three main SAS windows: program, log and output. If a window is accidentally shut down, a new window can be opened by typing 'log', 'output' or 'program' in the command line. To view a data file, double click the appropriate icon in the explorer window. Programs cannot be run on the data file while it is open.

SAS users may find useful an article by Dr Ronald Cody from New Jersey entitled, *Longitudinal Data Techniques: Looking Across Observations* available at:
http://www.ats.ucla.edu/Stat/sas/library/nesug00/ad1002.pdf.

General Implementation Notes for Stata Users

In general, Stata is more similar to SPSS than is SAS. If you find that you are running out of memory while running syntax, increase it using the command: set memory .

Sorting files in Stata, eg, sort rootlpno, will not necessarily keep records with equal sort variable values in the same order as those records were prior to the sort. To ensure that an original file order is preserved within groups of records with equal sort variable values requires a ***stable sort***; eg, sort rootlpno, stable.

Stata users should be aware that the creators of Stata have shown a propensity historically to create new versions of the package that are not compatible with syntax written for older versions. Thus depending on which version of Stata that you are using, it may be that some of the Stata solutions provided as examples with this course may cause grammatical syntax errors. Examples have included the grammar for merge procedures in Stata and the use of continuation signs of the form, \\\. The latter grammatical feature of Stata has even been withdrawn and then reintroduced between different versions of Stata.

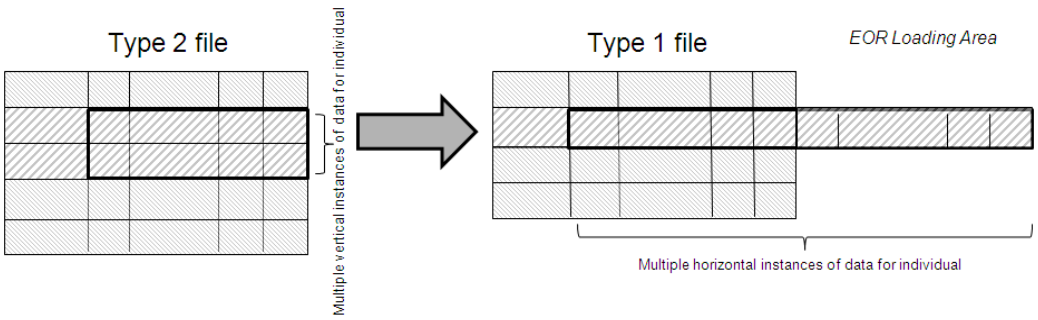
The use of temporary (scratch) variables is complicated in Stata, requiring local macros. Thus it is easier to create temporary variables as normal variables in the file and drop them from the data set afterwards.

General Implementation Notes for R Users

Copy the R data object (appropriately named file with the ".rda"file extension) into your chosen working directory (e.g., C:/ … /IALHD_course/Lab1/vashmds.rda). Create a new blank R script using R studio and save it to this working directory (e.g., C:/ … /IALHD_course/Lab1/IALHD_Lab1.R). Close R studio, then navigate to this working directory using Windows Explorer (or Finder if using a Mac) and open the R script by double-clicking the .R file – if installed using the default options, R studio should open the script in that working directory. An alternative way to specify or change the working directory once in R studio is to use the setwd() function. See Syntax solutions manual for further details and tips. Also see general notes about R, setting up R and R studio, and for downloading and importing the required R libraries of additional functions, as provided in the appendix.

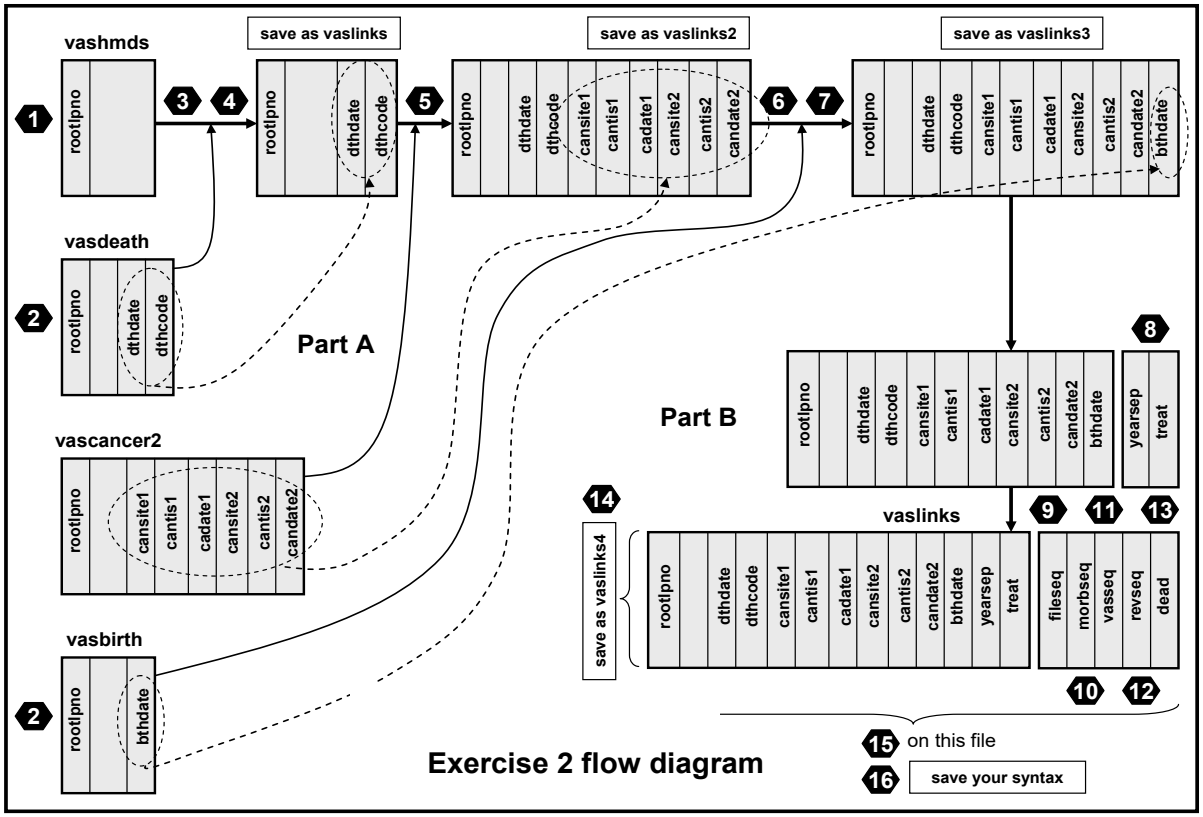
Converting a Type 2 File into a Type 1 File

In practical terms, a type 2 file can be distinguished from a type 3 file, because in the former it is practical to convert the data into a type 1 file. This is done by reshaping the file so that multiple instances of data for an individual, originally found in two or more vertically adjacent records, are reconfigured to occur horizontally on a single record for each individual. Note that to convert a type 3 file into a type 1 file is impracticable and this is a distinguishing feature of a type 3 file.



Exercise Flow Diagrams

At the end of each set of exercise instructions you will find a flow diagram, showing a step-by-step outline of the files that you will be using and how they are to be developed. We will say more in training session two about schematic and conceptual tools that can be used to assist you in performing your syntax programming and file management activities.



Overview of Exercise 1

<u>PART A</u>: Creating a morbseq variable to explore inpatient data

Step 1. Open vashmds file.

Step 2. Determine record number in data set.

Step 3. Create yearsep variable and determine number of separations by year.

Step 4. Create a morbseq variable.

Step 5. Use mobseq to determine number of patients in dataset and number of patients with single or multiple hospital records.

Step 6. Determine mean patient age at first admission.

Step 7. Use aggregate command (or SAS and Stata equivalent) to determine the mean, median and maximum number of hospital records per patient.

PART B: Converting a type II to a type I file

Step 8. Open vascancer file.

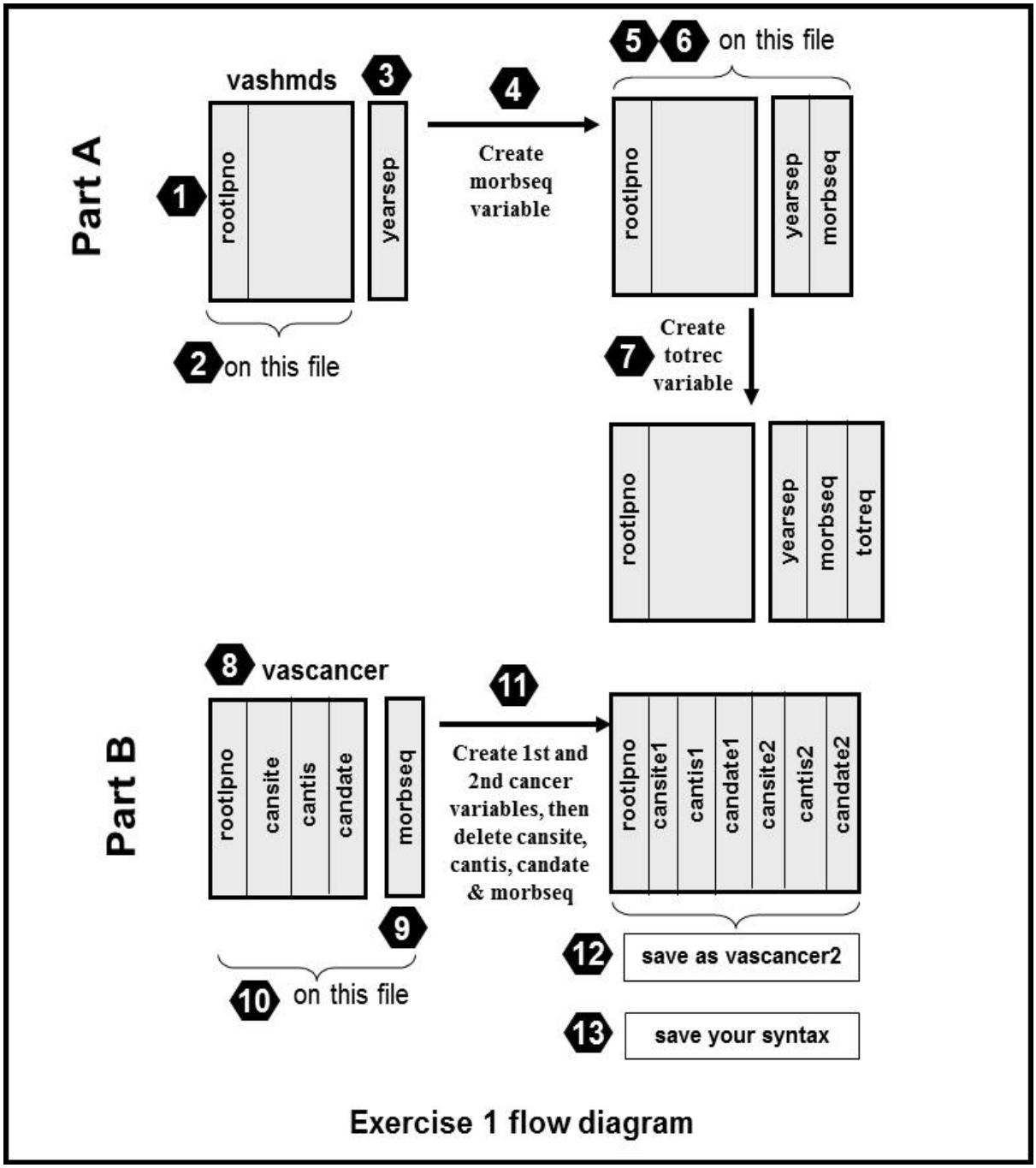
Step 9. Create a morbseq variable.

Step 10. Assess distribution of morbseq.

Step 11. Reconstruct the file as one record per individual.

Step 12. Save the reconstructed file as vascancer2.

Step 13. Save your syntax.



Exercise 1 flow diagram

Exercise Instructions

IMPORTANT: Please note that the methods you will use in this exercise are not necessarily the approach that you will take once you are more experienced and have more skills with writing syntax.

PART A: Creating a morbseq variable to explore inpatient data

In Part A of this exercise you will use syntax to explore the hospital morbidity data set. You will also calculate some basic descriptive parameters for the study sample from this file. Your aim is to create syntax to calculate basic summary statistics to complete the following table.

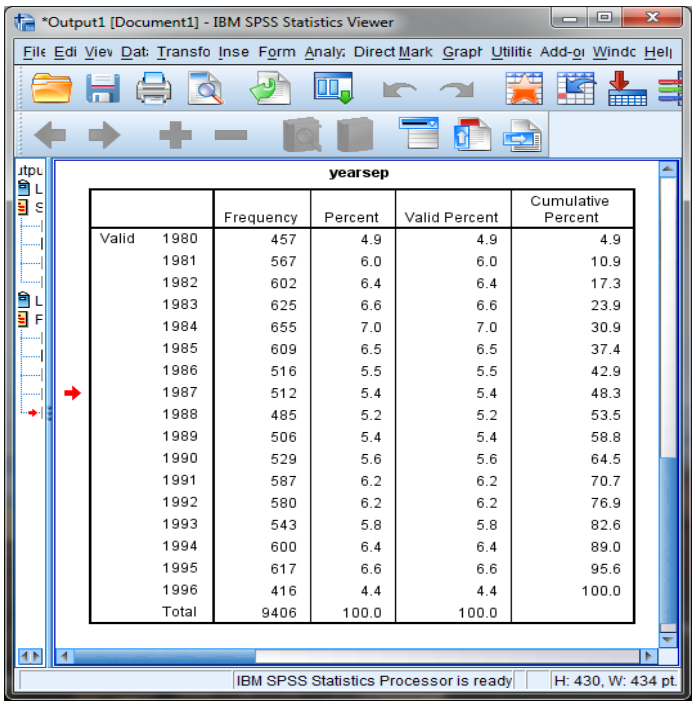
Characteristic	Result
Number of records in data set	
Year range of records in data set	
Number of people in data set	
Number of people with only a single hospital record	
Number of people with multiple hospital records	
Mean (±SD) age at first-time admission	
Age range at first-time admission	
Mean number of admissions per patient	
Median number of admissions per patient	
Maximum number of admissions per patient	

1. Open the vashmds data file in your preferred statistical software package.
2. Compose and execute syntax that will return the total number records in the vashmds file. Check that you have 9406 records. Note that for small files this can be achieved easily by scrolling to (or searching for) the end record in the data set. However, when working with large data sets containing thousands or millions of records this is not always so easily achieved.

   Useful tip: The use of syntax to return total case number in a data set is a quick way to check you are working on file with the correct number of records, especially after you have cut files down in size with previous analyses.

3. Now compose and execute syntax that will create a new variable called yearsep, derived from sepdate using a date function to extract the year. Run a frequency on yearsep to look at the number of hospital separations per year across the study period. The results should look like the output in the adjacent tables. Do the results look like what you would expect? What decisions might be necessary before proceeding with any further analysis?

SPSS



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1980	457	4.9	4.9	4.9
	1981	567	6.0	6.0	10.9
	1982	602	6.4	6.4	17.3
	1983	625	6.6	6.6	23.9
	1984	655	7.0	7.0	30.9
	1985	609	6.5	6.5	37.4
	1986	516	5.5	5.5	42.9
	1987	512	5.4	5.4	48.3
	1988	485	5.2	5.2	53.5
	1989	506	5.4	5.4	58.8
	1990	529	5.6	5.6	64.5
	1991	587	6.2	6.2	70.7
	1992	580	6.2	6.2	76.9
	1993	543	5.8	5.8	82.6
	1994	600	6.4	6.4	89.0
	1995	617	6.6	6.6	95.6
	1996	416	4.4	4.4	100.0
	Total	9406	100.0	100.0	

SAS

Results Viewer - sashtml

The SAS System

The FREQ Procedure

yearsep	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1980	457	4.86	457	4.86
1981	567	6.03	1024	10.89
1982	602	6.40	1626	17.29
1983	625	6.64	2251	23.93
1984	655	6.96	2906	30.90
1985	609	6.47	3515	37.37
1986	516	5.49	4031	42.86
1987	512	5.44	4543	48.30
1988	485	5.16	5028	53.46
1989	506	5.38	5534	58.83
1990	529	5.62	6063	64.46
1991	587	6.24	6650	70.70
1992	580	6.17	7230	76.87
1993	543	5.77	7773	82.64
1994	600	6.38	8373	89.02
1995	617	6.56	8990	95.58
1996	416	4.42	9406	100.00

Stata

Stata/IC 13.1 - H:\Teaching\PUBH - Linked data courses\M...

File Edit Data Graphics Statistics User Window Help

Variables

Year of Separation	Freq.	Percent	Cum.
1980	457	4.86	4.86
1981	567	6.03	10.89
1982	602	6.40	17.29
1983	625	6.64	23.93
1984	655	6.96	30.90
1985	609	6.47	37.37
1986	516	5.49	42.86
1987	512	5.44	48.30
1988	485	5.16	53.46
1989	506	5.38	58.83
1990	529	5.62	64.46
1991	587	6.24	70.70
1992	580	6.17	76.87
1993	543	5.77	82.64
1994	600	6.38	89.02
1995	617	6.56	95.58
1996	416	4.42	100.00
Total	9,406	100.00	

Command

H:\Teaching\PUBH - Linked data courses\Materials - Introductory Course\IALHI

4. Now that you have explored the number of records in the data set, you should investigate some basic descriptive characteristics for the number of patients. Compose and execute syntax that will create a new variable called morbseq (for *morbidity sequence*) in the EOR loading area, which is assigned the value '1' for the first hospital record for an individual, '2' for the second record in the same individual, and so on. **SAS users**: may employ by-group processing. Visually inspect the file to ensure that your syntax appears to have produced the expected result.
Useful tip: always visually inspect the data after executing syntax to check that the observed and expected results are the same. For complex syntax, it is wise to check the results for several different individuals with variations in characteristics that required the syntax to operate in different ways.

5. Using your newly created morbseq variable, select cases where morbseq=1 and run a frequency to output the number of patients in the dataset. Check that you have 2933 patients with at least one record in the data set. Now use the morbseq variable to determine the number of patients with multiple hospital separations in the data set? There should be 1772 with multiple hospital separation records during the observation period, leaving 1161 patients with a single record.

6. Now reselect cases where morbseq=1 and run a simple descriptive analysis to obtain the mean, standard deviation and range of ages in men at the time of their first hospital separation in the data set. Your results should be as shown in the output tables below.

SPSS

*Output1 [Document1] - IBM SPSS Statistics Viewer

File Edi View Dat Transfo Inse Form Analy Direct Marke Graph Utilitie Add-or Windo Help

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age in Years	2933	20	83	32.72	6.866
Valid N (listwise)	2933				

IBM SPSS Statistics Processor is ready

SAS

Results Viewer - sashtml

The SAS System

The MEANS Procedure

first_hospitalisation=.

Analysis Variable : AGE Age in Years				
N	Mean	Std Dev	Minimum	Maximum
6473	37.0812606	9.8294107	20.0000000	91.0000000

first_hospitalisation=1

Analysis Variable : AGE Age in Years				
N	Mean	Std Dev	Minimum	Maximum
2933	32.7207637	6.8662009	20.0000000	83.0000000

Stata

Stata/IC 13.1 - H:\Teaching\PUBH - Linked data courses\Materials - Introductory Course\IAL...

File Edit Data Graphics Statistics User Window Help

Variable	Obs	Mean	Std. Dev.	Min	Max
age	2933	32.72076	6.866201	20	83

Command

H:\Teaching\PUBH - Linked data courses\Materials - Introductory Course\IALHD Stata data files CAP NUM

7. Now compose and execute syntax to derive the mean, median and maximum number of admissions per patient in the data set. There are multiple ways you can do this which will be explored over the course, however you will implement one specific simple approach for this exercise. **SPSS users**: the simplest approach is to use the aggregate command. **Stata users**: Use the egen command. **SAS users**: You will need to create syntax using the retain statement.

Use the relevant software package command(s) to generate a new variable called totrec that returns the highest morbseq number for each patient to each record of that particular patient. Now select where morbseq=1 and calculate the mean, median and maximum number of separations per patient. If you have done this correctly, your results should look like the output tables below.

SPSS

*Output1 [Document1] - IB...

F E V D Tra In Fo An Direc Gr Ut Ad Wi H

Morbidity Sequence

N	Valid	2933
	Missing	0
Mean		3.21
Median		2.00
Std. Deviation		15.772
Minimum		1
Maximum		839
Sum		9406

IBM SPSS Statistics Processor is ready

SAS

Results Viewer - sashtml

Basic Statistical Measures			
Location		Variability	
Mean	3.206955	Std Deviation	15.77160
Median	2.000000	Variance	248.74331
Mode	1.000000	Range	838.00000
		Interquartile Range	3.00000

Stata

Stata/IC 13.1 - H:\Teaching\PUBH - Linked data courses\Materials - Introductory Course\IALHD Stata data files\vashmds.dta - [...

File Edit Data Graphics Statistics User Window Help

variable	mean	p50	sd	N	sum	min	max	sum
totrec	3.206955	2	15.7716	2933	9406	1	839	9406

Command

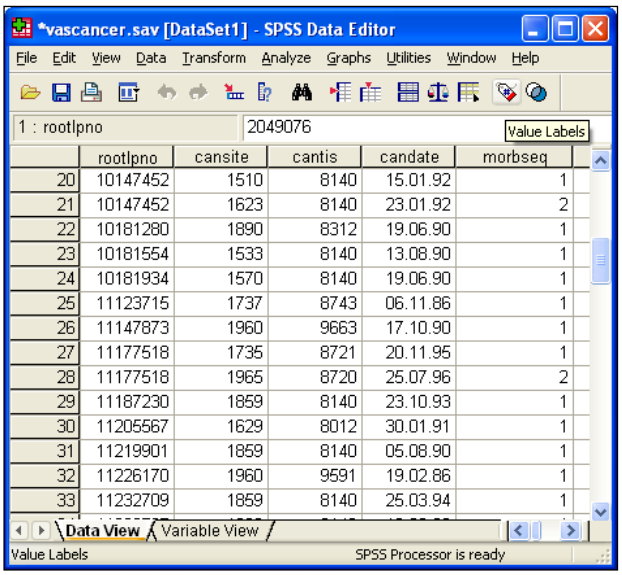
H:\Teaching\PUBH - Linked data courses\Materials - Introductory Course\IALHD Stata data files CAP NUM OVR

PART B: Converting a type II to a type I file

8. Open the correct vascancer data file for your preferred statistical software package.

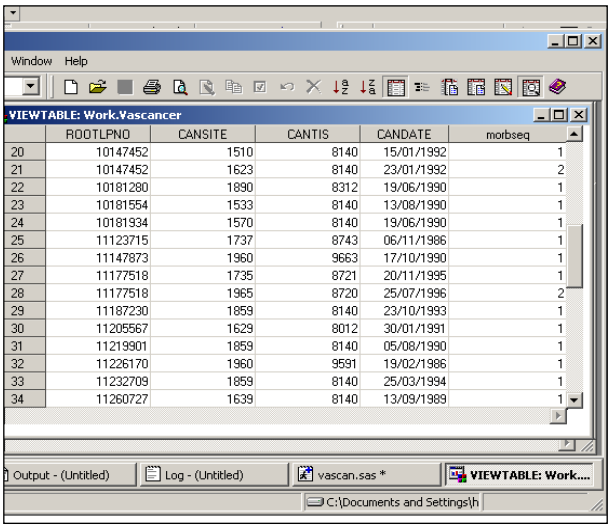
9. Compose and execute syntax that will create a new variable called morbseq (for *morbidity sequence*) in the EOR loading area, which is assigned the value '1' for the first cancer for an individual, '2' for the second cancer in the same individual, and so on. **SAS users**: may employ by-group processing. Visually inspect the file (especially rootlpno 11177518) to ensure that your syntax appears to have produced the expected result. **Useful tip**: always visually inspect the data after executing syntax to check that the observed and expected results are the same. For complex syntax, it is wise to check the results for several different individuals with variations in characteristics that required the syntax to operate in different ways.

SPSS



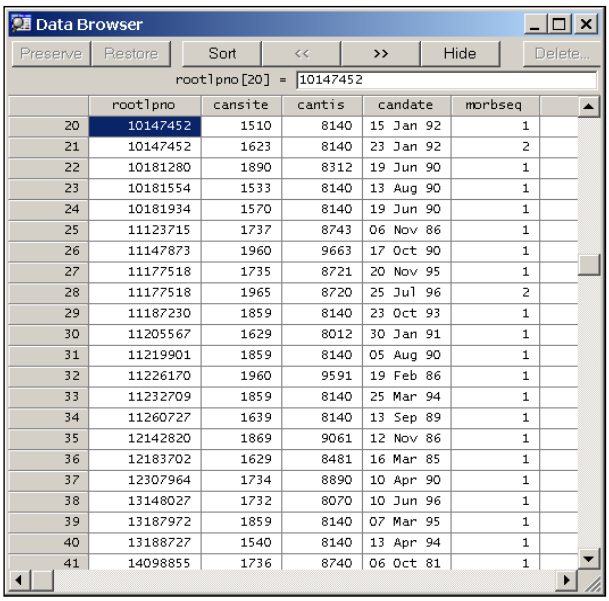
	rootlpno	cansite	cantis	candate	morbseq
20	10147452	1510	8140	15.01.92	1
21	10147452	1623	8140	23.01.92	2
22	10181280	1890	8312	19.06.90	1
23	10181554	1533	8140	13.08.90	1
24	10181934	1570	8140	19.06.90	1
25	11123715	1737	8743	06.11.86	1
26	11147873	1960	9663	17.10.90	1
27	11177518	1735	8721	20.11.95	1
28	11177518	1965	8720	25.07.96	2
29	11187230	1859	8140	23.10.93	1
30	11205567	1629	8012	30.01.91	1
31	11219901	1859	8140	05.08.90	1
32	11226170	1960	9591	19.02.86	1
33	11232709	1859	8140	25.03.94	1

SAS ↓



	ROOTLPNO	CANSITE	CANTIS	CANDATE	morbseq
20	10147452	1510	8140	15/01/1992	1
21	10147452	1623	8140	23/01/1992	2
22	10181280	1890	8312	19/06/1990	1
23	10181554	1533	8140	13/08/1990	1
24	10181934	1570	8140	19/06/1990	1
25	11123715	1737	8743	06/11/1986	1
26	11147873	1960	9663	17/10/1990	1
27	11177518	1735	8721	20/11/1995	1
28	11177518	1965	8720	25/07/1996	2
29	11187230	1859	8140	23/10/1993	1
30	11205567	1629	8012	30/01/1991	1
31	11219901	1859	8140	05/08/1990	1
32	11226170	1960	9591	19/02/1986	1
33	11232709	1859	8140	25/03/1994	1
34	11260727	1639	8140	13/09/1989	1

Stata ➔

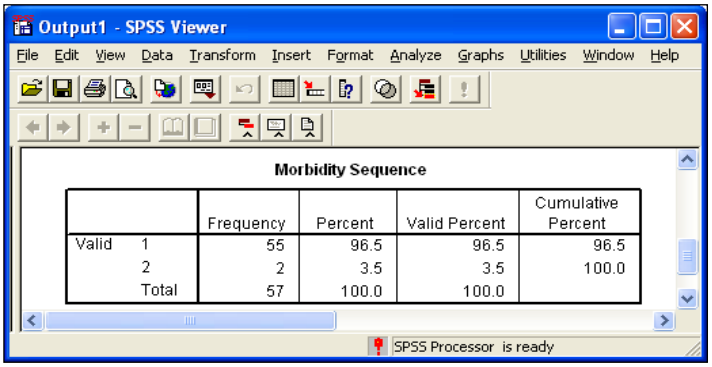


	rootlpno	cansite	cantis	candate	morbseq
20	10147452	1510	8140	15 Jan 92	1
21	10147452	1623	8140	23 Jan 92	2
22	10181280	1890	8312	19 Jun 90	1
23	10181554	1533	8140	13 Aug 90	1
24	10181934	1570	8140	19 Jun 90	1
25	11123715	1737	8743	06 Nov 86	1
26	11147873	1960	9663	17 Oct 90	1
27	11177518	1735	8721	20 Nov 95	1
28	11177518	1965	8720	25 Jul 96	2
29	11187230	1859	8140	23 Oct 93	1
30	11205567	1629	8012	30 Jan 91	1
31	11219901	1859	8140	05 Aug 90	1
32	11226170	1960	9591	19 Feb 86	1
33	11232709	1859	8140	25 Mar 94	1
34	11260727	1639	8140	13 Sep 89	1
35	12142820	1869	9061	12 Nov 86	1
36	12183702	1629	8481	16 Mar 85	1
37	12307964	1734	8890	10 Apr 90	1
38	13148027	1732	8070	10 Jun 96	1
39	13187972	1859	8140	07 Mar 95	1
40	13188727	1540	8140	13 Apr 94	1
41	14098855	1736	8740	06 Oct 81	1

10. Request a frequency distribution on morbseq. The results should show that there are 55 first-time and 2 second-time cancers in the file.

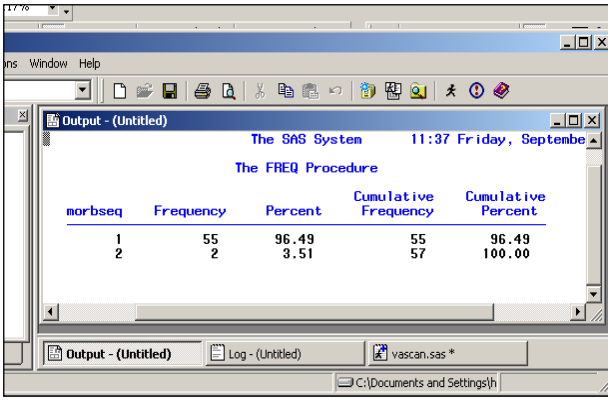
11. Now reconstruct this file such that each individual only has one record containing <u>all</u> of their available cancer data.

SPSS

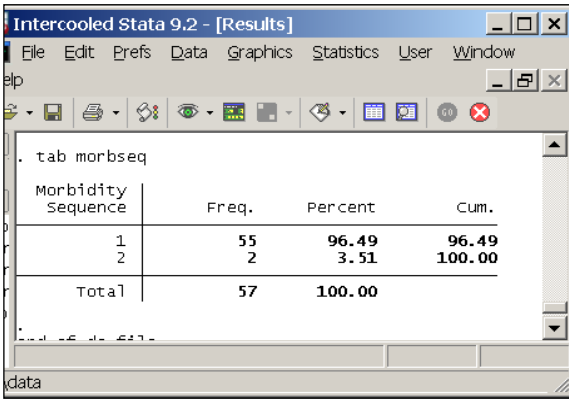


		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	55	96.5	96.5	96.5
	2	2	3.5	3.5	100.0
	Total	57	100.0	100.0	

SAS



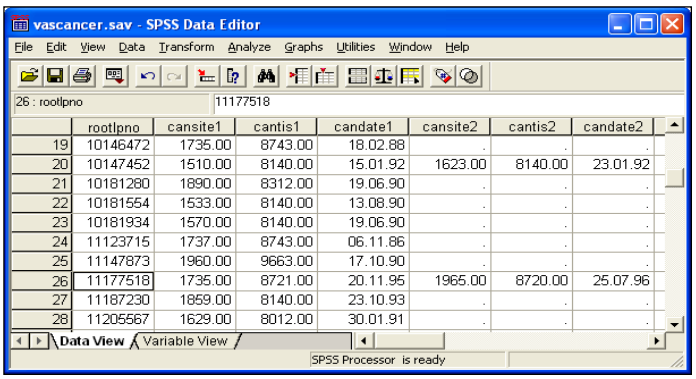
Stata



Important note for everyone: Please follow the instructions below carefully and do not waste your time by trying to work out a better way to perform the task. The exercise seeks to demonstrate a principle by asking you to take a bad approach to the task – a method that you will never use again once you have acquired a few more skills over the days to come.

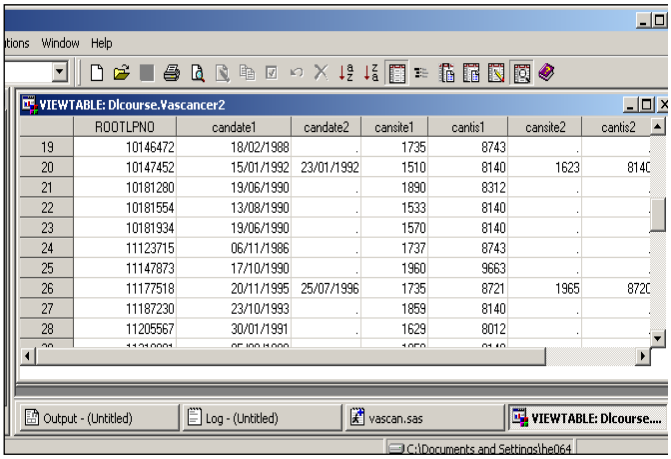
Compose and execute syntax that will create six new variables in the EOR loading area: cansite1, cantis1, candate1, cansite2, cantis2 and candate2. For records of first-time cancers, the syntax should assign the values of cansite, cantis and candate to cansite1, cantis1 and candate1 on the <u>same</u> record; while leaving cansite2, cantis2 and candate2 blank. For records of second-time cancers, the syntax should assign the values of cansite, cantis and candate <u>found on the previous record</u> to cansite1, cantis1 and candate1; while assigning the values of cansite, cantis and candate found on the <u>same</u> record to cansite2, cantis2 and candate2.

SPSS



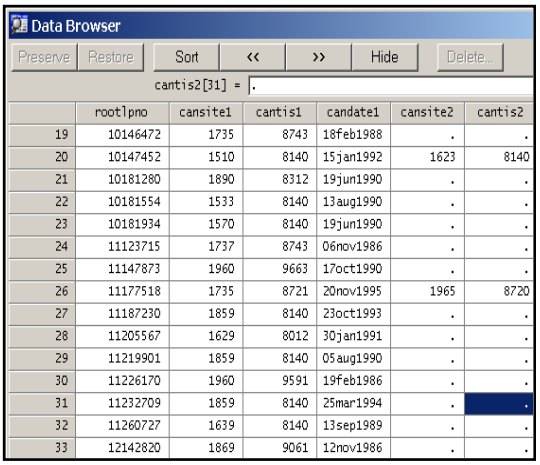
	rootlpno	cansite1	cantis1	candate1	cansite2	cantis2	candate2
19	10146472	1735.00	8743.00	18.02.88	.	.	.
20	10147452	1510.00	8140.00	15.01.92	1623.00	8140.00	23.01.92
21	10181280	1890.00	8312.00	19.06.90	.	.	.
22	10181554	1533.00	8140.00	13.08.90	.	.	.
23	10181934	1570.00	8140.00	19.06.90	.	.	.
24	11123715	1737.00	8743.00	06.11.86	.	.	.
25	11147873	1960.00	9663.00	17.10.90	.	.	.
26	11177518	1735.00	8721.00	20.11.95	1965.00	8720.00	25.07.96
27	11187230	1859.00	8140.00	23.10.93	.	.	.
28	11205567	1629.00	8012.00	30.01.91	.	.	.

SAS



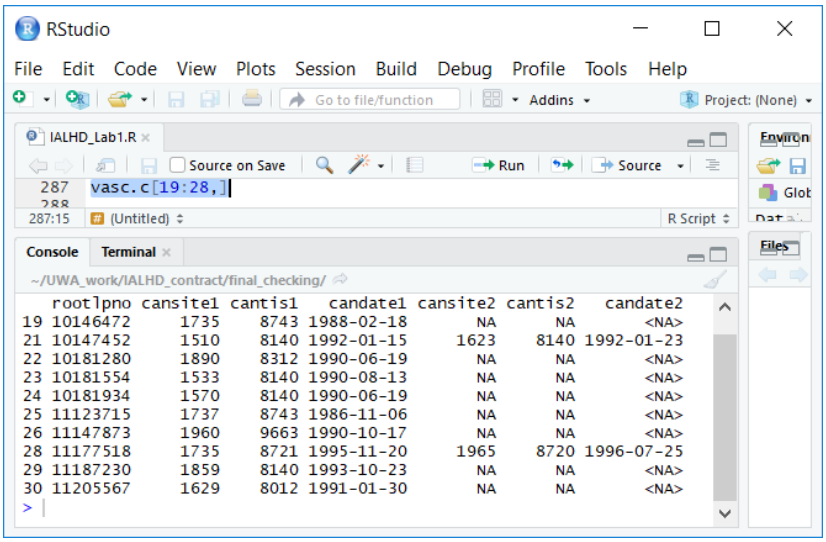
	ROOTLPNO	candate1	candate2	cansite1	cantis1	cansite2	cantis2
19	10146472	18/02/1988	.	1735	8743	.	.
20	10147452	15/01/1992	23/01/1992	1510	8140	1623	8140
21	10181280	19/06/1990	.	1890	8312	.	.
22	10181554	13/08/1990	.	1533	8140	.	.
23	10181934	19/06/1990	.	1570	8140	.	.
24	11123715	06/11/1986	.	1737	8743	.	.
25	11147873	17/10/1990	.	1960	9663	.	.
26	11177518	20/11/1995	25/07/1996	1735	8721	1965	8720
27	11187230	23/10/1993	.	1859	8140	.	.
28	11205567	30/01/1991	.	1629	8012	.	.

Stata



	rootlpno	cansite1	cantis1	candate1	cansite2	cantis2
19	10146472	1735	8743	18feb1988	.	.
20	10147452	1510	8140	15jan1992	1623	8140
21	10181280	1890	8312	19jun1990	.	.
22	10181554	1533	8140	13aug1990	.	.
23	10181934	1570	8140	19jun1990	.	.
24	11123715	1737	8743	06nov1986	.	.
25	11147873	1960	9663	17oct1990	.	.
26	11177518	1735	8721	20nov1995	1965	8720
27	11187230	1859	8140	23oct1993	.	.
28	11205567	1629	8012	30jan1991	.	.
29	11219901	1859	8140	05aug1990	.	.
30	11226170	1960	9591	19feb1986	.	.
31	11232709	1859	8140	25mar1994	.	.
32	11260727	1639	8140	13sep1989	.	.
33	12142820	1869	9061	12nov1986	.	.

R Users: the rows with names 20 and 27 are now missing, since they have been deleted, as per instructions (this is different to what is displayed in the other software packages).



For the sake of aesthetics, you may delete the original cansite, cantis and candate variables and the morbseq variable from the file. You should also manually delete the original first-time cancer records for rootlpno 10147452 and rootlpno 11177518. The appearance of the data file should now be like that shown above with the two sets of cancer information.

As already emphasised, the methods you have just used in step 4 are poor practice and not the approach that you will ever use again once you have more skills. It should never be necessary to delete records manually in a problem of this type. The reason you are 'forced' to do it here is because for the individuals with two cancer records, the record receiving the aggregated information on both cancers was the second record and not the first record. This leaves us in the situation where we have no syntactical basis on which to define which first records are followed by a second record for the same individual and may thus be deleted. You had to look at the file visually to work this out and then delete manually. If the file was much bigger and there was the possibility of 1-5 cancer records per patient, this visual and manual approach would become quickly impracticable.

It is possible to write more generic syntax that copes with a file of any size and a larger, multiple number of cancers and to perform the task in a way that loads all of the information onto the EOR area of the first-time cancer record, thus enabling all first-time records to be easily selected using morbseq = 1 for saving as a new file, rather than manually deleting the unwanted records. Method for achieving this result involve either the use of an 'upside-down file' or a 'vector file' and you will learn about these later in the course.

12. Save the reconstructed data file using the file name vascancer2. You will need it for later work.
13. Finally, save your syntax written for exercise 1, as some of it will be used again. **Useful tip**: add documentation by way of non-executable comments or remarks to your syntax. Most computing languages allow this. Documentation on syntax will prove invaluable when you may wish to use it again several years later or if you wish to share your syntax with a colleague.

Reference

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

Cody R. **Longitudinal data techniques: looking across observations.** Available at: http://www.ats.ucla.edu/STAT/sas/library/nesug00/ad1002.pdf.

REVIEW 1:
'INTRODUCTION TO DATA LINKAGE SYSTEMS'

Please check that you are familiar with the following topics:

- The Acheson and Weed traditions of health data integration ☐
- The use of data linkage systems as research infrastructure ☐
- The ways that data linkage systems may be developed ☐
- The potential scope of a data linkage system ☐
- The benefits of a data linkage system ☐
- International variations in data linkage systems ☐
- Desirable design criteria of comprehensive data linkage systems ☐

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

- Rectangular file structure ☐
- End of record loading ☐
- Preferred sequence of records ☐
- Block of records for an individual ☐
- Index event ☐
- Syntax editor ☐
- Date functions ☐
- Lag functions ☐
- Assignments ☐
- Loops ☐
- Conditionals ☐
- Scratch variable (SPSS only) ☐
- Sequence variable ☐

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

- Open and check data files in SPSS, SAS, Stata or R ☐
- Create a morbidity sequence variable ☐
- Reconstruct a single-record-per-person data file from a multiple-record-per-person data file (in the simple case of limited double records in a type 2 file) ☐

2. THEORY OF DATA LINKAGE METHODS AND APPLICATIONS

2.1 Definition of Data Linkage

In this course we define data linkage as follows:

> *Data linkage is the bringing together from two or more different sources, data that relate to the same individual, family, place or event.*

This definition has evolved somewhat from earlier versions. For example, Hobbs and McCall in 1970 offered the definition, *'the bringing together in a single file, of records derived from different sources, but relating to the same individual or event'*. Substitution of the word 'data' for 'record' embraces a broader concept of information sources that goes beyond conventional paper records and electronic records in the health and welfare sectors. The data sources may include, for example, spatially referenced geographic information systems, where data are not necessarily structured as distinct 'records', as well as images and sound. Even biospecimens may be viewed as containing data in a biological form that can be read by scientific instrumentation. The increased utility of family linkage is also taken into account, in view of recent developments in genetic epidemiology.

Harold Newcombe of the Atomic Energy Commission of Canada undertook pioneering work on medical record linkage in the 1950s, and Sir Don Acheson established the first medical record system in England in 1962, the Oxford Record Linkage System. Newcombe's *Handbook of Record Linkage Methods for Health and Statistical Studies, Administration, and Business* and Baldwin, Acheson and Graham's *Textbook of Medical Record Linkage* were two of the foundation textbooks on the subject.

2.2 Practical Steps in the Process of Data Linkage

The process of data linkage consists of five practical steps:

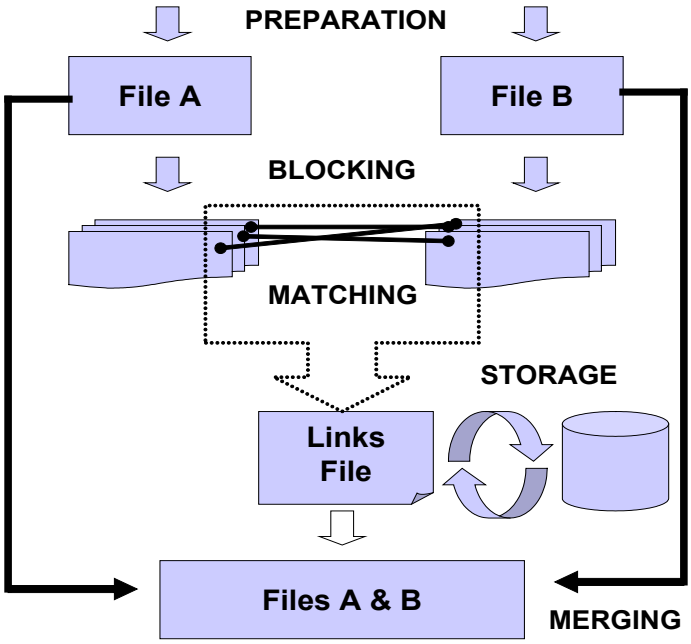
1. ***Preparation***: Prior to commencement of the more technical steps of data linkage comes the very important phase of *data preparation*. This involves searching for and correcting invalid values in the data fields; frequency listing and cross-checking; recoding and transforming data into a standard format; production of phonetic codes from names using name compression algorithms; and file explosion to take account of alternative surnames and given names.

2. ***Blocking***: Files of records are ordered or blocked in particular ways to increase the efficiency of searching for matches. For unique matching, blocking simply involves sorting the files into the same order of the unique identifier. For probabilistic matching, use is typically made of multiple blocking ('multiple passes'), such as by a name compression algorithm and by date of birth. The WA Data Linkage Project blocks on unit medical record number (unique only to teaching hospitals), surname, first given name and initial, date of birth, sex and address. The blocking strategies for each pass should be independent to the extent possible. The fields that are best suited for blocking are those with the most number of values possible and highest reliability.

   Please note that the above is a different use of the word 'block' from how we use the word in the practical sessions to describe blocks of records in files pertaining to the same individual.

3. ***Matching***: This is the process by which potentially linkable pairs of records are systematically compared against all other 'candidate' records (ie, those in the same block) to determine whether or not they relate to the same person. In a probabilistic matching process, each matching field provides some information and, taken together, the fields determine the status of the pair being examined. Some fields provide more reliable information than others, and thus in probabilistic matching a system of weights is used to reflect the value of the comparison in each field. The end result of the matching process is a file of accepted links between the data sources.

4. ***Storage***: This phase is especially distinctive of a *data linkage system*, as distinguished from *ad hoc data linkage* for a single research project. The file of links arising from completion of the matching phase is stored for later retrieval for use in future data extraction and merging.
5. ***Merging***: The linkable data sets may be brought together using the links arising directly from or stored after the completion of the matching phase. The data from different sources are assembled in such a way that they can be analysed as a set of composite records for each individual, family, place or event. This phase may include validation checks, error detection and correction of linkage errors.



Please note that in the above illustration, files A and B may represent two different files of data or they could represent the original and a copy of the same file with multiple records per person, which is to undergo an internal linkage of records pertaining to each unique individual.

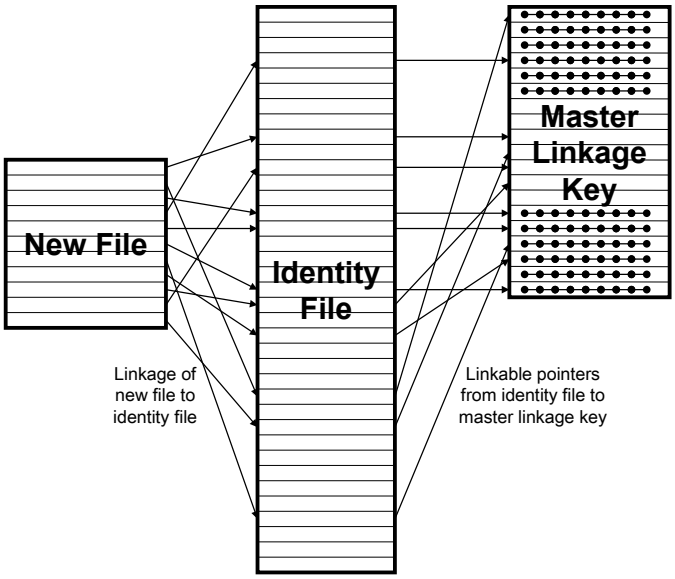
2.3 Data Preparation Techniques

Phonetic Compression and File Explosion

Methods of ***phonetic compression*** of names have been developed to reduce missed matches caused by variation in spelling and errors in recording of names. They also simplify the process of blocking. Typically they reduce names to a short fixed format code by eliminating vowels, regarding certain consonants as silent, taking others as equivalent and regarding a set number of significant consonants as the name. Examples are the Soundex and NYSIIS name compression algorithms; eg, the surnames Moroni and Moreno are both compressed to M650 in Soundex and to MARAN in NYSIIS.

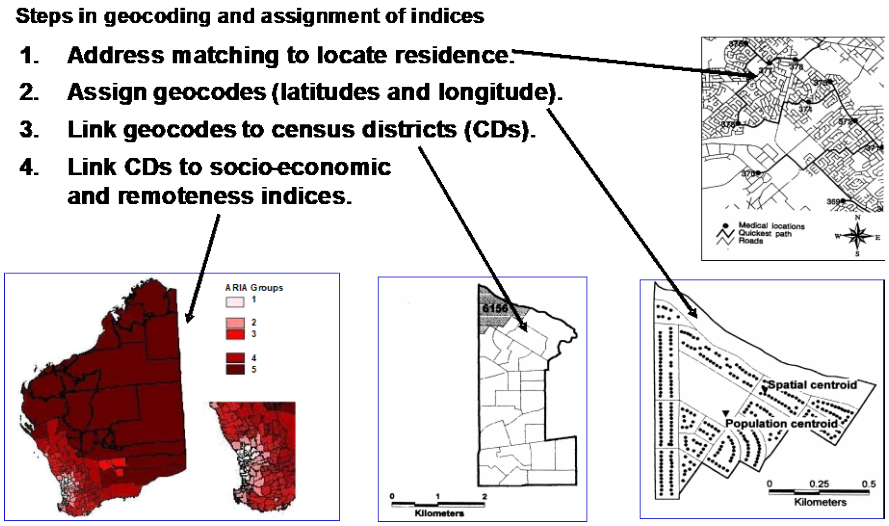
Recognised contractions and variants of forenames and surnames may be referred to an ***equivalencing dictionary***; eg, Liz, Lizzy, Bess, Betty, Beth, Bethwyn and El are variants of Elizabeth; Mac or Mc in surnames. These equivalent names and the alternative use of maiden surname for married women may be taken into account in the blocking phase by allowing records to appear in more than one block using the alternative names. This is known as ***file explosion***.

After file explosion, many individuals will have two or more identities (an ***identity*** is a set of partial identifiers for an individual) and thus the total number of identities covered by the data linkage system will exceed the number of chains of links in the master linkage key. It therefore becomes necessary to store the identities in a separate ***identity file*** with linkable pointers from that file to the master linkage key. Any new file of partial identifiers needing to be brought into the system is first matched and linked to the identity file and thereby ***mapped*** to the master linkage key so that new links can be added to the correct chain of links for each individual represented in the new file. More explanation will come later concerning the concepts of a chain of links and a master linkage key.



Geocoding

Latitude, longitude and census areas may be pre-assigned to individual records within data collections using ***address parsing software*** to extract ***tokens*** (eg, street number and name, suburb and postcode) that are referred to spatially referenced look-up tables (eg, those provided by the WA Land Information System). Although not essential to the process of data linkage, these geocoded references are valuable in research using linked health data. They can be used to associate health events with environmental attributes and to provide much improved accuracy in the assignment of social disadvantage (eg, Australian SEIFA indices) and remoteness scores (eg Australian ARIA or ARIA-plus indices). They enable the geographic distribution of individuals to be defined at a given point in time and, combined with linked health data, are capable of supporting a cross-sectional view of socio-demographic determinants of health outcomes, and a longitudinal view of residential histories, for example, in patients with chronic conditions to study illness-related migration.



2.4 Data Matching

Identifiers

Identifiers are data fields that have predictive value in assigning the correct identity to an individual, family, place or event. Some are intended to be ***unique identifiers*** in that each individual, family place or event is assigned a unique value, such as a unit medical record number or health insurance registration number. Others, such as a person's name or date of birth, or a street address, are not intended to be completely unique, but are intended to have a positive predictive value in identification. These are known as ***partial identifiers***. In practice, even the most rigorous 'unique' identifiers (eg, tax file numbers; genomic profiling used in paternity investigations) are at best very high performance partial identifiers due to human or machine error in use or statistical considerations. On the other hand, it is possible to pretend that even a relatively weakly performing partial identifier can be used as if it were a unique identifier.

Methods of Data Matching

Linkage across data sources requires access to, and assessment of, unique or partial identifiers that are common to each set. In practice, this may involve comparing all available identifiers in one set against those in a master file or the comparison of similar identifiers in two large data files *de novo*. Either of two general techniques is employed in the matching phase of data linkage - the methods may be ***deterministic*** or ***probabilistic***, and each of these methods has subvariants:

Deterministic matching: When data are only linked when all selected identifiers from each data source match exactly, or match within defined limits, the method is termed 'deterministic', a term that belies the fact that the process may only identify 80-90% of true matches due to recording errors. The method may be particularly unreliable if recording errors or multiple registrations of individuals are common. It may also be used when identifiers are separately or jointly encrypted prior to matching to overcome privacy concerns. On the other hand, deterministic matching may be sufficiently reliable if special measures or technologies are employed to ensure accuracy; for example, through the use of smart cards or bar coded cards with photographic ID.

There are two variants of deterministic matching:

(i) ***Unique matching*** is used predominantly when the range of identifying information is restricted to a supposedly unique identifier such as an insurance or medical record number. The files are sorted into the same order, and matched within blocks containing either a matched pair or a lone record. This merging process may only identify 80-85% of the true matches due to recording errors unless special measures or technologies are employed. Performance of unique matching may also be improved by safeguards such as ***check digits*** to help prevent the misrecording of unique identifiers.

(ii) ***Fuzzy matching*** makes use of partial identifiers such as names, date of birth, sex, postcode or place of birth, often allowing for a margin of error by linking data that are almost the same. Computer programs either present a choice of matches to the user relies on a scoring system. Typically this technique identifies 85-90% of true matches. Another form of fuzzy matching involves the use of cyphers, taking specified characters from partial identifiers.

Probabilistic matching: This form of matching, as developed by Newcombe and others, allows identifiers to vary between sources. Candidates for linkage are assessed for similarity and then classified as probable or improbable links. The assessment of similarity is based on sophisticated decision rules that include weights derived from the probabilities of similarities occurring by chance. There are two variants:

(i) ***Automated computerised matching*** (also know as *statistical matching*) leaves the decisions on links to be made automatically by the computer. Decision rules are built into a software package based on the probability of two records being from different subjects having the same identifiers. These probabilities are aggregated into a score and checked against a user-defined threshold to determine whether a link should be made. Such techniques typically identify 95-99% of true matches, while 1-2% of matches are false positives.

(ii) ***Probabilistic matching with clerical review***: Clerical review is applied to a group of doubtful matches between acceptance and rejection user-defined thresholds assessed probabilistically. Often this clerical review process is minimal, but it can also be extensive, especially when an archive of previously linked records is available. Under such circumstances, the levels of false positive and false negative errors can be reduced to <0.1%.

2.5 Theory of Probabilistic Matching

Missed Matches and Mismatches

The following table shows the four possibilities that exist when the values of an identifier are compared between two records that may or may not relate to the same individual, family, place or event:

		Records are from the same subject	
		Yes	No
Identifier values agree	Yes	A = true positives = ***true matches***	B = false positives = ***mismatches***
	No	C = false negatives = ***missed matches***	D = true negatives = ***true non-matches***

Match Weights

Each identifier has two probabilities associated with it:

m = the probability that the values of an identifier agree given that the record pair is from the same subject [= A/(A+C) = sensitivity = 1 – prob(*missed match*)].

u = the probability that the values of an identifier agree given that the record pair is not from the same subject [= B/(B+D) = 1 – specificity = prob(*mismatch*)].

Note that because there are so many more true non-matched pairs possible than true matched pairs, u is effectively the probability that the identifier values agrees at random.

The ***match weight*** (or *score*), being the log-likelihood ratio for an identifier that agrees for a paired comparison, is computed as the logarithm to the base 2 of the ratio of *m* to *u*,

$$\mathbf{w_i = \log_2(m_i/u_i)}$$

whereas in the case where the identifier values disagree,

$$\mathbf{w_i = \log_2[(1-m_i)/(1-u_i)]}$$

<u>Example</u> *(from Jaro, 1995): Assume sex does not match in 10% of pairs or records where it should (ie, 90% sensitivity) and social security number does not match in 40% of such pairs (60% sensitivity), and that chance agreement has a probability of 0.5 for sex (assumes equal males and females; 50% specificity)) and 1 in 10 million for SSN (99.99999% specificity).*

For sex, m = 0.9 and u = 0.5, ∴w for agreement on sex = $\log_2(0.9/0.5) = \ln(0.9/0.5)/\ln(2) = 0.85$.
and w for disagreement on sex = $\log_2(0.1/0.5) = \ln(0.1/0.5)/\ln(2) = -2.32$.

For SSN, m = 0.6 and u = 0.0000001, ∴w for agreement on SSN = $\log_2(0.6/0.0000001) = 22.51$.
and w for disagreement on SSN = $\log_2(0.4/0.9999999) = -1.32$.

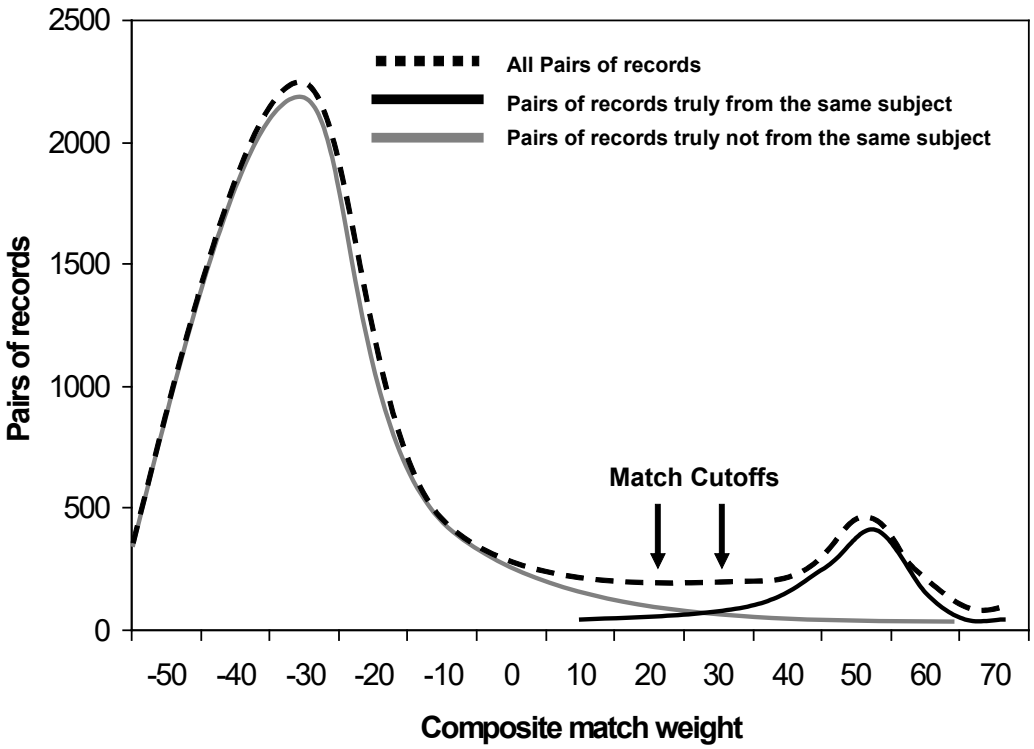
Note in this example how agreement on sex is not particularly discriminatory (low +ve match weight), whereas disagreement on sex is highly discriminatory (high –ve match weight). For SSN the opposite is the case; agreement on SSN is very informative, whereas disagreement on SSN is not. The discriminating power of different identifiers is reflected by the extent to which *m* exceeds *u* across the range of possible values.

Composite Match Weights

For each record pair compared, a ***composite match weight*** is computed as the sum of the individual weights for all field comparisons, ie, $\mathbf{W = \Sigma w_i}$. The higher this score, the more likely the record pair is to arise from the same subject. In some matching protocols, eg that used by the Oxford Record Linkage Study, the composite match weight is compartmentalised into a name score and a non-name score, being the separate sums of match weights for name and non-name identifiers.

Distribution of Composite Match Weights

The distribution of composite weights is generally bimodal. Many pairs of records have high negative weights, because most pairs are not from the same subject. There is another mode at highly positive weights where the pairs of records that are from the same subject tend to cluster. In fact, the observed bimodal distribution is the sum of two underlying distributions of composite match weights, one for the pairs of records that are truly from the same subjects and the other for pairs of records that are truly not from the same subjects.



Clerical review (manual checking) of the ***grey zone*** or *valley of indecision* in the bimodal distribution of composite match weights will increase the likelihood of classifying those pairs of records correctly as true matches or mismatches. The extent of clerical review is determined by setting the ***match cutoffs***, which are assigned using a combination of theoretical considerations (sometimes supported by the matching software) and trial and error. The region below the grey area is the region of unconditional rejection of matches and that above is the region of unconditional acceptance of matches. The WA Data Linkage Unit typically performs clerical checking on about 2% of all potential links, and one-quarter of these may require detailed checking.

Estimation of the u_i and the m_i

Reasonable estimates of u_i can be obtained by considering only the probability of chance agreement on the identifier. Data linkage techniques often take advantage of the distribution of the values of an identifier. Eg, agreement on a surname like *Padereski* is more likely to represent the same individual than agreement on *Smith*. Frequency tables of value states are used to alter the u_i probability. Eg, if there are 1,000 records and *Smith* appears 200 times, the initial estimate of u_i for surname given the value of *Smith* is 0.2. Initial estimation of a general m_i for a field is made as a judgement call, eg, 0.9. Distribution effects are often population-specific; eg, Tan in Singapore vs Andersen in Denmark.

In some matching software packages, the u_i and m_i are updated on a value-specific basis, using the number of times a value has agreed in a linked pair and the number of times it has participated in a linked pair, based on the assumption that the majority of matching decisions are correct. This process converges to obtain a stable set of parameter estimates.

2.6 Evaluating the Performance of Data Matching

Typically, evaluation of the performance of matching is undertaken either by:

i. ***Match weight sampling and review techniques***: Pairs of records receiving different composite match weights are sampled in the regions before, within and after the grey zone. The pairs of records are then subjected to a very detailed clerical review to determine, as accurately as possible, their true matching status.

   Example: The WA Data Linkage Unit has assessed the quality of hospital morbidity data linkage by a sampling technique. Pairs of records were selected from 21 different composite match weight strata, and the results of a careful clerical review were compared with matching results previously obtained using the routine processes of the Project. The proportions of mismatches (1 - specificity) and missed matches (1 – sensitivity) were both estimated by this method at 0.11%.

ii. ***Chain sampling and review techniques***: Chains of linked records assessed by the routine process to belong to the same individuals are sampled and subjected to a detailed clerical review to determine if any of the links in the chain relate to mismatched records.

   Example: An audit by the WA Data Linkage Unit, involving sampling and detailed clerical review of linked chains with up to 2000 links per chain, resulted in an estimate of <0.3% of chains containing one or more links to mismatched records.

iii. ***Use of validation data sets***: This may consist of conducting a matching run on records for which the true matching status is known in advance; or observing the level of internal matches obtained on files of supposedly unique records (eg, death registrations).

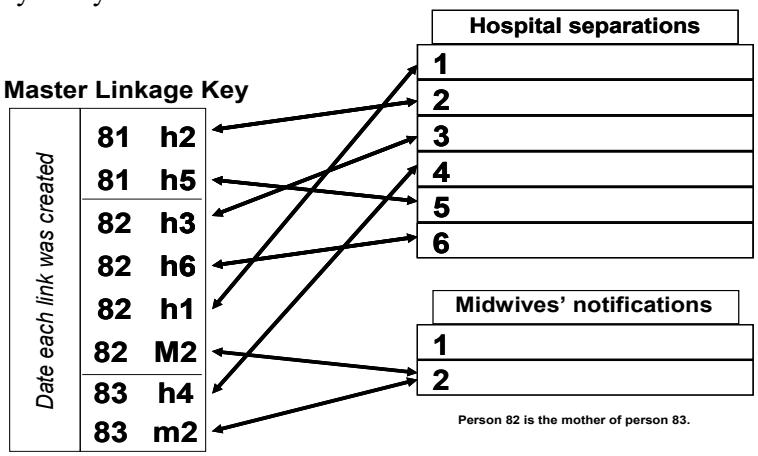
In practice, thresholds for defining matches must be tuned to achieve a balance between optimal technical performance (ie, minimal linkage errors) against the clerical resources needed to perform what can amount to thousands of reviews per day in a large system. The number of doubtful links, and the proportion accepted as links varies among datasets and depends on the quality of the identifiers, the expected match rate and whether other reference sources are available. Use of an archive of previously linked records available during the clerical review phase may improve technical performance but at the cost of considerable personnel time.

Technical performance is also determined by the matching strategy, characterised by the number and sequence of matching passes through the data, and the ways that the data are blocked, how partial identifiers are selected, how match weights are assigned and thresholds are set. An experienced data linkage unit will use different matching strategies for each linkage that relate directly to the quantity and quality of the available identifiers.

2.7 Storage, Updating and Retrieval of Links

Chains of Links

One method of storing links arising from the data matching process is in the form of ***chains of links*** (in fact, pointers enabling retrieval of source data elements), which are stored in a database or file known as a ***master linkage key*** (see figure below). Note that a single record may be in more than one chain. For example, a midwives' notification record will be linked in the chain for the mother and a chain for the infant. An advantage of the chain of links structure is that chains can be readily broken and re-joined as new information arises, either for correction of existing links or insertion of new links to newly acquired data sources. The ease with which new links can be inserted or erroneous links split dynamically from chains allows continuous improvement of linkage quality. It also allows for ready expansion of a data linkage system with no limitations on the time period or the number of data sets that can be covered by the system.



The tool sets used to manage a master key, enabling the chains to be dynamically created, extended, shortened or broken (eg, if it turns out that what was initially thought to be data on one individual turns out to be the combination of data on two individuals), are what distinguish such systems from 'snapshot' linkage projects that produce once-off linked data files for analysis.

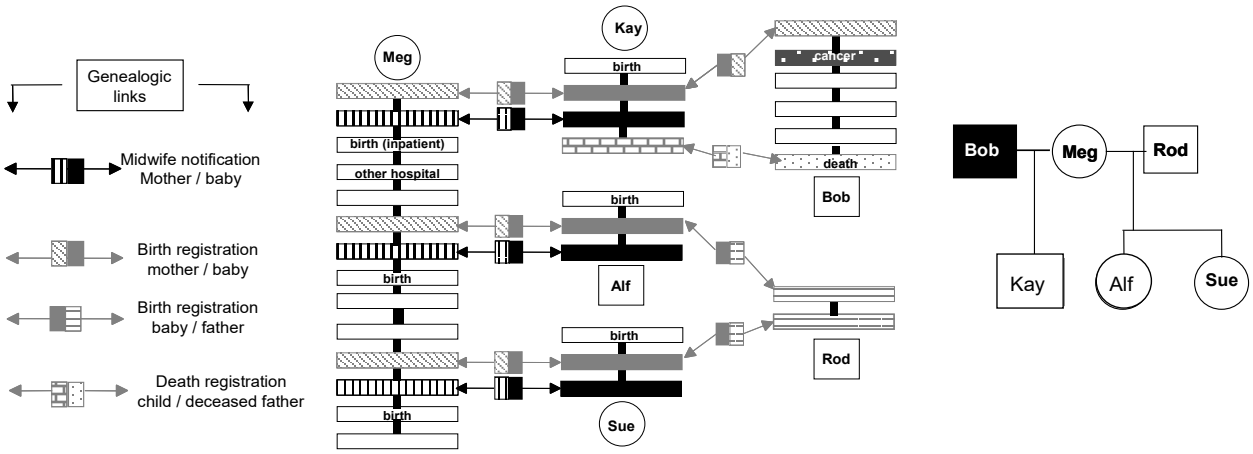
Risk Reduction Subsystems

A sophisticated data linkage system may incorporate one or more subsystems that store and use information to reduce risks of linkage errors, disaster impacts or privacy breaches. A ***no-links subsystem*** enforces rules to ensure that linkage results from clerical review are not inappropriately overridden by later automatic processes. A ***links history subsystem*** date-stamps and records transactions involving the master linkage key to support a ***re-wind facility*** so that the state of linkage on any previous date can be re-created for disaster recovery in the event of degradation of the system by corrupt data. It also enables extracts of chains consistent with the 'state of play' when a past extract was created. Links to particularly sensitive data sources may motivate an ***extra links subsystem*** to isolate sensitive links and enable their temporary connection to the master linkage key as needs arise.

Storage of Genealogical Links

Family links may be used in combination with linked health records to support genetic and human genome research, including utilisation and outcomes of genetic services. Links between genealogically related individuals may be primarily identified using parent-child information from birth registrations, supplemented with other sources such as death, marriage and midwives' notifications. For each individual identified on these records (for example, as either the decedent, parents or offspring on a death registration), a pointer to an entry in a ***genealogical links subsystem*** can be added to their chain of links in the master linkage key. Within the genealogical links subsystem are records representing genealogical relationships between parents and children from which other relationships can be implied. Pedigrees can then be constructed, with selection options such as a threshold based on the coefficient

of relatedness (the proportion of alleles held in common by two related individuals). The figure below shows an example of the structure of genealogical links and the corresponding pedigree chart of two nuclear families with a common mother, based on midwives' notifications, birth and death registrations.



2.8 Contribution of Data Linkage to Privacy Protection

Privacy-Protection Effect of a Comprehensive Data Linkage System

A comprehensive data linkage system reduces the invasion of privacy associated with use of name-identified health information in research because personal details can be removed from data released to researchers, while allowing research based on multiple data streams to be performed. There is a need to promote a better understanding of this important benefit of data linkage. Pivotal to the issue is an appreciation that the activity of data linkage is distinct from the creation and maintenance of source data collections. If the data already exist, then data linkage will allow their application in research in a way that protects privacy.

Distributed vs Centralised Data Repositories

A design construct that enables a data linkage system to enhance privacy is the separation of linkage processes from those operating on sensitive clinical data. The master linkage key is created using only a limited number of identifiers and information retained in the master linkage key is not the actual data used for research analysis, but pointers to the source data elements, the majority of which are held by custodians external to the data linkage unit who retain responsibility for approving the release of the data for research purposes. This design, known as a ***distributed data repository***, is the antithesis of the popular 'big brother' image of powerful interests accumulating a massive data bank of health information on unsuspecting members of the public; ie, a ***centralised data repository***.

You should note, however, that the distributed data repository paradigm comes at a cost to the efficiency of a data linkage system, because of the much greater levels of administrative overhead, social challenges and even technical challenges that arise when every potential research project requires a separate negotiation with each data custodian and subsequent extraction of data by the custodians, who may vary in their commitment to system, as well as their available resources and technical capacity. For these reasons, a ***hybrid data repository*** may be considered necessary as compromise, where a number of frequently used ***core data sets*** are held centrally by agreement the data custodians, whereas less frequently used data sets remain at the periphery.

The suitability of data sets to be held centrally will usually depend on the sensitivity of the information that they contain. For example, death and birth registrations may be considered relatively less sensitive than a database of mental health service contacts. Some data linkage systems get around these problems by creating an ***anonymous centralised data repository***, which is a standard set of anonymised core data sets, covering a pre-determined period of time, that are held centrally and which can be merged together using encrypted ID numbers. Subsets of these data are then made available to researchers for use in particular research projects.

2.9 Enhancing Anonymity in Data Linkage Research

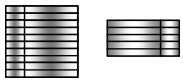
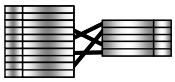
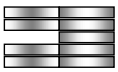
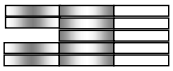
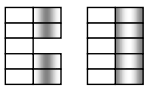
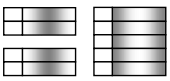
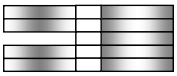
There are three areas of enhancements that stand to weigh the risk-benefit calculus even more in favour of anonymity and health research than is already the case:

Enhancing Anonymity by Governance

1. ***Privacy-sensitive protocols***: At the simplest level, these involve the use of an ***honest broker***, ***trusted third party*** or ***integrating authority***, through which outgoing data from all sources must pass. The honest broker integrates data before making the combined resources available in anonymous form with a coded identifier to a researcher. The coded identifier system is specific to each project and is not, therefore, a unique personal identifier. This form of ***pseudonymisation*** is technically reversible and hence the data are re-identifiable, although agency-level policies, contractual obligations or statutory prohibitions may bar this possibility.

   A more advanced version is the ***best practice protocol***, developed in WA and adopted by Australia's Population Health Research Network. As shown below, a 'data linkage unit' receives solely personal identifiers (names, gender, date of birth, postcode) and uses these to return encrypted project-specific identifiers to two or more data custodians. The custodians associate the project identifiers with their sensitive data and provide them in an anonymous form directly to the researcher, who assembles the sensitive data with identifiers that lack the quality of unique personal identity. This protocol ensures that:

   - no sensitive data are provided to the data linkage unit, but only limited identifiers;
   - sensitive data are not shared by data custodians;
   - identifying information is not released to researchers; and
   - breaking the anonymisation requires an improper collusion between at least three parties.

1	**Obtain A and B identity files**	
2	***Link the identity files***	
3	**Strip off the identifying data**	
4	**Add an arbitrary 'project id'**	
5	**Create 'project id <-> local id' files and return to data custodians (A and B)**	
6	**Data custodians A and B generate 'project id <-> clinical data' files**	
7	**Data custodians send clinical data files to researcher to link via 'project id'**	

2. ***Secure enclaves***: These are physically or virtually secure analysis facilities that limit and police the personnel who may have access to anonymous data and the conditions under which access is permitted. In its strictest form, the facility is a physical one, where research analysts may not make any written notes and may only collect their statistical results after a checking process to ensure that all conditions have been observed. There are trade-offs with ***research functionality***, especially if analysts are restricted in their access by time or place; capacities to create and manipulate files or run appropriate software packages; or the timeliness of checking.

Enhancing Anonymity by Technology

3. ***Advanced encryption methods***: Mathematically irreversible ***hash algorithms*** and other technologies can be used to encrypt identification numbers and even nominal identifiers such as names in a way that preserves an identity across subject matter, whilst placing the encrypted identifiers beyond any form of foreseeable re-identification process, including what is known as ***dictionary attack***. There are potential trade-offs with research functionality, such as the researcher's capacity to ask data custodians to check likely errors in provided materials.
4. ***Masking and restricted field views***: Although trimming displayed health information beyond a point that is detrimental to research in counter-productive, there is scope for the use of restricted views in situations where researchers may need to run computer syntax on fine-grained values, but not see the full data set on any one linked but anonymous identity. It is possible also to mask or aggregate sensitive information in datasets in a way that makes them less distinctive to a human observer, but still available digitally for computer analysis as the detailed values. These solutions are most readily implemented in a secure enclave and require close consultation with researchers to ensure that the approach is one that does not compromise research functionality.

Enhancements by Law

5. ***Civil law obligations not to re-identify***: Researchers who use anonymous patient-related subject matter can be contractually obliged, as a legal condition of their employment, and their employing institutions can be obliged, as a legal condition of their receipt of funding or source data, never to attempt to ascertain the personal identity associated with any health information by any means whatsoever.
6. ***Criminal law reform*** would be possible to make it a criminal offence for a researcher to attempt to ascertain a personal identity contrary to the operating rules of an approved anonymisation protocol. According to expert legal opinion, this would settle the question of reasonable ascertainability under *all* circumstances. To be fully effective, the criminalisation of improper re-identification procedures would have to be complemented by ***transborder flow*** restrictions on the transfer of anonymous data to an ***international data haven***.

Community Representation and Community Participation

Consumer representation on governance bodies overseeing data linkage and a broader commitment to community participation through communication with the media about data linkage activities, and engagement of consumer advocates and community groups in research projects all help to ensure ethical practice and transparency, and to engender public confidence in a data linkage system. Some of the indicators of a strong community participation program include:

- a dedicated consumer advocate on staff;
- budgeted participation activities in research grant applications;
- a well developed governance and policy framework for participation;
- training events relevant to participation for researchers and consumer representatives; and
- integrative forums bringing researchers, other professions and consumer representatives together.

2.10 Applications of Linked Health Data in Research

Increasingly, epidemiologists are studying diseases associated with behavioural and environmental exposures that occur over long periods of time and necessitate long term follow-up of individuals. Health services researchers are concerned with the evaluation of patterns of service provision or outcomes of interventions based on very large administrative data sets. Data linkage provides a technology to facilitate these research activities.

Epidemiologic and Human Services Research

Applications of linked data in epidemiological research, health and social services research and health and social service management may include studies in the following areas:

- ***Surveillance and needs analysis**:* This may include the collation and cross-validation of information from multiple sources to identify first-time incident events or prevalent clinical populations for needs analysis. Examples of the latter include the increasing 'active prevalence' of cancer due to wider use of chemotherapy to postpone death; and adverse mortality trends in mental health patients with prevalent ischaemic heart disease.
- ***Aetiologic research**:* Data linkage may be used to identify health events during the longitudinal follow-up of exposed cohorts, eg, a cohort-crossover study of travellers' thrombosis in passengers arriving on long-haul flights; or a risk assessment used to establish occupational exposure standards for an airborne hazard such as crystalline silica.
- ***Service utilisation research**:* This category may include studies of patterns of care in clinical populations to identify differences between subgroups, time trends or causal relationships in the processes of service delivery. Examples include inequities in access to breast reconstruction in women having undergone mastectomy; increasing readmissions driven by new technology in the treatment of ureteric calculi; and the impact of government financing policies on the use of private health insurance.
- ***Service outcomes research**.* Data linkage may be used to support safety research such as post-marketing surveillance for adverse effects of new technology such as minimally invasive surgery, assisted reproduction or new therapeutic substances; or economic outcomes such as the impact of survival of low birth weight neonates on cerebral palsy and health care costs; or efficacy and prognostic research such as improvements in survival after cancer treatments. By enabling the evaluation of clinical outcomes, data linkage can be used to evaluate new and existing treatments involving medical devices, drugs and surgical procedures.

Methodologic Research

New developments bring with them new methodologic challenges that often exceed existing textbook-based knowledge. Examples of methodologic developments arising from the WA Data Linkage System include a backcasting model to correct for the prevalent pool artefact in first-time incidence rates based on linked multi-event data; the use of cohort-crossover designs to estimate the effects of health care events on short-term readmission risks; a couplet methodology to study patterns of care and health insurance use; improved methods of risk adjustment; and case-crossover designs to study adverse drug reactions.

2.11 Administrative Data

Types of Administrative Databases

Administrative databases usually originate from systems that are either mandated by law for social or public health reasons (eg, cancer registrations) or provide or finance health care to patients (eg, insurance claims). To be useful for research, such databases must be accurate and complete; serve a well-defined population and contain information on all persons in the study population; include records for identifying disease occurrence or other study outcomes; and include other data needed to analyse a specific health services issue. The following are types of administrative databases, which relate to patient populations, commonly used in health services research:

1. ***Civil information***: These include population census collections, electoral rolls pertaining to registered voters and the records of public utilities (eg, telephone directories, water and electricity accounts) and licensing systems (eg, drivers' licences). In Australia, the smallest unit of census information available to researchers are data on collectors' districts (CDs), which are geographic groupings of about 200 households.
2. ***Births, deaths and marriages***: These records are typically held by registrar generals of births, deaths and marriages. Death records are based on medical certificates of cause of death and/or death information papers. In Australia, the Australian Institute of Health and Welfare (AIHW) hosts the National Death Index, which collates death registrations from the States and Territories.
3. ***Disease registers***: Most jurisdictions in western societies have a cancer registry supported by statutory notifications by pathologists or other providers. In Australia, these data are also aggregated nationally by AIHW into the National Cancer Clearing House. Other population-based disease registers may include those concerned with birth defects, cerebral palsy and intellectual handicap, notifiable infectious diseases, diabetes mellitus, ischaemic heart disease and stroke, chronic renal failure and other chronic disabling conditions.
4. ***Professional encounter data***: These databases are created by systems that require a certain professional group to notify a central authority each time they provide a particular class of service. An example in Australia are the midwives' notifications, whereby registered midwives must notify to state health authorities information about each confinement they have attended, whether a hospital or home birth. Another example is the statutory requirement in Western Australia for medical practitioners to notify particular classes of assisted reproduction technology (ART) services. Data collected from aged care assessment teams is another example.
5. ***Organisational encounter data***: These databases centre on organisations rather than professional groups. The most important for data linkage systems are the hospital morbidity databases, which collect abstracts of demographic and clinical information on separations from (usually) acute care hospitals. Another example in Western Australia is the Psychiatric Online Information System (PSOLIS), which records information on encounters with inpatient and public outpatient psychiatric services. Other institutional encounter systems include those concerned with waiting lists, acute hospital outpatient services, public laboratory services, home and community care services, domiciliary nursing services, hostel and nursing home residents. Data collected by non-health organisations, such as road accident data held by a police department, are also useful where appropriate in health services research.
6. ***Claims data***: These administrative databases may be found in the government or non-government sectors. In Australia, claims data generated by the Medical Benefits Scheme (MBS) and the Pharmaceutical Benefits Scheme (PBS) are the two largest systems run by Medicare Australia. Residential care subsidies paid to nursing homes and hotels are another example. Claims data held by private health insurance organisations are the most significant example in the non-government sector.

Coding of Administrative Health Data

Clinical information, such as principal diagnosis, co-morbid conditions and procedures, are usually coded to the ***International Classification of Diseases***. Coding practices change over time and the administrative data generally retain the coding used at the time when they were created. For example, in Western Australia in July 1999, the ICD 10th Revision Australian Modification (ICD-10-AM) replaced the ICD 9th Revision Clinical Modification (ICD-9-CM). The Australian MBS and PBS data have been through many complex modifications of the applicable coding systems. Researchers using early data therefore need to refer to the older coding manuals to check on changes in codes and the additions of new codes. New diseases, procedures and drug treatments may not at first be adequately coded at all, especially if there is no ICD code available. There have also been changes in coding conventions over time and, unfortunately, these are often not always well documented. Experienced clinical coders are often the best source of information about such matters.

Example: *Percutaneous coronary angioplasty (PTCA) is a frequently performed myocardial revascularisation procedure. Prior to 1988, there was no specific code for PTCA – however the ICD-9 code '5-363 : Other heart revascularisation' was the code used to indicate PTCA. After 1988, several ICD-9-CM codes were employed depending on the number of coronary vessels involved and the use or otherwise of thrombolytic agents (36.01 – 36.05). After 1995 – additional codes for coronary artery stenting were also added (36.06 and 36.07).*

Limitations of Administrative Data

Problems in the use of administrative data collections in research include the following:

Missing or incorrect data: A number of validation studies have assessed the quality of data in administrative databases, such as hospital morbidity data. While there are no hard and fast rules, the following conclusions about hospital morbidity data are generally supported by validation research:

- Missing records are generally not a serious problem, especially when data collections have a well-organised statutory framework or submission of data is motivated by financial rewards.
- Missing demographic data items are usually very uncommon; <1%.
- Data on procedures are usually recorded and accurately coded in 90%+ of cases. Data on non-procedural treatments that do not have financial implications for the provider (eg, inpatient chemotherapy) are usually poorly recorded.
- Primary diagnoses are usually recorded and accurately coded in 90%+ of cases.
- Secondary diagnoses and complications are recorded with variable accuracy depending on the nature of the condition. The main problem is the sensitivity of recording some conditions. Conditions such as diabetes and cancer tend to be relatively well recorded as secondary diagnoses (≈80% sensitivity); ischaemic heart disease is moderately recorded (≈70%) and conditions such as dementia and ulcers are poorly recorded (≈10%).

Two important actions arise from the consideration of the quality of administrative data. First the need for validation studies when important conclusions are to be derived from diagnostic fields other than the principal diagnostic field in a data set; and second the need for caution in the interpretation of outliers (e.g., unusually low or high utilisation or outcomes). While outliers may represent practice patterns that need correction or are models of excellence, experience has taught many researchers that data quality problems should be considered as the first explanation for an outlier.

Example: *Injury to the common bile duct is a serious complication of laparoscopic cholecystectomy. Researchers in WA have found that only a low proportion of these injuries, identified by review of hospital charts, were recorded and coded in the hospital morbidity data system. Thus the quality of administrative data was unsuitable for comparing the occurrence of this adverse event between laparoscopic cholecystectomy and open cholecystectomy, whereas differences in length of stay and case fatality rates are supported by administrative data of adequate quality.*

Insufficient subgroup specification: The level of detail achieved by the classification systems used in administrative databases may be insufficient to define important subgroups of patients, who are either the specific object of study or who must be considered as separate strata for study results to be interpretable. Coding systems such as ICD-9-CM do not always provide the level of detail that is required and even fairly basic information, such as the laterality of a procedure (e.g. whether a right or left herniorrhaphy) may be undefined and reduce the utility of the database without supplementary data collection.

Example: A frequent problem with data contained in cancer registries is the lack of information on clinical staging. Because clinical staging usually has a profound effect on cancer survival, it is often inappropriate to use cancer registry data alone for health services research into outcomes of cancer treatment. This is one of the motivations for the establishment of hospital-based cancer registries that feed into a population-based master register.

Non-ascertainment of confounders: Constructing valid comparisons that are sufficiently free of selection bias, confounding and information bias is a challenge for all forms of non-experimental health research and especially for that based on administrative data sets. The capacity for risk adjustment may be limited by lack of recording of important data items. For example, smoking and other lifestyle variables that can confound statistical associations are seldom recorded in administrative databases. The methods include restriction, stratification and modelling. When detailed data on confounders are not recorded, sometimes useful improvements in validity can still be achieved using restriction.

Example: A study of the effect of splenectomy on the long-term risk of septicaemia could be undertaken using hospital morbidity data and a comparison group of people admitted to hospital, but who did not have a splenectomy. Even with stratification by age and sex, the potential for confounding by sociodemographic and lifestyle factors and by intention to treat (the two main indications for splenectomy are ruptured spleen and myeloproliferative disorders) is highly problematic. Greater validity of comparison could be obtained by restricting the study design to patients who sustain multiple injuries from motor vehicle accidents and stay in hospital at least seven days. These patients are then divided into those who did and did not require a splenectomy, but are likely to be much more homogenous with respect to extraneous confounders.

Non-ascertainment of study outcomes: Administrative databases are usually adequate to detect outcomes that reliably result in diagnoses or treatments recorded in the database. The more severe the condition the more complete ascertainment is likely to be. Outcomes associated with a procedure, readmission or death may be readily ascertained. Those involving subclinical or ambulatory health problems, functional status or quality of life are unlikely to be detected through the use of administrative data and review of patient records and/or special surveys may be required.

Example: The use of administrative data to study outcomes of carotid endarterectomy may be adequate to establish patterns of survival and the risk of stroke following the procedure, but evaluation of outcomes in terms of a change in transient ischaemic attacks and quality of life would most likely require special surveys to be conducted.

Data checking and cleaning

Most established data linkage systems will routinely undertake some form of quality assurance checking on the linkage and extraction of data to be provided to researchers. However, the focus and extent of these checking procedures varies widely between systems and often rely on a generic set of algorithms that are not necessarily tailored towards individual projects.

Regardless of the services offered by data linkage systems, it is good research practice to always undertake a period of data checking and cleaning of any data sets received before embarking on the analyses for the project proper. These processes should consider intra- as well as inter-dataset checking to ensure accurate and consistent data exists both within and between all linked data sources received. Depending on the specific project scope, number of data sets and amount of data, this process can take anywhere from a few days to a number of months.

A good starting point is to create a checklist of items to be reviewed which can then be built upon as more projects are undertaken. Such checklists can be provided to analysts in a research group commencing work on a new project to ensure data cleaning follows a standardised and acceptable process across researchers and projects. Some of the more common checks performed with data cleaning are as follows:

- Review of descriptive statistics for expected patient demographic characteristics
- Determination of any missing data
- Review of outliers
- Consistency of each data field both within and between datasets
- Checking that all records are within the correct study observation period
- Review of records out of the study area
- Checking for gaps in data for individual patients, possibly indicative of migration.
- Checking if units are out of scope or incorrectly classified
- Validation of the administrative data against other sources if possible
- Cohort check to ensure that each person has at least one record that matches the cohort selection criteria
- Identification of impracticable circumstances such as a service contact after death
- Clinical consistency checks between variables or related data sets

References:

Aboagye-Safro P, Mai Q, Sanfilippo FM, Preen DB, Stewart L, Fatovich D. **A comparison of multivariate and univariate time series approaches to modelling and forecasting Emergency Department demand in Western Australia.** *The Journal of Biomedical Informatics* 2015; 57; 62-73.

Acheson ED. **Medical Record Linkage.** London: Oxford University Publications, 1967.

Acheson ED (Ed). **Record Linkage in Medicine. Proceedings of the International Symposium, Edinburgh, July 1967.** London: ES Livingstone, 1968.

Baldwin JA, Acheson ED, Graham WJ (eds). **Textbook of Medical Record Linkage.** Oxford: OUP, 1987.

Bartu A, Freeman NC, Gawthorne GS, *et al*. **Mortality in a cohort of opiate and amphetamine users in Perth, Western Australia.** *Addiction* 2004; 99: 53-60.

Brameld K, Holman D, Thomas M, *et al*. **Use of a State data bank to measure incidence and prevalence of a chronic disease: end-stage renal failure.** *Am J Kidney Disease* 1999; 34: 1033-1039.

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 Com Health* 1998; 2: 740-744.

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, Threllfall 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, Holman CDJ. **Estimation of excess risk of readmission to hospital after an index inpatient separation.** *Med Care* 2003; 41: 693-697.

Brameld KJ, Holman CDJ. **The effect of locational disadvantage on hospital utilisation and outcomes in Western Australia.** *Health and Place* 2006; 12: 490-502.

Brameld KJ, Holman CDJ. **The use of end-quintile comparisons to identify under-servicing of the poor and over-servicing of the rich: a longitudinal study of the effect of socioeconomic status on hospital utilisation and outcomes.** *BMC Hlth Serv Res* 2005; 5: 61.

Calver J, Wiltshire A, Holman CDJ, *et al*. **Does health assessment improve health outcome in indigenous people? A RCT with 13 years of follow up.** *ANZ J Public Health* 2005; 29: 107-111.

Calver J, Preen DB, Sanfilippo FM, Bulsara MK. **Stimulant prescribing for the treatment of ADHD in Western Australia: socio-economic and remoteness differences.** *Medical Journal of Australia* 2007, 186(3):124-127.

Clark A, Preen DB, Ng JQ, *et al*. **Is Western Australia representative of other Australian States and Territories in terms of key socio-demographic and health economic indicators?** *Aust Hth Rev* 2010; 34: 210-215.

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.

Einarsdóttir K, Preen DB, Emery J, *et al*. **Regular primary care decreases the likelihood of mortality and morbidity in older people with chronic disease.** *Long Life Stud Int J* 2010; 1 Suppl: 59 (S3.4.2).

Einarsdottir K, Preen DB, Emery J, *et al*. **Regular primary care plays a significant role in secondary prevention of ischemic heart disease in a Western Australian cohort.** *J Gen Int Med* 2011; 26: 1092-1097.

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 Pub Hth* 2010; 10: 385, doi:10.1186/1471-2458-10-385.

Fellegi IP, Sunter AB. **A theory for record linkage.** *J Am Stat Assoc* 1962; 5: 563-566.

Fletcher DR, Hobbs MST, Tan P, *et al*. **Complications following cholecystectomy. Risks of the laparoscopic approach and protective effects of operative cholangiography: a population based study.** *Annals Surg* 1999; 229:449-457.

Gibson DAJ, Moorin R, Preen DB, *et al*. **Can increasing physician fee-for-service payments improve service regularity in elderly patients with chronic disease?** *Long Life Stud Int J* 2010; 1 Suppl: 60 (S3.4.3).

Gibson DAJ, Moorin RA, Preen D, *et al*. **Enhanced primary care improves GP service regularity in older patients without impacting on service frequency.** *Aust J Prim Care* 2012; doi:10.1071/PY11050.

Gibson DAJ, Moorin RE, Preen DB, *et al*. **Effects of the Medicare enhanced primary care program on primary care physician contact in the population of older Western Australian with chronic diseases.** *Aust Health Rev* 2011; 35: 334-340.

Gill L, Goldacre M, Simmons H, *et al*. **Computerised linking of medical records: methodological guidelines.** *Epidemiol Community Health* 1993; 47: 316-319.

Glover J, Rosman DL, Tennant S. **Unpacking analyses relying on area-based data: are the assumptions supportable?** *Int J Health Geog* 2004; 3: 30.

Hall SE, Holman CDJ, Finn J, *et al*. **Improving the evidence base for promoting quality and equity of surgical care using population-based linkage of administrative health records.** *Int J Quality in Health Care* 2005; 17: 415-420.

Hall SE, Holman CDJ, Wisniewski ZS, *et al*. **Prostate cancer: socio-economic, geographical and private-health insurance effects on care and survival.** *BJU Int* 2005; 95: 51-58.

Hall SE, Holman CDJ. **Inequalities in breast cancer reconstructive surgery according to social and locational status in Western Australia.** *Eur J Surg Oncol* 2003; 29: 519-525.

Hansen M, Kurinczuk JJ, Bower C, *et al*. **The risk of major birth defects after instracytoplasmic sperm injection and in vitro fertilization.** *NEJM* 2002; 346: 725-730.

Hobbs MST, McCall MG. **Health statistics and record linkage in Australia.** *J Chron Dis* 1970; 23: 375-381.

Holman CDJ. **Anonymity & Research: Health Data and Biospecimen Law in Australia**. ISBN 978-0-646-57417-2. Perth: UniPrint, The University of Western Australia, 2012.

Holman CDJ, Bass J, Rouse IL, *et al*. **Population-based linkage of health records in Western Australia: development of a health services research linked database.** *ANZ J Public Health* 1999; 23: 453-459.

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 Int* 2002; 90: 7-15.

Ingarfield SL, Finn JC, Jacobs IG, *et al*. **Use of emergency departments by older people from residential care: a population based study.** *Age and Ageing* 2009; 1-5, doi: 10.1093/ageing/afp022.

Jaro MA. **Advances in record linkage methodology as applied to matching the 1985 census of Tampa, Florida.** *J Am Stat Assoc* 1989; 84: 16-22.

Jaro MA. **Probabilistic linkage of large public health data files.** *Stats Med* 1995; 14: 491-498.

Kelman CW, Bass AJ, Holman CDJ. **Research use of linked health data - a 'best practice' protocol.** *ANZ J Public Health* 2002; 26: 251-255.

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

Laurvick CL, Semmens JB, Leung Y, *et al*. **Ovarian cancer in Western Australia (1982-98): trends in surgical intervention and relative survival.** *Gynae Oncol* 2003; 88: 141-148.

Lawrence D, Holman CDJ, Jablensky AV, *et al*. **Excess cancer mortality in Western Australian psychiatric patients due to higher case fatality rates.** *Acta Psych Scand* 2000; 101: 382-388.

Lawrence DM, Holman CDJ, Jablensky AV, *et al*. **Death rate from ischaemic heart disease in Western Australian psychiatric patients 1980-1998.** *Br J Psych* 2003; 182: 31-36.

Moorin R, Gibson D, Bulsara CE, *et al*. **Challenging the perceptions of cancer service provision for the disadvantaged: evaluating utilisation of cancer support services in Western Australia.** *Supp Care Cancer* 2012: doi 10.1007/s00520-011-1262-1.

Moorin R, Holman CDJ. **Impact of the evolution of surgical procedures for low back pain: a population based study of patient outcomes and hospital utilisation**. *ANZ J Surg* 2009; 79: 610-618.

Moorin RE, Brameld KJ, Holman CDJ. **The influence of federal health care policy reforms on the use of private health insurance in disadvantaged groups.** *Aust Health Rev* 2006; 30: 241-251.

Moorin RE, Holman CDJ, Garfield C, *et al*. **Health related migration: evidence of reduced "urban-drift".** *Health and Place* 2006; 12: 131-140.

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.

Newcombe HB, Kennedy JM, Axford SJ, *et al*. **Automated linkage of vital records.** *Science* 1959; 130: 954-959.

Newcombe HB. **Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business.** Oxford: Oxford University Press, 1988.

Patterson KM, Holman CDJ, English DR, *et al*. **First-time hospital admissions with illicit drug problems in indigenous and non-indigenous Western Australians: an application of record linkage to public health surveillance.** *ANZ J Public Health* 1999; 23: 460-463.

Preen DB, Holman CDJ, Lawrence DM, Bayhnam H, Semmens JB. **Hospital chart review provided more accurate comorbidity information than data from a general practitioner survey or an administrative database.** *Journal of Clinical Epidemiology* 2004; 57: 1295-1304.

Preen DB, Emery J, Holman CDJ. **The influence of frequency and periodicity of primary health care contact on disease progression, unplanned hospitalisation and mortality in patients with type 2 diabetes.** *Long Life Stud Int J* 2010; 1 Suppl: 181 (T5.5.3).

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.

Preen DB, Holman CDJ, Lawrence DM, Semmens JB. **Modelling comorbidity recency, duration and severity with risk adjustment for mortality and morbidity outcomes with administrative data.** *Australasian Epidemiologist* 2004**; 11(1): 12-19.**

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.

Qun Mai, Holman CDJ, Sanfilippo FM, *et al*. **Mental illness related disparities in diabetes prevalence, quality of care and outcomes: a population-based longitudinal study.** *BMC Med* 2011; 9: 118, doi:10.1186/1741-7015-9-118.

Qun Mai, Holman CDJ, Sanfilippo FM, *et al*. **The impact of mental illness on potentially preventable hospitalisations: a population-based cohort study.** *BMC Psych* 2011; 11: 163, doi:10.1186/1471-244X-11-163.

Ray, WA. **Policy and program analysis using administrative databases.** *Ann Int Med* 1997; 127: 712-718.

Roos LL, Mustard CA, Nicol JP, *et al*. **Registries and administrative data: organisation and accuracy.** *Med Care* 1993; 31: 201-212.

Roos LL, Nicol P. **A research registry: uses, development, and accuracy.** *J Clin Epidemiol* 1999; 52: 39-47.

Roos LL, Roos N. **Of space and time, of health care and health.** *J Epidemiol Community Health* 2001; 6: 120-122.

Roos LL, Wajda A, Nicol JP. **The art and science of record linkage: methods that work with few identifiers.** *Comput Biol Med* 1986; 16: 45-57.

Roos NP, Black CD, Frohlich N, *et al.* **A population-based health information system.** *Medical Care* 1995; 33: DS13-20.

Rosman DL, Knuiman MW. **A comparison of hospital and police road injury data.** *Accident Anal Prev* 1994; 26: 215-222.

Semmens JB, Aitken RJ, Sanfilippo FM, *et al*. **The Western Australian audit of surgical mortality advancing surgical accountability.** *Med J Aust* 2005; 183: 504-508.

Semmens JB, Lawrence-Brown MMD, Fletcher DR, *et al*. **The Quality of Surgical Care Project: a model to evaluate surgical outcomes in Western Australia using population-based record linkage.** *ANZ J Surg* 1998; 68: 397-403.

Sibthorpe B, Kliewer E, Smith L. **Record linkage in Australian epidemiological research: health benefits, privacy safeguards and future potential.** *Aust J Public Health* 1995; 19: 250-6.

Smith FJ, Holman CDJ, Moorin RE, *et al*. **Lifetime risk of undergoing surgery for pelvic organ prolapse.** *Obstet Gynaecol* 2010; 116: 1096-1100.

Smith M, Semmens J, Rosman D, *et al*. **International Health Data Linkage Network.** *Healthcare Policy* 2011; 6: 94-96.

Smith ME, Newcombe HB. **Automated follow-up facilities in Canada for monitoring delayed health effects.** *Am J Public Health* 1980; 70:1261-1268.

Stanley F, Blair E, Alberman E. **Cerebral Palsies: Epidemiology & Causal Pathways.** Cambridge: Mac Keith Press, 2000.

Stanley FJ, Croft ML, Gibbons J, *et al*. **A population database for maternal and child health research in Western Australia using record linkage.** *Paed Perinatal Epidemiol* 1994; 8: 433-447.

Stanley FJ, Read AW, Kurinczuk JJ, *et al*. **A population maternal and child health research database for research and policy evaluation in Western Australia.** *Semin Neonatol* 1997; 2: 195-201.

Stewart LM, Holman CDJ, Hart R, *et al*. **How effective is in vitro fertilization and how can it be improved?** *Fert Steril* 2011; 95: 1677-1683.

Wajda A, Roos LL. **Simplifying record linkage: software and strategy.** *Comput Biol Med* 1987; 17: 239-248.

TRAINING SESSION 2: PREPARING A LINKED FILE AND CHARACTERISING A POPULATION

IMPORTANT NOTICE

The practice data sets provided for training purposes in this course are synthetic. All results based on analysis of these synthetic/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

For your training session on preparing a linked file for analysis and characterising a population, you will replicate the preparatory work of a Western Australian research team who used linked health data to study outcomes following in-hospital vasectomy and vasovasostomy (vasectomy reversal). The fictitious data sets modelled on this study offer an excellent resource for training purposes, because they involve four separate types of data links: (i) internal linkage of hospital morbidity data; (ii) external linkage of death registrations to hospital morbidity data; (iv) external linkage of cancer notifications to hospital morbidity; and (iv) a genealogical linkage of hospital morbidity to paternity information on birth registrations.

Training Session Resources

The following data files are used in this training session:

If you are using this statistical package …	*… you should use this save file.*
SPSS	vashmds.sav vasdeath.sav vascancer2.sav vasbirth.sav
SAS	vashmds.sas7bdat vasdeath.sas7bdat vascancer2.sas7bdat vasbirth.sas7bdat
Stata	vashmds.dta vasdeath.dta vascancer2.dta vasbirth.dta
R	vashmds.rda vasdeath.rda vascancer2.rda vasbirth.rda

Note that you should have saved the vascancer2 file from exercise 1.

Each data set contains a rootlpno, being the unique identifier for the same individual in all of these files. The presence of rootlpno in vashmds, vasdeath and vascancer2 simulates the situation where the links have been created by the routine procedures of the WA Data Linkage Unit. In the case of vasbirth, the links simulate the situation where they have been created from a special *ad hoc* linkage of men who had a vasovasostomy to the paternity field on birth registrations, which captured the birth date of the first child legally fathered by the man after the date of the vasovasostomy.

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

Procedure	Coding System	Codes
Vasectomy	ICPM (pre-1988)	5-636, 5-981
	ICD-9-CM (1988 to June 1999)	63.7*X*
Vasovasostomy	ICPM (pre-1988)	5-634, 5-637
	ICD-9-CM (1988 to June 1999)	63.8*X*

Note that '*X*' is a *wild* character, which may represent 0-9. **Useful tip**: be aware that diagnostic and procedural codes do not always come in data sets exactly as they do in the coding manuals. Sometimes decimal points and dashes are omitted or shifted; eg, the code 5-636 may appear in a data set as '5-636' (a string variable), 5636 (string or numeric), 5.636 or even 56.36 (numeric or string).

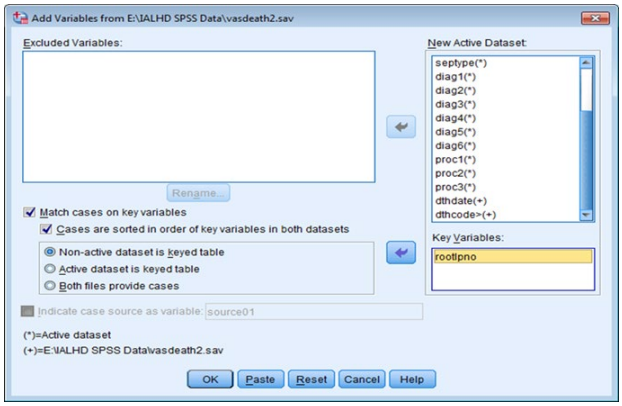
Merging Files

In order to perform generic external loading of data from an external file into the EOR loading area of a working file, you will need to be familiar with the procedure used in your statistical software package to merge files so as to add variables. The procedures are as follows:

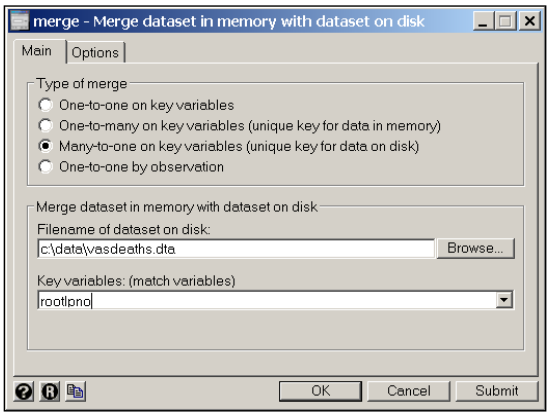
Package:	SPSS	SAS	Stata	R
Merge files so as to add variables from an external file to a working file.	MATCH FILES (*data* | *merge files* | *add variables* in pull-down menu mode)	merge …; by ……;	merge [varlist] using filename, options (*data* | *combine datasets* | *merge two datasets* in pull down menu mode)	left_join(dataframe1,dataframe2, by= c(dataframe1 linking field name = dataframe2 linking field name), keep = FALSE)

These procedures will rely on the use of rootlpno as the key variable and it is essential that both the working file and the external file are sorted in the same ascending order of rootlpno. For the type of merge that we will perform it is also essential that the external file (but not the working file) has a maximum of only one record for each individual; ie, no more than one record per rootlpno. In some packages, the external file in this arrangement is termed a ***keyed table*** because for each record in the working file, the new information to be added to the EOR loading area is 'looked-up' in the external file, using the rootlpno as a key to indicate the correct information to be transferred. The window opposite shows what this procedure looks like in pull-down menu mode in SPSS (you will be doing this procedure in the exercise). The windows below show what the equivalent procedures look like in SAS and one version of Stata.

SPSS



Stata



SAS	R
Merging in SAS is not done by pull down menus data hmds2; merge dlcourse.vashmds (in=mor) dlcourse.vasdeath (in=dth); by rootlpno; if mor then output; run;	**Merging in R is not done by pull down menus** vaslinks <- left_join(vashmds,vasdeath, by=c("rootlpno"="rootlpno"), keep = FALSE)

Tagging Records Mentioning a Target Condition or Procedure

Often it is desired to identify all records in a file that mention one or more target conditions or procedures. If there are multiple diagnosis or procedure fields per record, a ***loop command*** is used to proceed through each of the fields in turn and to check for the presence of any of the target codes. If a target code is found, the record is 'tagged' by assigning a designated value to a new variable in the EOR loading area. Below are simplified examples of syntax, in each of the three languages, designed to tag records mentioning a target code ('XXX.X') in any of six diagnostic fields. You should note how the loop command is constructed in your language: DO REPEAT in SPSS; array/do-end in SAS and foreach in Stata.

SPSS

```
COMPUTE tag=0.
DO REPEAT dummy=diag1 TO diag6.
  IF (dummy EQ XXX.X) tag =1.
END REPEAT.
EXECUTE.
```

SAS

```
data file2;
set file1;
tag = 0;
array dummy{6} diag1-diag6;
  do i = 1 to 6;
    if dummy{i}  = XXX.X then tag = 1;
  end;
drop i;
run;
```

Stata

```
generate tag=0
foreach x in diag1 diag2 diag3 diag4 diag5 diag6 {
recode tag(0=1) if (`x'==XXX.X)
}
```

R

```
data_frame$tag <- 0
diag1_6 <- c(paste("diag",seq(1,6),sep=""))
for (x in diag1_6){
  for(i in 1:nrow(data_frame)){
      if(data_frame[i, x] == XXX.X) data_frame$tag[i] <- 1
  }
}
```

Changes in Coding System

A further aspect of tagging records that mention target conditions and procedures is the likelihood that your syntax will need to deal with introductions of new coding systems. This is achieved by use of conditionals based on the values of a relevant date variable, which indicates what coding system is applicable to each record. In general, for hospital morbidity data it will be the date of separation, and not the date of admission, that indicates which coding manual is applicable.

```
Pseudo-syntax
  Set tag to 0
  REPEAT for dummy = diag 1 TO diag6
    IF date is within 1st period and dummy = 1st period target value THEN set tag to 1
    IF date is within 2nd period and dummy = 2nd period target value THEN set tag to 1
  END
```

Tag Variables with Positional Attribution

Note that there are more sophisticated ways of accounting for the possibility of multiple target codes for conditions or procedures occurring on the same record. One such approach is to combine all of the information into a single tag variable by use of ***positional attribution***. This is a way of constructing a tag variable value such that each numeric position in the value represents the occurrence of another target code on the record. For example, if you are searching for two target conditions in several diagnostic fields, a tag variable could summarise the result of the search using the values: blank, '1', '2', '11', '12', '21' or '22', '111', '112', etc. The value '21' would mean that the first target condition was the first to be mentioned in any field, whereas the second target condition was the second to be mentioned. The value '111' would mean that the first target condition was mentioned three times on the record. The utility of this approach is that a frequency distribution of the tag variable values will reveal all of the combinations of target conditions or procedures occurring in the data. A second advantage is that these string-like numeric codes can be operated on in different ways to yield any number of desired groupings.

The following ***pseudo-syntax*** (ie, a generic form of computing syntax that is neither specifically SPSS, SAS, Stata nor R) is an example of how a tag variable with positional attribution can be built up using a loop command when searching across six diagnostic fields for either of two target condition codes. Note how the exponentiation of the counter variable provides the basis for positional attribution of any target codes that are found. For example, if the first target value is found first, then at that point in the loop tag $= 0 + 1 \times 10^0 = 1$. If the second target value is found next, the value of the tag variable updates to tag $= 1 + 2 \times 10^1 = 21$. This option is outlined only for information; you need not implement it in this short course.

```
Pseudo-syntax
  Set tag to 0
  Set counter to 0
  REPEAT for dummy = diag 1 TO diag6
    IF dummy = 1st target value THEN tag = tag + 1 × 10^counter
    IF dummy = 2nd target value THEN tag = tag + 2 × 10^counter
    IF dummy = any target value THEN counter = counter + 1
  END
```

Sequence Variables

The analysis of a file of linked data is often facilitated by the addition in the EOR loading area of a number of variables with ordinal integer values, representing the sequences of records in various ways. Most of these ***sequence variables*** are premised on a file structure in which records pertaining to the same individual are grouped together in a block in preferred sequential order. In the first exercise you were introduced to one of these sequence variables, being the morbidity sequence variable. In fact, there are several types, of which the following are the most useful:

1. A ***file sequence variable*** (suggested variable name: fileseq), which identifies the ordinal sequence of the entire file in its preferred sequential state. Usually the preferred sequence is individuals in ascending order of rootlpno and records for individuals in ascending chronological order, such as by hospital separation or admission date. The fileseq variable is useful in sorting a file back to its original preferred sequence, after the sequence has been disrupted.
2. A ***morbidity sequence variable*** (morbseq) or *block sequence variable*, which identifies the ordinal sequence of records in the file pertaining to the same individual. Each time a new individual is encountered in the file the morbseq variable is reset to 1.
3. An ***index sequence variable*** (indexseq), which identifies the ordinal sequence of records in the file pertaining to the same individual, but commencing with the value of 1 for the index record. The ***index record*** is the first to mention the target condition, procedure or event that is the object of the research. The value of indexseq assigned to an individual's records encountered prior to the index record may be assigned as blank, 0, or sometimes it is preferable to number them as -1, -2, -3, and so on, moving backwards from the index record. The following pieces of syntax illustrate different syntactical approaches in each of the three languages: use of a scratch variable in SPSS, by-group processing in SAS and derivation of indexseq from a 0-1 tag variable in Stata.

SPSS (scratch variable)

```
IF (morbseq EQ 1) #flag1 = 0.
IF (#flag1 EQ 1) indexseq = lag(indexseq) + 1.
DO IF (tag EQ 1 AND #flag1 EQ 0).
  COMPUTE indexseq = 1.
  COMPUTE #flag1 = 1.
END IF.
IF (#flag1 EQ 0) indexseq = 0.
EXECUTE.
```

SAS (by-group processing)

```
data file2;
set file1;
by rootlpno;
retain indexseq;
if first.rootlpno then indexseq = 0;
if tag = 1 or indexseq>0 then indexseq = indexseq + 1;
run;
```

Stata (derivation from 0-1 tag variable)

```
generate indexseq=tag
replace indexseq=indexseq[_n-1]+1 if rootlpno==rootlpno[_n-1] & indexseq[_n-1]~=0
```

R

```
flag1 = 0 # proxy for SPSS scratch variable
vaslinks$vasseq <- 0
attach(vaslinks)
for(i in 1:nrow(vaslinks)){
  if(morbseq[i]==1) flag1 = 0
  if(flag1 == 1) vaslinks$vasseq[i] = vaslinks$vasseq[i-1] + 1
  else
    if(treat[i]==1 && flag1==0){
      vaslinks$vasseq[i] = 1
      flag1 = 1
    }
  else
    if(flag1==0) vaslinks$vasseq[i] = 0
}
detach(vaslinks); rm(flag1)
```

SPSS users: The syntax above may seem unduly complex and, indeed, it is possible to avoid the use of a scratch variable by using conditionals similar to those implied and expressed in the SAS syntax; or an identical approach to that in the Stata syntax if the records with target codes are marked by a simple 0-1 tag variable. The utility of the more complex approach is that it enjoys greater flexibility to cope with different forms of tag variables and with variations in the design of index sequence variables. Most newcomers to SPSS syntax will find this syntax confusing at first. In particular, the second line "IF (#flag1 EQ 1) indexseq = lag(indexseq) + 1." seems non-sensical given that in the first line #flag1 has been set to 0 at the beginning of the block. The key to understanding it is to appreciate ***file processing order***: the computer applies all the syntax to the first record in the file, then all the syntax to second record, and so on. The fifth line of syntax, "COMPUTE #flag1 = 1.", has the capacity to change the value of #flag1 to 1 once the index record has been found. If this has occurred during processing of an earlier record, the conditional in the second line of syntax will be satisfied.

SAS users: Note the use of the retain statement, which has the effect of carrying over and assigning the value of a variable from one record to the next unless further processing of the value takes place.

Stata users: Note that in today's practical exercise, you will find that the tag variable, called treat, is not initially in a 0-1 format because treat is used to tag two different types of index record simultaneously. Thus, some recoding is required in conjunction with the syntax above.

R users: In this example, we iterate through an additional loop: one for each row (indexed *i*). This is likely inefficient from a programming perspective, but it is one way of allowing calculations to be done on a row-by-row basis, as is required for many linked health data applications. This step is not required to be specified as such in the other software programs, which make use of a default *file processing order* of operations or similar method. (Some alternatives to loops for R users are provided in the Syntax Solutions).

Overview of Exercise 2

Step 1. Open vashmds file.

Step 2-4. Merge death data from vasdeath file.

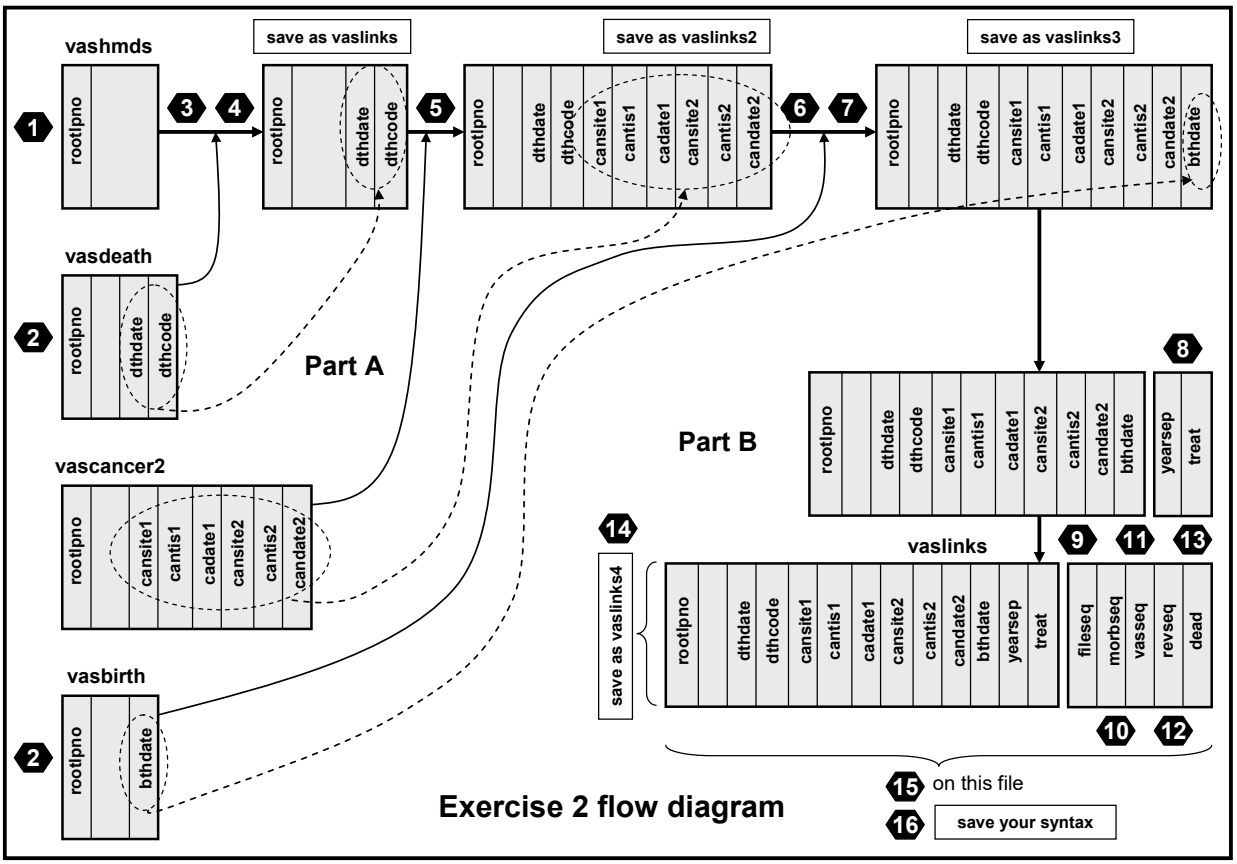
Step 5. Merge cancer data from vascancer2 file.

Step 6-7. Merge birth data from vasbirth file.

Step 8. Tag hospital records with target procedure codes.

Step 9-12. Create fileseq, morbseq and indexseq (×2).

Step 13-14. Characterise the clinical population.

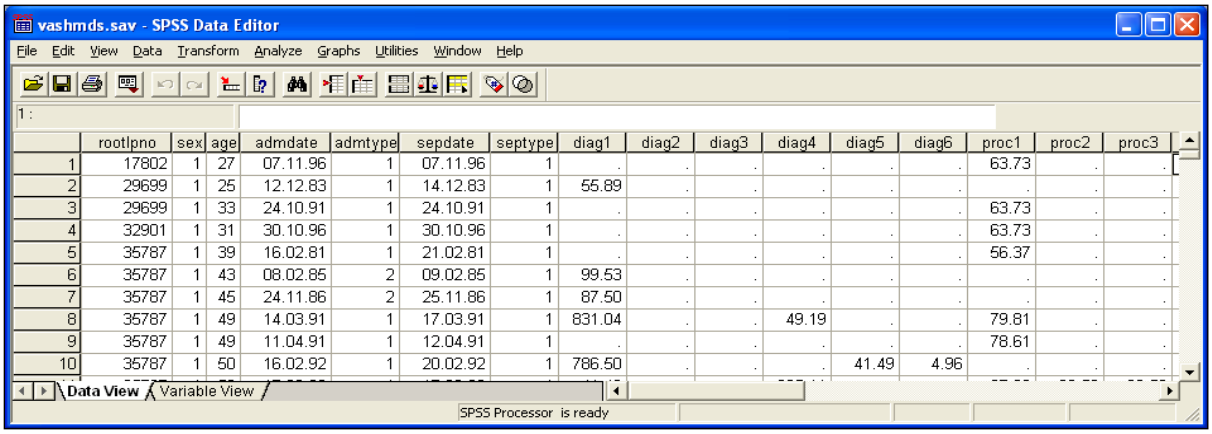


Exercise Instructions Part A: Merging Files

In this exercise, you will be loading data onto a working file, taken from a series of external files, in order to create a composite linked file of all the data needed for later analysis, including future exercises in the course.

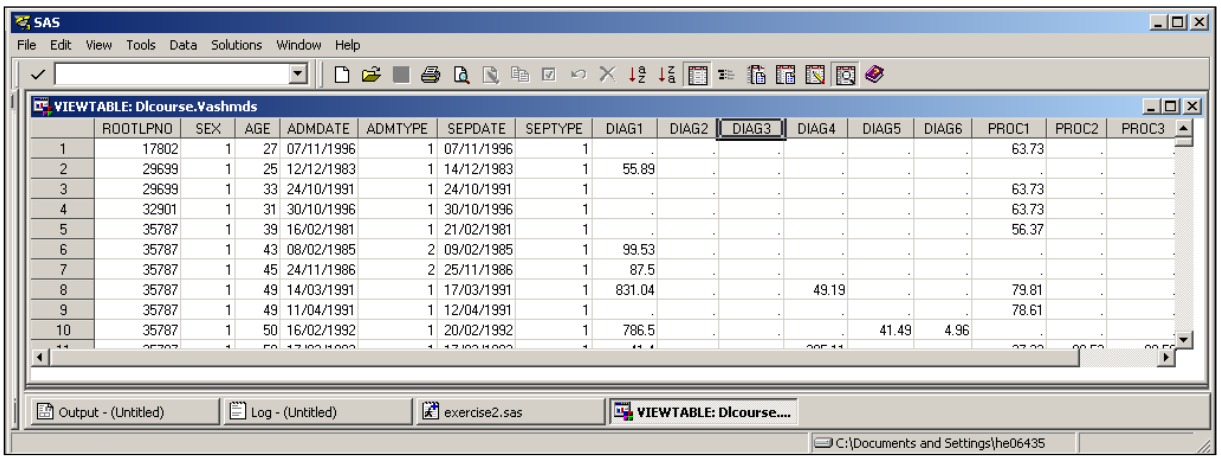
1. Open the vashmds data file (or vashmds1 file created yesterday) as your working file and examine the file's layout. It is a typical linkable hospital morbidity data file of 9,406 records, containing a rootlpno, some demographics (sex and age), admission and separation dates and types (admdate, admtype, sepdate, septype), six diagnosis variables (diag1 to diag6) and three procedure variables (proc1 to proc3). Most real-life hospital morbidity data files will contain more fields than this for diagnoses and procedures, and may also contain fields for coding external causes.

SPSS



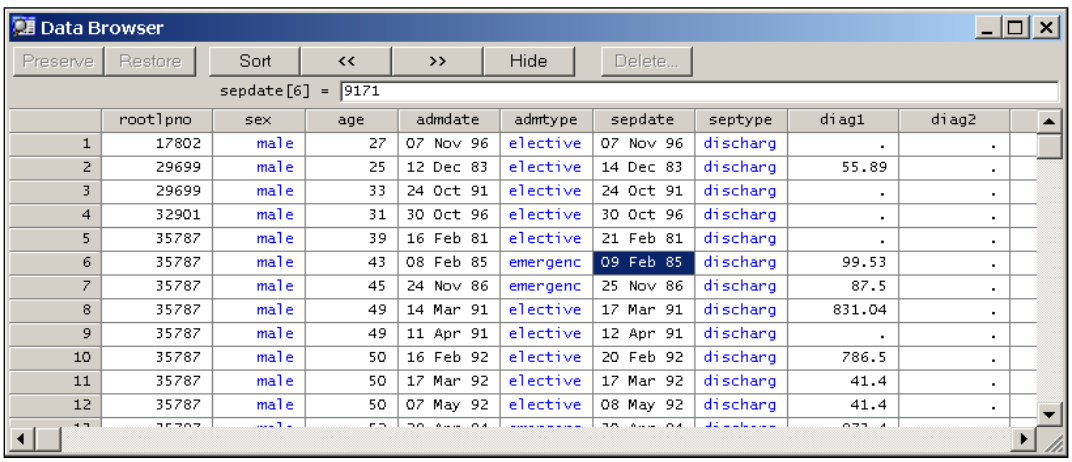
	rootlpno	sex	age	admdate	admtype	sepdate	septype	diag1	diag2	diag3	diag4	diag5	diag6	proc1	proc2	proc3
1	17802	1	27	07.11.96	1	07.11.96	1	.	.	.	.	.	.	63.73	.	.
2	29699	1	25	12.12.83	1	14.12.83	1	55.89	.	.	.	.	.	.	.	.
3	29699	1	33	24.10.91	1	24.10.91	1	.	.	.	.	.	.	63.73	.	.
4	32901	1	31	30.10.96	1	30.10.96	1	.	.	.	.	.	.	63.73	.	.
5	35787	1	39	16.02.81	1	21.02.81	1	.	.	.	.	.	.	56.37	.	.
6	35787	1	43	08.02.85	2	09.02.85	1	99.53	.	.	.	.	.	.	.	.
7	35787	1	45	24.11.86	2	25.11.86	1	87.50	.	.	.	.	.	.	.	.
8	35787	1	49	14.03.91	1	17.03.91	1	831.04	.	.	49.19	.	.	79.81	.	.
9	35787	1	49	11.04.91	1	12.04.91	1	.	.	.	.	.	.	78.61	.	.
10	35787	1	50	16.02.92	1	20.02.92	1	786.50	.	.	.	41.49	4.96	.	.	.

SAS



	ROOTLPNO	SEX	AGE	ADMDATE	ADMTYPE	SEPDATE	SEPTYPE	DIAG1	DIAG2	DIAG3	DIAG4	DIAG5	DIAG6	PROC1	PROC2	PROC3
1	17802	1	27	07/11/1996	1	07/11/1996	1	.	.	.	.	.	.	63.73	.	.
2	29699	1	25	12/12/1983	1	14/12/1983	1	55.89	.	.	.	.	.	.	.	.
3	29699	1	33	24/10/1991	1	24/10/1991	1	.	.	.	.	.	.	63.73	.	.
4	32901	1	31	30/10/1996	1	30/10/1996	1	.	.	.	.	.	.	63.73	.	.
5	35787	1	39	16/02/1981	1	21/02/1981	1	.	.	.	.	.	.	56.37	.	.
6	35787	1	43	08/02/1985	2	09/02/1985	1	99.53	.	.	.	.	.	.	.	.
7	35787	1	45	24/11/1986	2	25/11/1986	1	87.5	.	.	.	.	.	.	.	.
8	35787	1	49	14/03/1991	1	17/03/1991	1	831.04	.	.	49.19	.	.	79.81	.	.
9	35787	1	49	11/04/1991	1	12/04/1991	1	.	.	.	.	.	.	78.61	.	.
10	35787	1	50	16/02/1992	1	20/02/1992	1	786.5	.	.	.	41.49	4.96	.	.	.

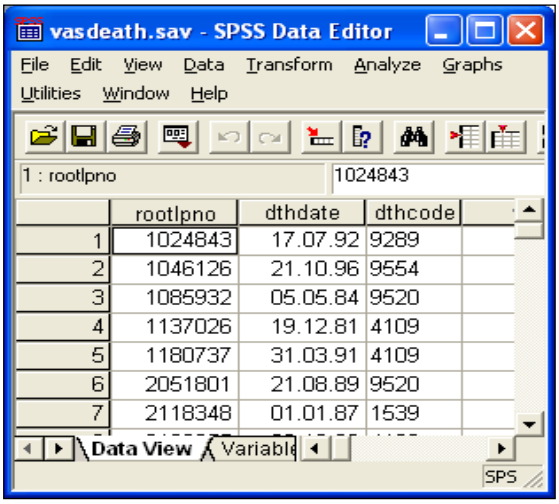
Stata



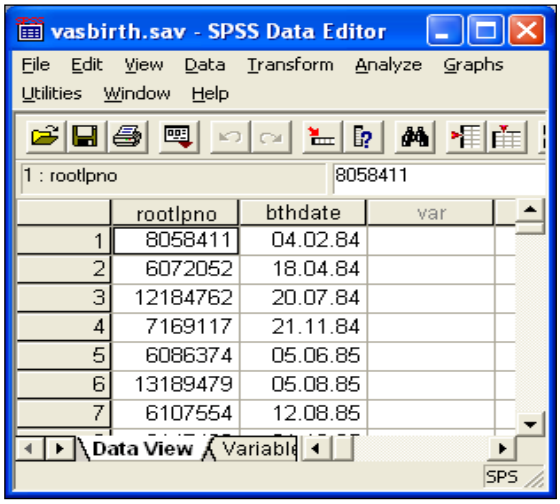
	rootlpno	sex	age	admdate	admtype	sepdate	septype	diag1	diag2
1	17802	male	27	07 Nov 96	elective	07 Nov 96	discharg	.	.
2	29699	male	25	12 Dec 83	elective	14 Dec 83	discharg	55.89	.
3	29699	male	33	24 Oct 91	elective	24 Oct 91	discharg	.	.
4	32901	male	31	30 Oct 96	elective	30 Oct 96	discharg	.	.
5	35787	male	39	16 Feb 81	elective	21 Feb 81	discharg	.	.
6	35787	male	43	08 Feb 85	emergenc	09 Feb 85	discharg	99.53	.
7	35787	male	45	24 Nov 86	emergenc	25 Nov 86	discharg	87.5	.
8	35787	male	49	14 Mar 91	elective	17 Mar 91	discharg	831.04	.
9	35787	male	49	11 Apr 91	elective	12 Apr 91	discharg	.	.
10	35787	male	50	16 Feb 92	elective	20 Feb 92	discharg	786.5	.
11	35787	male	50	17 Mar 92	elective	17 Mar 92	discharg	41.4	.
12	35787	male	50	07 May 92	elective	08 May 92	discharg	41.4	.

2. Separately open the vasdeath and the vasbirth data files, similarly, to inspect them visually and become familiar with their structure and content. These are simple files containing no more than the rootlpno, dates of the relevant events and, in the case of the death file, a cause of death code. You are already familiar with the vascancer2 data file.

SPSS

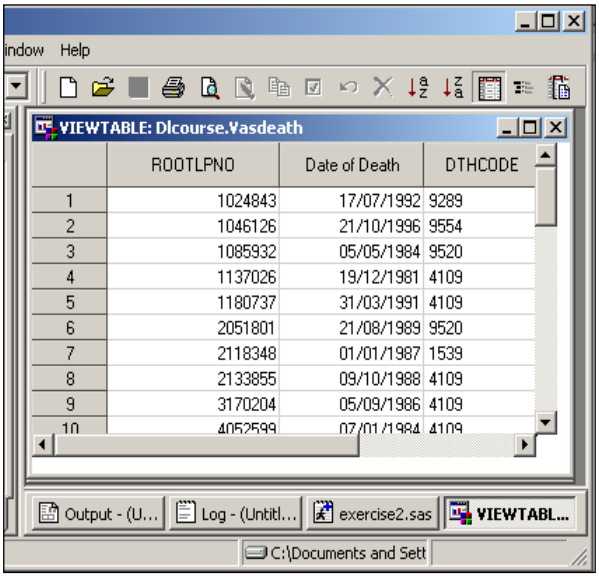


	rootlpno	dthdate	dthcode
1	1024843	17.07.92	9289
2	1046126	21.10.96	9554
3	1085932	05.05.84	9520
4	1137026	19.12.81	4109
5	1180737	31.03.91	4109
6	2051801	21.08.89	9520
7	2118348	01.01.87	1539

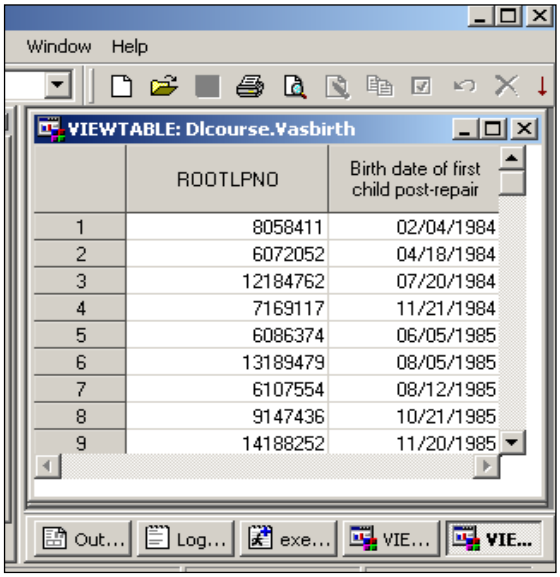


	rootlpno	bthdate	var
1	8058411	04.02.84	
2	6072052	18.04.84	
3	12184762	20.07.84	
4	7169117	21.11.84	
5	6086374	05.06.85	
6	13189479	05.08.85	
7	6107554	12.08.85	

SAS

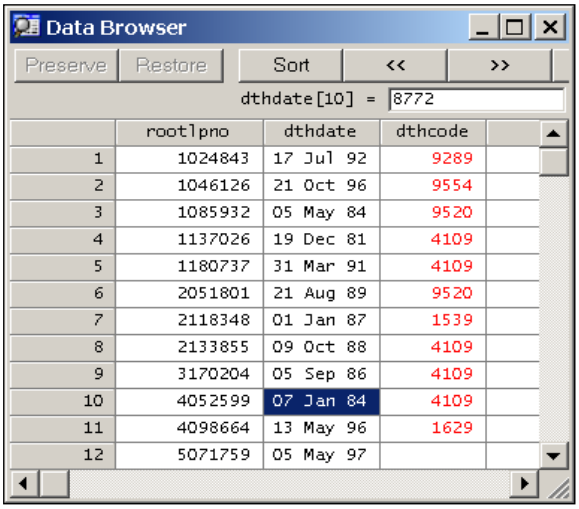


	ROOTLPNO	Date of Death	DTHCODE
1	1024843	17/07/1992	9289
2	1046126	21/10/1996	9554
3	1085932	05/05/1984	9520
4	1137026	19/12/1981	4109
5	1180737	31/03/1991	4109
6	2051801	21/08/1989	9520
7	2118348	01/01/1987	1539
8	2133855	09/10/1988	4109
9	3170204	05/09/1986	4109
10	4052599	07/01/1984	4109

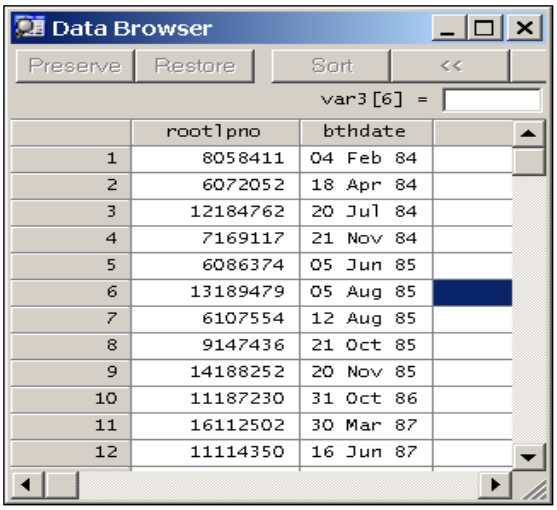


	ROOTLPNO	Birth date of first child post-repair
1	8058411	02/04/1984
2	6072052	04/18/1984
3	12184762	07/20/1984
4	7169117	11/21/1984
5	6086374	06/05/1985
6	13189479	08/05/1985
7	6107554	08/12/1985
8	9147436	10/21/1985
9	14188252	11/20/1985

Stata



	rootlpno	dthdate	dthcode
1	1024843	17 Jul 92	9289
2	1046126	21 Oct 96	9554
3	1085932	05 May 84	9520
4	1137026	19 Dec 81	4109
5	1180737	31 Mar 91	4109
6	2051801	21 Aug 89	9520
7	2118348	01 Jan 87	1539
8	2133855	09 Oct 88	4109
9	3170204	05 Sep 86	4109
10	4052599	07 Jan 84	4109
11	4098664	13 May 96	1629
12	5071759	05 May 97	



	rootlpno	bthdate
1	8058411	04 Feb 84
2	6072052	18 Apr 84
3	12184762	20 Jul 84
4	7169117	21 Nov 84
5	6086374	05 Jun 85
6	13189479	05 Aug 85
7	6107554	12 Aug 85
8	9147436	21 Oct 85
9	14188252	20 Nov 85
10	11187230	31 Oct 86
11	16112502	30 Mar 87
12	11114350	16 Jun 87

3. Use the merging procedure available in your statistical software package to merge the dates of death (dthdate) and causes of death (dthcode) from vasdeath as the external file used as a keyed table with rootlpno as the keying variable. *You should read on to step 4 at this point so you know what is coming next.*

4. Why doesn't step 3 work at first? Remember that for merging to work, two essential conditions must be met: (i) both the working file and the external file must be sorted in order of the keying variable, in this case rootlpno; and (ii) there must be no more than one record per individual, ie, per rootlpno, in the external file. **SPSS and SAS users** should investigate, identify the problem and fix it, so that you can successfully load the death data from vasdeath.sav. For **Stata users** there are two problems that you should identify and fix, so that you can successfully load the death data from the vasdeath file. **All users** should note that for the person with two possible death dates, only one date is feasible given the same person's history of hospital separations. Once you have succeeded in rectifying the problem, save the merged file of hospital morbidity and death data using the file name, vaslinks.

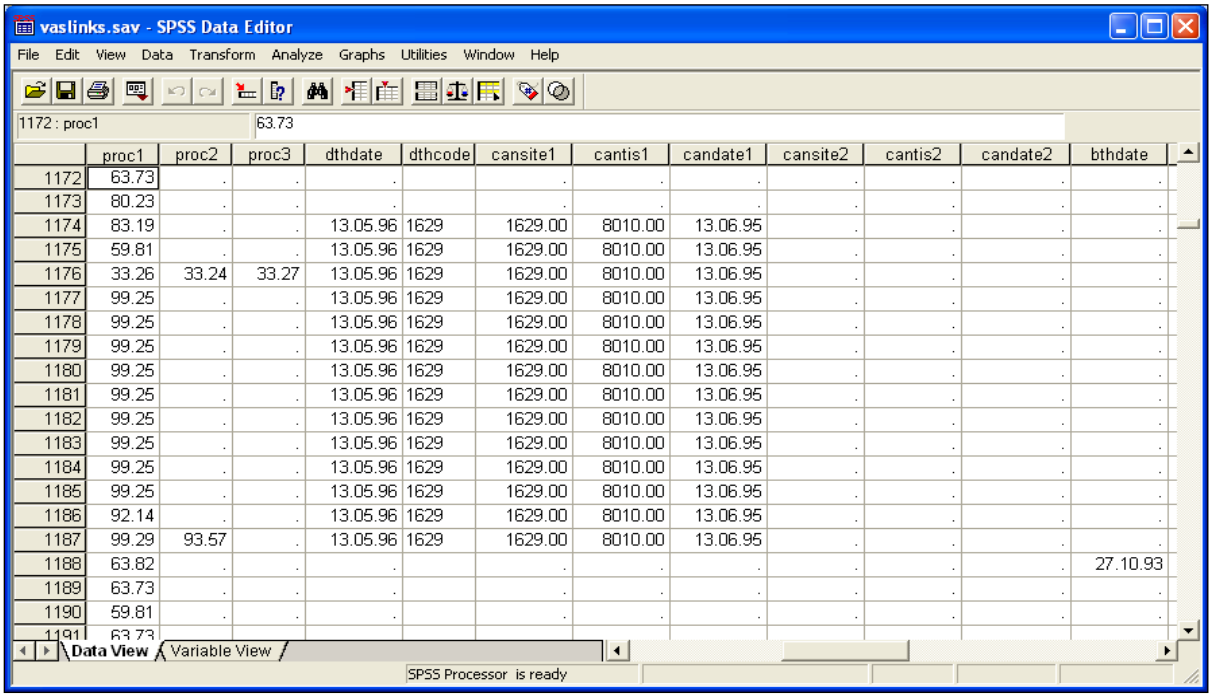
5. Next are the cancer data. Open the vaslinks data file, if not already open. Now repeat the same type of merging procedure as performed in steps 3-4, but this time with vascancer2 as the external file. Load all of the six cancer site, tissue type and date variables onto the EOR loading area of the vaslinks data file. This procedure should run smoothly. Save the file again at this point as vaslinks2. **SAS users:** Your published syntax solution uses different file numbering.

6. Now load the birth registration data from vasbirth onto the EOR area of the vaslinks2 data file using the same merging procedure.

7. Why does step 6 fail the first time? **SPSS and SAS users** should investigate and remedy the reason, which is different from what caused the earlier problem with the deaths. **Stata users** should note that the Step 6 does not necessarily fail the first time as newer versions of Stata are able to cope with merge files not in ascending order. However, processing will be faster if the file is already sorted and, as a general rule, it is good practise to always work with files sorted in the correct/desired order.

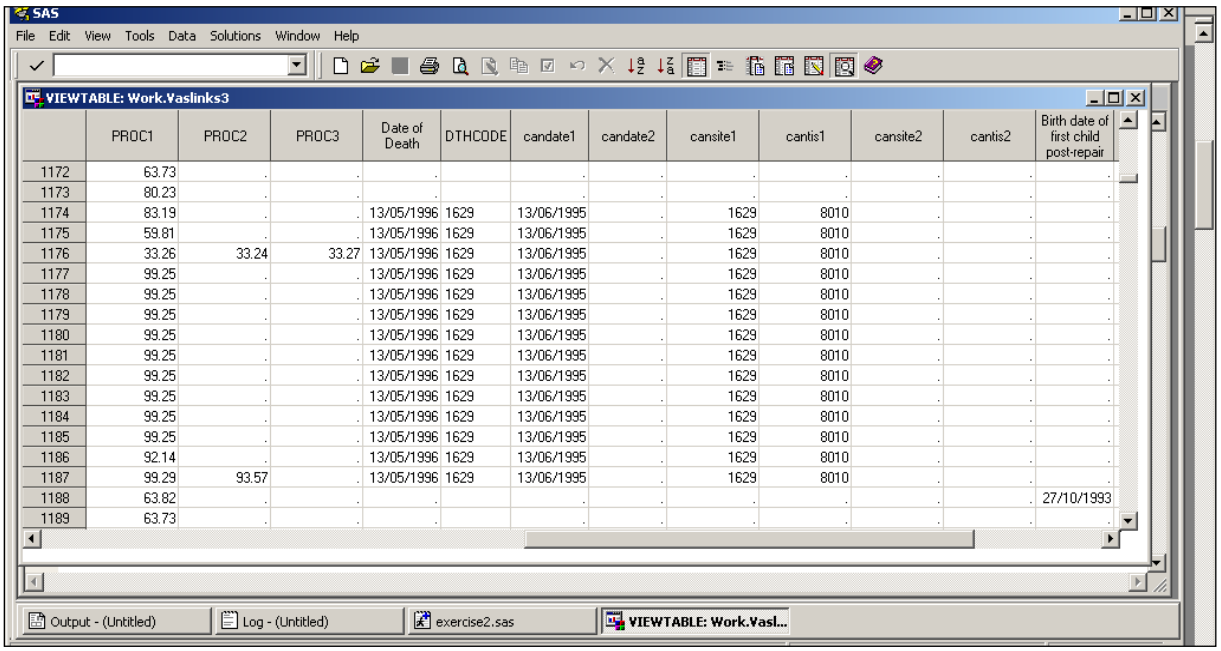
   Once you have succeeded, save the merged file of linked hospital morbidity, death, cancer and birth data as vaslinks3. The EOR loading area of your file should look like this:

SPSS



	proc1	proc2	proc3	dthdate	dthcode	cansite1	cantis1	candate1	cansite2	cantis2	candate2	bthdate
1172	63.73	.	.	.		.	.	.	.	.	.	.
1173	80.23	.	.	.		.	.	.	.	.	.	.
1174	83.19	.	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1175	59.81	.	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1176	33.26	33.24	33.27	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1177	99.25	.	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1178	99.25	.	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1179	99.25	.	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1180	99.25	.	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1181	99.25	.	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1182	99.25	.	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1183	99.25	.	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1184	99.25	.	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1185	99.25	.	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1186	92.14	.	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1187	99.29	93.57	.	13.05.96	1629	1629.00	8010.00	13.06.95	.	.	.	.
1188	63.82	.	.	.		.	.	.	.	.	.	27.10.93
1189	63.73	.	.	.		.	.	.	.	.	.	.
1190	59.81	.	.	.		.	.	.	.	.	.	.
1191	63.73											

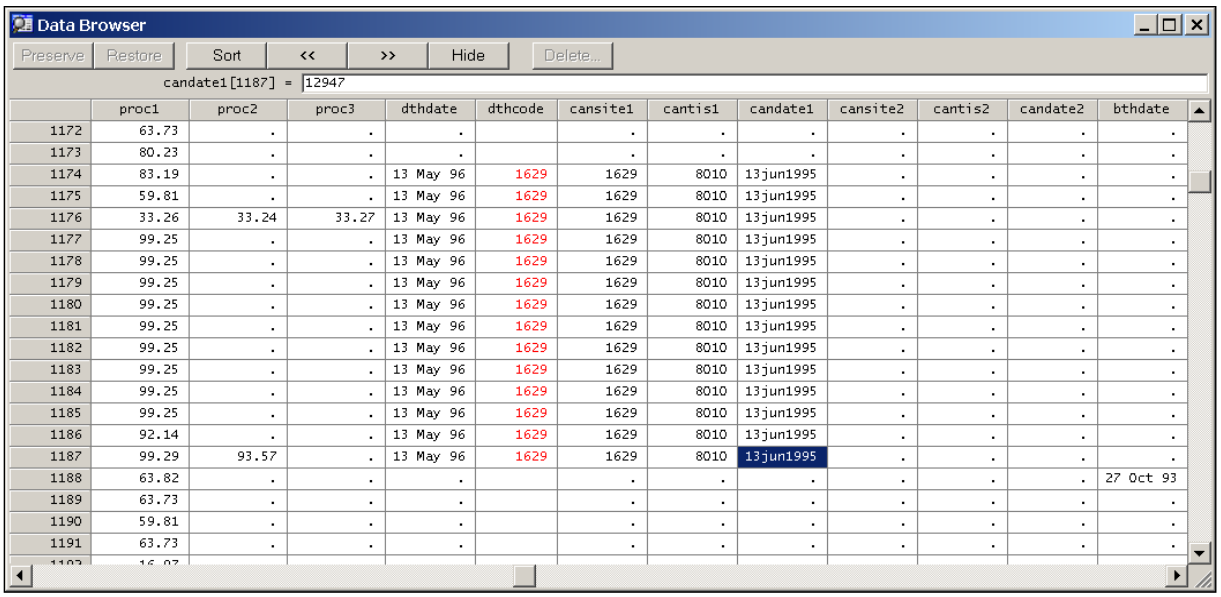
SAS



VIEWTABLE: Work.Vaslinks3

	PROC1	PROC2	PROC3	Date of Death	DTHCODE	candate1	candate2	cansite1	cantis1	cansite2	cantis2	Birth date of first child post-repair
1172	63.73	.	.	.		.	.	.	.	.	.	.
1173	80.23	.	.	.		.	.	.	.	.	.	.
1174	83.19	.	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1175	59.81	.	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1176	33.26	33.24	33.27	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1177	99.25	.	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1178	99.25	.	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1179	99.25	.	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1180	99.25	.	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1181	99.25	.	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1182	99.25	.	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1183	99.25	.	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1184	99.25	.	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1185	99.25	.	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1186	92.14	.	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1187	99.29	93.57	.	13/05/1996	1629	13/06/1995	.	1629	8010	.	.	.
1188	63.82	.	.	.		.	.	.	.	.	.	27/10/1993
1189	63.73	.	.	.		.	.	.	.	.	.	.

Stata



Data Browser

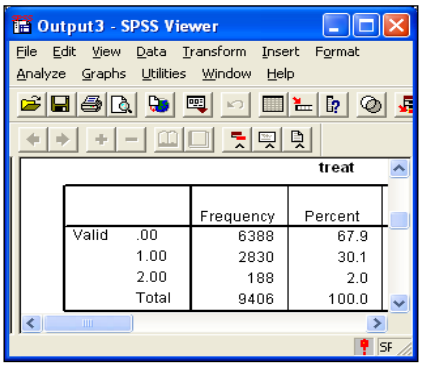
candate1[1187] = 12947

	proc1	proc2	proc3	dthdate	dthcode	cansite1	cantis1	candate1	cansite2	cantis2	candate2	bthdate
1172	63.73	.	.	.		.	.	.	.	.	.	.
1173	80.23	.	.	.		.	.	.	.	.	.	.
1174	83.19	.	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1175	59.81	.	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1176	33.26	33.24	33.27	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1177	99.25	.	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1178	99.25	.	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1179	99.25	.	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1180	99.25	.	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1181	99.25	.	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1182	99.25	.	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1183	99.25	.	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1184	99.25	.	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1185	99.25	.	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1186	92.14	.	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1187	99.29	93.57	.	13 May 96	1629	1629	8010	13jun1995	.	.	.	.
1188	63.82	.	.	.		.	.	.	.	.	.	27 Oct 93
1189	63.73	.	.	.		.	.	.	.	.	.	.
1190	59.81	.	.	.		.	.	.	.	.	.	.
1191	63.73	.	.	.		.	.	.	.	.	.	.

Exercise Instructions Part B: Sequence Variables

8. Now add a year of separation variable, yearsep, derived from sepdate using a date function to extract the year. Tag all records that mention a vasectomy by setting a new tag variable, treat, to '1' (with default = '0'). Also, tag all of the records that mention a vasovasostomy by setting the value of treat to '2'. You will need to search for the target procedure codes in proc1, proc2 and proc3, but you need not be concerned about multiple instances per record of the same code. However, you will need to allow for the change in coding system from 1988 onwards. The frequency distribution of treat is shown opposite.

SPSS

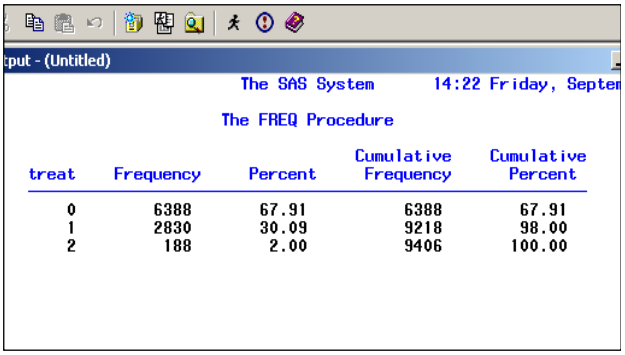


Output3 - SPSS Viewer

treat

		Frequency	Percent
Valid	.00	6388	67.9
	1.00	2830	30.1
	2.00	188	2.0
	Total	9406	100.0

SAS

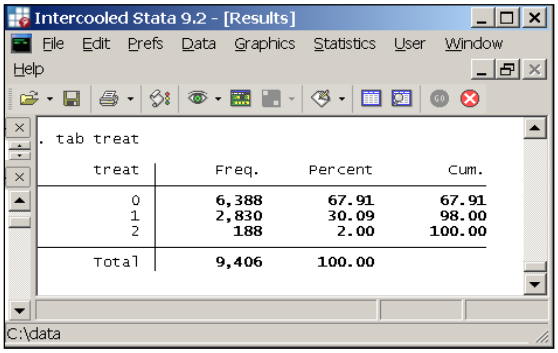


The SAS System 14:22 Friday, Septem

The FREQ Procedure

treat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6388	67.91	6388	67.91
1	2830	30.09	9218	98.00
2	188	2.00	9406	100.00

Stata

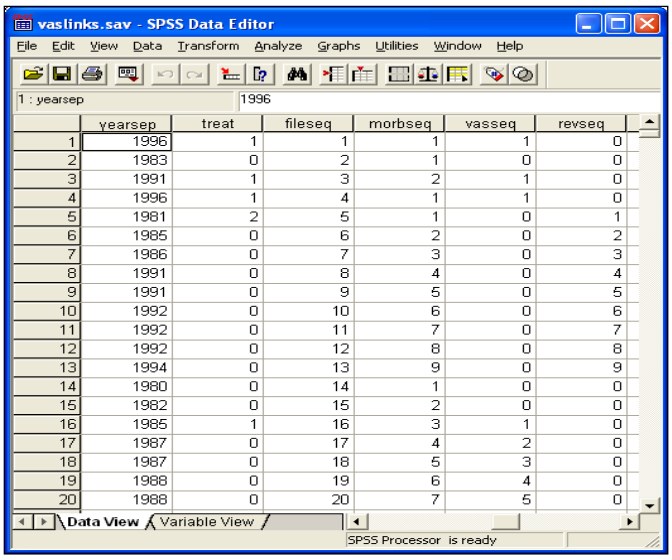


. tab treat

treat	Freq.	Percent	Cum.
0	6,388	67.91	67.91
1	2,830	30.09	98.00
2	188	2.00	100.00
Total	9,406	100.00	

9. Create a fileseq variable in the EOR loading area, numbering all records from 1 to 9,406.
10. Create a morbseq variable at the end of each record. You learnt how to do this in exercise 1.
11. Next compose and run syntax that loads an index sequence variable onto the end of each record, which you should name vasseq. It is assigned the value of 1 for the first record in a block for an individual where any code for a vasectomy is mentioned. Subsequent records in the block are numbered with incremented integers; the values for earlier records in the block are numbered 0.
12. Do the same as in step 11, but this time regard the index record as the first one to mention any code for vasovasostomy (ie vasectomy reversal), and call the index sequence variable revseq. Your file should now look something like that shown opposite.
13. Now add a dead variable to the file, which is set to '0' for records pertaining to living individuals and '1' for those pertaining to dead individuals.
14. Save your files as vaslinks4 with the inclusion of new variables in the EOR loading area.

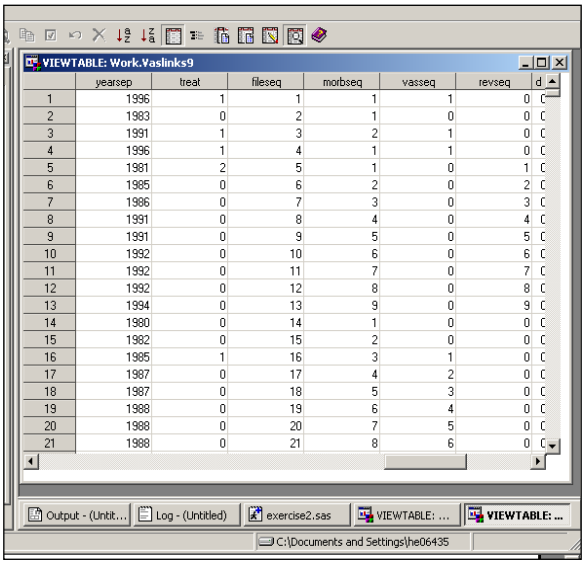
SPSS



vaslinks.sav - SPSS Data Editor

	yearsep	treat	fileseq	morbseq	vasseq	revseq
1	1996	1	1	1	1	0
2	1983	0	2	1	0	0
3	1991	1	3	2	1	0
4	1996	1	4	1	1	0
5	1981	2	5	1	0	1
6	1985	0	6	2	0	2
7	1986	0	7	3	0	3
8	1991	0	8	4	0	4
9	1991	0	9	5	0	5
10	1992	0	10	6	0	6
11	1992	0	11	7	0	7
12	1992	0	12	8	0	8
13	1994	0	13	9	0	9
14	1980	0	14	1	0	0
15	1982	0	15	2	0	0
16	1985	1	16	3	1	0
17	1987	0	17	4	2	0
18	1987	0	18	5	3	0
19	1988	0	19	6	4	0
20	1988	0	20	7	5	0

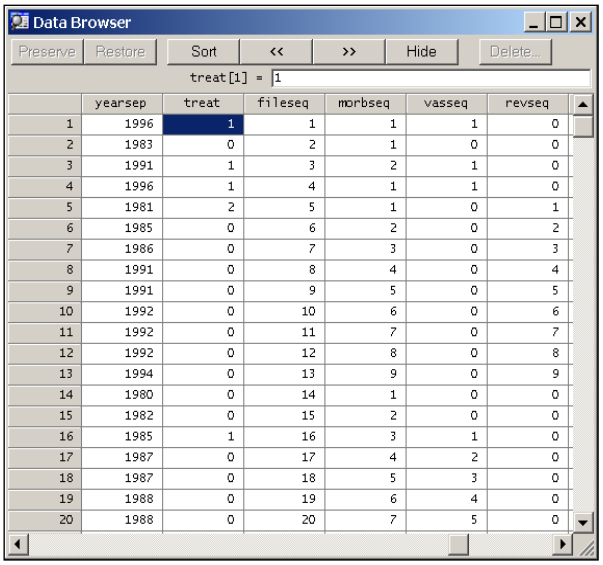
SAS



VIEWTABLE: Work.Vaslinks9

	yearsep	treat	fileseq	morbseq	vasseq	revseq
1	1996	1	1	1	1	0
2	1983	0	2	1	0	0
3	1991	1	3	2	1	0
4	1996	1	4	1	1	0
5	1981	2	5	1	0	1
6	1985	0	6	2	0	2
7	1986	0	7	3	0	3
8	1991	0	8	4	0	4
9	1991	0	9	5	0	5
10	1992	0	10	6	0	6
11	1992	0	11	7	0	7
12	1992	0	12	8	0	8
13	1994	0	13	9	0	9
14	1980	0	14	1	0	0
15	1982	0	15	2	0	0
16	1985	1	16	3	1	0
17	1987	0	17	4	2	0
18	1987	0	18	5	3	0
19	1988	0	19	6	4	0
20	1988	0	20	7	5	0
21	1988	0	21	8	6	0

Stata



Data Browser

treat[1] = 1

	yearsep	treat	fileseq	morbseq	vasseq	revseq
1	1996	1	1	1	1	0
2	1983	0	2	1	0	0
3	1991	1	3	2	1	0
4	1996	1	4	1	1	0
5	1981	2	5	1	0	1
6	1985	0	6	2	0	2
7	1986	0	7	3	0	3
8	1991	0	8	4	0	4
9	1991	0	9	5	0	5
10	1992	0	10	6	0	6
11	1992	0	11	7	0	7
12	1992	0	12	8	0	8
13	1994	0	13	9	0	9
14	1980	0	14	1	0	0
15	1982	0	15	2	0	0
16	1985	1	16	3	1	0
17	1987	0	17	4	2	0
18	1987	0	18	5	3	0
19	1988	0	19	6	4	0
20	1988	0	20	7	5	0

15. Use your linked file to characterise the population by completing the missing entries in the table below. **Hint:** Obtaining the correct numbers of 1st-time vasovasostomies before and after 1st–time vasectomy is not as straightforward as it at first appears. Draw a diagram of a file with two blocks of records, one with a vasovasostomy occurring before a vasectomy and the other with the opposite order. Show what the vasseq and revseq variable values would look like in each block. This should help you understand what syntax is required.
16. Finally, save your syntax.

Characteristics of men undergoing first-time inpatient vasectomy and vasovastomy in Western Australia in 1980-1996 based on an analysis of fictitious linked health data

Characteristic	Men who underwent vasectomy (N =)	Men who underwent vasovasostomy (N =)
Mean age in years ± SD		
Number dead (%)		
Number with a 1st cancer registration (%)		
Number with a 2nd cancer registration (%)		
Number of 1st-time vasovasostomies: Before 1st-time vasectomy (%) After 1st-time vasectomy (%)		
Number of paternity registrations (%)		

Reference

Holman CDJ, Wisniewski ZS, Semmens IL, Rouse IL, Bass AJ. **Population-based outcomes after 28246 in-hospital vasectomies and 1902 vasovasostomies in Western Australia.** *BJU International* 2000; 86: 1043-1049.

REVIEW 2:
'CHARACTERISING POPULATIONS USING LINKED DATA'

Please check that you are familiar with the following topics:

- The definition of data linkage ☐
- The practical steps of data linkage ☐
- Theory of data matching ☐
- Evaluation of the performance of data matching ☐
- Storage, updating and retrieval of links ☐
- Legal frameworks for patients' and other interests in health information ☐
- Measures to enhance anonymity in data linkage research ☐
- Community participation in data linkage research ☐
- Types of administrative data used in linked data in research ☐
- Limitations of administrative health data ☐

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

- Phonetic compression and file explosion ☐
- Address parsing and geocoding ☐
- Unique identifier, partial identifier and cypher ☐
- Deterministic matching, unique matching, fuzzy matching ☐
- Probabilistic matching with and without clerical review ☐
- Equivalencing and file explosion ☐
- Missed matches and mismatches ☐
- Match weights and match weight distribution ☐
- Match cutoffs and manual (clerical) checking ☐
- Blocking and multiple blocking ☐
- Links, chain of links and master linkage key ☐
- Genealogical links ☐
- Privacy-sensitive linkage protocol ☐
- ICD-9, ICPM, ICD-9-CM and ICD-10-AM ☐

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

- Merging files using a linkage key, including troubleshooting ☐
- Tagging records mentioning a target procedure ☐
- Creation of sequence variables ☐
- Selection of index records and use of linked data to characterise a population ☐

3. THEORY OF MEASUREMENT OF HEALTH EVENTS AND STATES

3.1 What You Must Know

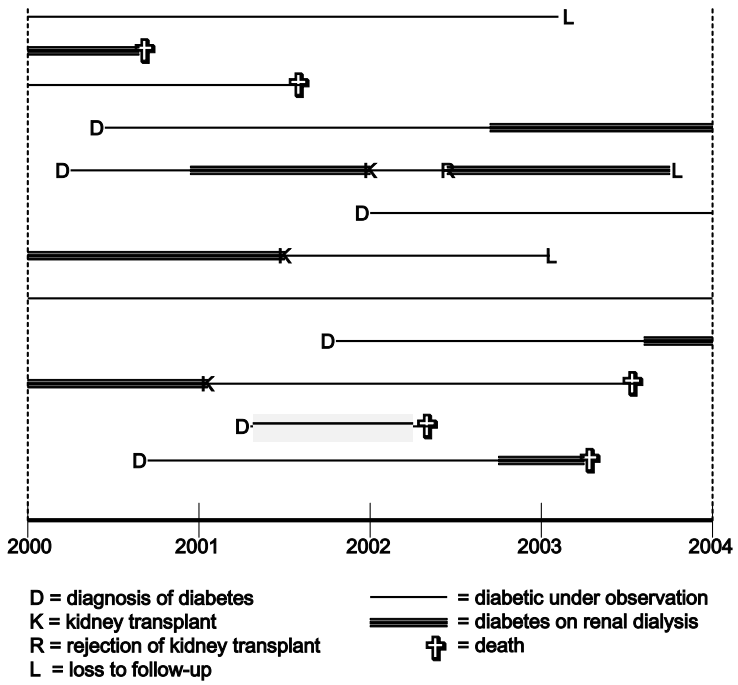
A number of theoretical concepts central to the discipline of epidemiology are used to construct measures of health services resources, utilisation, outcomes and adverse events, especially when they relate to populations in the general community or populations of patients needing or receiving services. Ideally, the researcher who analyses linked health data will have mastered the principles and methods of the discipline of epidemiology, found in a comprehensive text, such as Rothman and Greenland's *Modern Epidemiology*. Failing that, there is a progression of basic epidemiological and biostatistical concepts that you *must* know to be a competent research analyst who uses linked health data. These are summarised as follows:

Basic measures
↓
Effects measures
↓
Estimation from 2x2 tables
↓
Confounding and effect modification in stratified 2x2 tables
↓
Implementation of concepts arising from the above in regression models

Here we deal with basic measures, effect measures and their estimation in 2x2 tables.

3.2 Illustration of the Importance of Measurement Theory

Consider the diagram below, which shows the time traces of 12 patients with severe diabetes, each at high risk of renal failure. The symbols and legend illustrate that several different events may happen, which can be ascertained from linked health data available from the four years from 1 Jan 2000 to 1 Jan 2004: diagnosis of diabetes; commencement of renal dialysis (a marker of chronic renal failure); renal transplant; transplant rejection; loss to follow-up and death.



A clinical collaborator might ask the following research questions:

1. What was the risk of a severe diabetic developing renal failure within one year of diagnosis?
2. What was the rate at which diabetics on renal dialysis received kidney transplants?
3. What prevalence of severe diabetics were on a dialysis program as at 30 June 2002?
4. How did that prevalence compare with the same measure a year earlier at 30 June 2003?
5. What was the mortality rate ratio in those having renal failure compared with those without?

It would be difficult for a research analyst to answer these questions correctly from studying the diagram, let alone to prepare the appropriate syntax to analyse the linked data, unless the analyst had a sound grasp of the foundations of epidemiologic measurement.

3.3 Basic Measures in Epidemiology

At the most basic level, epidemiology performs two tasks: ***measurement*** and ***comparison***. In an effort to find out whether a community or patient group is healthy or unhealthy, the epidemiologist first measures one or more indicators and then compares the results with another community or patient group. In each instance, an epidemiologic measure has at least two components: a numerator and a denominator. Numerators consist of counts of events or states of occurrence of health-related phenomena; denominators consist of populations in which events or states occur; and numerators are related to denominators through a variety of classes of mathematical expressions.

Events vs States of Health Occurrence

Event: A health *event* is an occurrence of a health or health service phenomenon that occurs at a particular point in time at a particular place to a particular member of a population. Examples include the event of attending a dental clinic, the event of developing acute appendicitis, the event of childbirth, the event of being admitted to hospital for pneumonia, the event of being readmitted for complications and the event of death.

State*:* A health *state* is an occurrence of a health or service phenomenon that is a particular condition or attribute (neither necessarily time-limited nor permanent) observed in an individual. Examples include the state of having diabetes mellitus, the state of being registered in a renal dialysis program, the state of being on a waiting list, and the state of being an inpatient of a long-stay psychiatric institution.

Types of Population

There are two basic types of population in which health-related events and states may be observed:

Cohort population*:* A *cohort population*, also known as a fixed population, is itself defined by one or more events, although not necessarily health-related events. Once the qualifying events have been experienced by a person, membership of the cohort becomes a permanent attribute of that person. An example of a cohort is the population of patients in WA first receiving a diagnosis of bladder cancer in 1980-1999.

Dynamic population*:* A *dynamic population* is defined by one or more states of being. Because membership only lasts while the qualifying states are present, membership may be temporary. An example of a dynamic population is the population of patients in WA with active (uncured) bladder cancer in 1980-1999. Note that all of the patients in the above cohort population also enter this dynamic population. So do all of the prevalent cases of active bladder cancer diagnosed prior to 1980, but still present in the population from 1980. Patients who die or migrate out of WA relinquish their membership of the dynamic population (but not the corresponding cohort) at that point in time.

Cross-section*:* A *cross-section* refers to the state of members of a population (whether cohort or dynamic) observed at one point in time, which can be either calendar time, age or follow-up time.

Classes of Mathematical Expression Relevant to Epidemiology

***Ratio**:* A *ratio* is a general class of mathematical expression that includes but is not restricted to a proportion or rate. A ratio is obtained by dividing one quantity by another without implying any specific relationship between the numerator and the denominator. Ratios may or may not have specific dimensions. For example, the ratio of 3.5 hospital beds per 1,000 population has specific dimensions; whereas a ratio of 2.3 in the risk of repeat prostatectomy following an initial closed vs an open procedure has no dimensions. The value of a ratio can range from minus infinity to plus infinity.

Proportion**:* A *proportion* is a type of ratio in which those included in the numerator must also be included in the denominator; ie, the numerator is a subset of the denominator. Proportions are often expressed as a percentage. They do not have dimensions and are constrained between 0 and 1 (0 to 100%). Two important proportions are prevalence and cumulative incidence. ***Prevalence is the proportion of members of a cohort or dynamic population in a given state at a specified point in time. ***Cumulative incidence*** is the proportion of members of a cohort at risk who experience a given event within a specified period of follow-up time. The cumulative incidence of death is known as ***cumulative mortality***.

Rate**:* A *rate*, defined strictly, is a ratio of the amount of change in one quantity relative to another. It implies a distinct relationship between numerator and the denominator, with often a specified time dimension in the denominator. The most important rates used in epidemiologic measurement have ***person-time in the denominator, representing the amount of population experience that is under observation; and the numerator represents the number of events arising in the population at risk.

Both the number of people (P) and the length of time that they are under observation (T) contribute to the amount of person-time (PT) that is observed. The most frequently used unit of person-time is the person-year. 100 PYs may arise from observing 100 people for one year or 50 people for two years, or 200 people for six months, and so on.

A rate in which the numerator is new cases of a disease or injury and the denominator is the population at risk expressed as person-time is known as an ***incidence rate***. The incidence rate of death in a general population is known as a ***mortality rate***, whereas the incidence rate of death in a diseased population is known as a ***case fatality rate***. The incidence rate of inpatient hospitalisation is known as a ***hospital morbidity rate***.

Often measures that are proportions or ratios are incorrectly called rates, albeit that this point is not without controversy. Some examples are survival rate (a proportion); response rate (a proportion); and infant mortality rate (strictly a ratio if calculated in the routine manner of dividing infant deaths by live births in any one year). Purists do not use the word 'rate' when referring to these measures.

The following table shows examples of ratios, proportions and rates used in health services research.

	Ratios	Proportions	Rates
Inputs/resources	Manpower/population ratio Bed/population ratio Per capita health expenditure Depreciated/replacement value	Coverage by insurance Tertiary bed fraction Proportion of service providers accredited	Claim rate
Tasks/services	Units of blood per procedure Attendances per session	Coverage by service Lifetime risk of service Same day fraction Occupancy	Attendance rate First admission rate Separation rate Procedure rate
Performance /outcomes	Cost per procedure Cost per casemix adjusted separation Cost per QALY Relative procedure rate	Unplanned return to theatre Patient satisfaction Readmission at 30d Survival at 5yr Avoidable deaths	Recurrence rate Revisional procedure rate Case fatality rate Mortality rate

Risk

The concept of risk in epidemiology is closely related to a proportion, and especially the proportion known as cumulative incidence. The difference is that risk is a probability measure that applies to individuals rather than a descriptive measure of what is observed in a whole population.

Incidence and Prevalence Measures

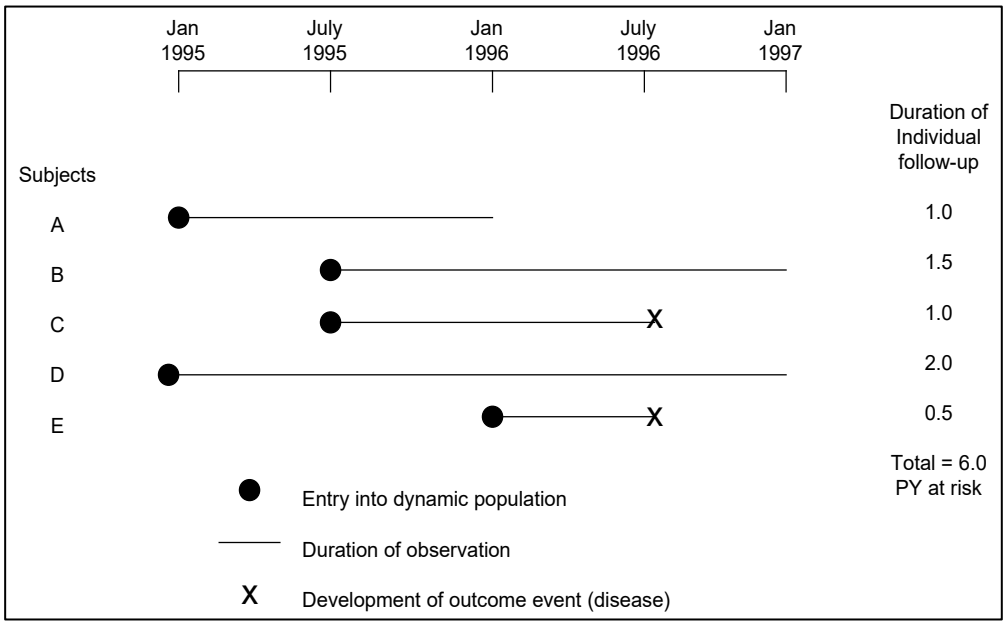
Most frequently, epidemiologic measures of states and events take the form of prevalence (for states) or incidence measures (for events). In addition, there are two types of incidence measure.

Prevalence (P = p/N) is an estimate of the proportion of individuals in the population with a given health state at a particular point in time (the ***point prevalence***) or at some point during a particular period of time (***period prevalence***).

Cumulative incidence (CI_t = a/N) is the proportion of people in a population at risk who have developed a particular health event at least once during a specific time period. The follow-up begins at a point in time known as ***time-zero*** or t_0. Cumulative incidence assumes that the entire population at risk has been followed from t_0 for the entire follow-up period. However, this may not be the case due to deaths, drop-outs or late entry of the subject relative to the end of the study period. A CI measure that ignores the problem of loss to follow-up, and is therefore potentially an underestimate of the true risk, is known as an ***unconditional cumulative incidence***. A CI measure that is adjusted for the problem of loss to follow-up, by modelling what would have been expected if follow-up had been complete, is known as a ***conditional cumulative incidence***.

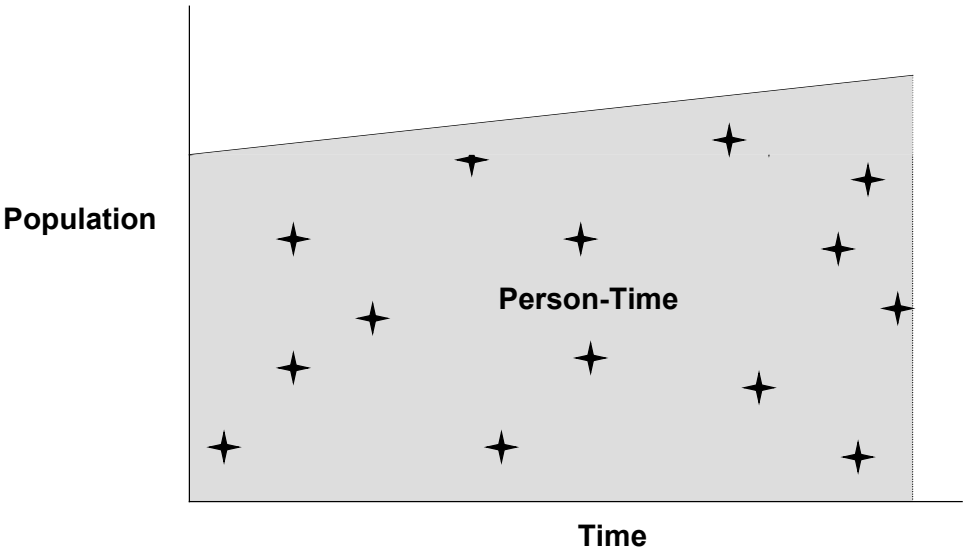
A measure closely related to CI is ***survival*** (S). Survival is also a proportion and is the unitary complement of a CI, most commonly the cumulative mortality; ie, $S_t = 1 - CM_t$. Clinical researchers often refer to the percentage survival from a disease at five or ten years after diagnosis or treatment.

Incidence rate (IR = a/PT) is a true rate and refers to the rate of development of disease or some other event in a population over time. For example, if a researcher observes 10 cases of disease and calculates the person-time at risk in the observed population to be 1,000 person-years, then the incidence rate is 0.01 case per PY, or 1 case per 100PY, or 100 cases per 10,000PY, etc. The figure below illustrates the calculation of person-time at risk based on a hypothetical example of five subjects observed during a follow-up period of two years, in which the incidence rate is 2 cases per 6 PY or 33.3/100PY.



Sometimes, incidence rate implies ***instantaneous incidence rate*** (= ***force of morbidity/mortality*** = ***hazard***), which is the theoretical limit of an incidence rate at a point in time.

The term, ***incidence density***, according to the IEA *Dictionary of Epidemiology* is "the average person-time incidence rate" during a period of observation, although its original and alternative use is to mean the instantaneous incidence rate. Taking the IEA definition, cases of disease (or other outcomes) can be viewed as occurring densely or sparsely in an area (or 'sea') of person-time, as depicted in this illustration:



The concept of incidence density is helpful for estimating person-time at risk. If the population is in ***steady-state*** with respect either to a constant overall size of population or a steady linear growth or decline in population, a cross-sectional census at the mid-point of the study period may be multiplied by the length of time the population is observed to obtain an estimate of the person-time at risk. The following table shows which classes of epidemiologic measures can be estimated in which types of study population.

	Proportions	Rates
Cohort population	Cumulative incidence $CI_t = a/N$	Incidence rate IR = a/PT
Dynamic population		Incidence rate IR = a/PT
Cross-section	Prevalence P = p/N	

3.4 Relationships Between Basic Measures

Cumulative incidence (CI) is dependent on incidence rate (IR) and length of follow-up time (t); ie, the proportion of individuals affected at least once increases if either (i) the rate of development increases or (ii) the time under observation increases; which makes intuitive sense. For a small cumulative incidence (<0.1) the relationship is well approximated by CI = IR.t. However, for larger CI's the following more accurate formula is required: $CI = 1 - e^{-IR.t}$.

Prevalence (P) is dependent on incidence rate (IR) and average duration of condition (D); ie, the proportion of individuals affected at a point in time increases if either the rate of development increases or the duration of the condition increases, which makes intuitive sense. For a small prevalence (<0.05) the relationship is well approximated by P = IR.D. However, for larger P's the following more accurate formula is required P/(1-P) = IR.D or P = IR.D/(1+IR.D).

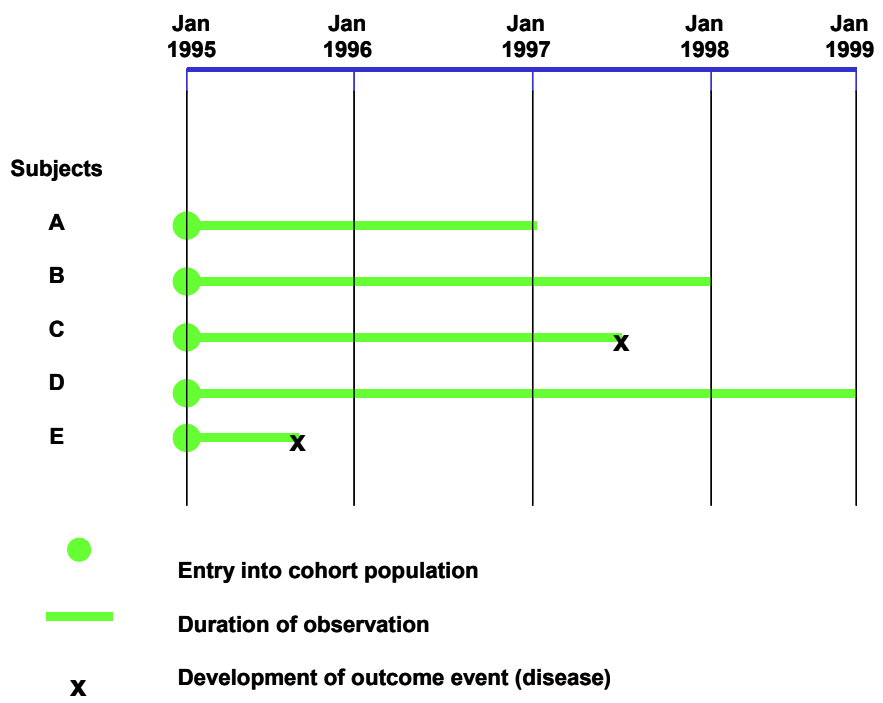
3.5 The Life Table, Actuarial Survival and Relative Survival

If a cohort population is followed over a period of time, whether measured as age in years or time elapsed since a particular health event, such as a diagnosis or procedure, it is usually the case that different subjects are followed for different lengths of time and some are lost to follow-up before the study ends. Thus the unconditional cumulative incidence of the outcome event, $CI_t = a/N$, will be a biased estimate of the true cumulative incidence. It is also often the case that the incidence rate of the outcome event is unsteady over time. For example, the general mortality rate increases with age; and the incidence rate of an acute adverse event may diminish with time elapsed since a procedure was performed. Under these circumstances, the formula for the conditional cumulative incidence, $CI = 1 - e^{-IR.t}$, is also invalid because IR is not constant.

Both problems are solved by dividing the follow-up time into a series of shorter intervals and studying the development of the outcome event during these shorter intervals in those who are still at risk at the commencement of each interval. This way, subjects lost to follow-up by the time a particular interval is reached do not affect the results in the next interval, and because the time intervals are short, the IR in each interval is more likely to be relatively stable. If $S_x = 1 - CI_x$ is the probability of surviving the outcome event during the xth time interval, and these probabilities across different time intervals are independent, then the cumulative survival to time t, passing through all of the time intervals from t_o up to t, is given by:

$$S_t = \Pi S_x, \text{ (ie } S_1 \times S_2 \times S_3 \times S_4 \text{ and so on)}$$

This formulation is the basis for the ***life table*** (or actuarial) method of survival analysis.



a = 2

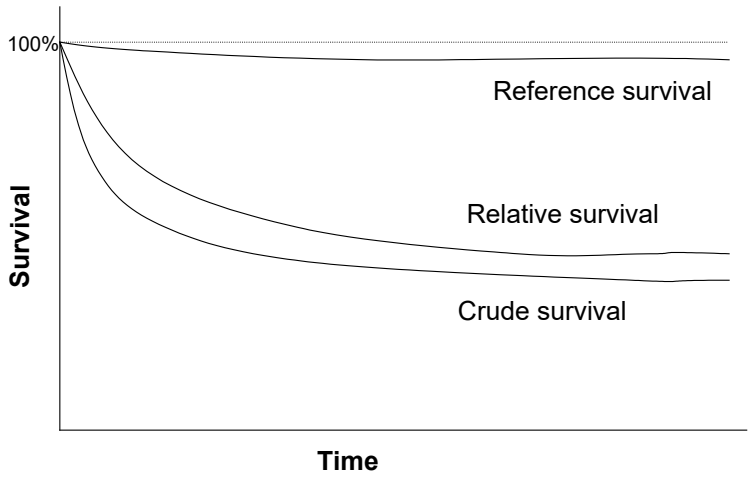
N = 5

Unconditional
$CI_{t=4}$ = 2/5 = 40%
and $S_{t=4}$ = 60%

IN FACT,
$S_{x=1}$ = 0.80
$S_{x=2}$ = 1.00
$S_{x=3}$ = 0.67
$S_{x=4}$ = 1.00

Conditional
$S_{t=4}$ = 0.8 × 0.67
= 54%
$CI_{t=4}$ = 46%

Relative survival is the ratio between a simple survival measure in a study population (known as ***crude survival*** in this context) and the corresponding survival measure in a reference population such as a theoretical or real sample from the general community with a similar demographic profile (sex, age, calendar time) to the study population. Because the reference population survival at any point in time must be ≤ 1.00, it follows that relative survival must be greater than or equal to the corresponding crude survival measure. This is shown graphically below:



In fact, relative survival is an example of an ***effect measure***.

3.6 Effect Measures (Measures of Association) in Epidemiology

Effect measures combine into a single expression a comparison between basic measures on two different populations being compared. This can be achieved in either of two ways: the difference between the basic measures, known as an absolute effect measure, and the ratio between the basic measures, known as a relative effect measure. Examples are as follows:

Absolute Effect Measures

Incidence rate different	$IRD = IR_1 - IR_0$
Cumulative incidence difference	$CID = CI_1 - CI_0$
Prevalence difference	$PD = P_1 - P_0$

Relative Effect Measures

Incidence rate ratio	$IRR = IR_1/IR_0$
Cumulative incidence ratio	$CIR = CI_1/CI_0$
Prevalence ratio	$PR = P_1/P_0$

Social epidemiologists are often interested in comparisons of utilisation measures of health care across different population subgroups defined by sociodemographic factors such as race, social disadvantage, remoteness, private health cover, gender, age and calendar time. Effect measures can be constructed and used as ***equity measures*** for this purpose.

Choose of absolute vs relative effect measures: Some study designs, notable case-control and case-distribution studies, area capable only of estimating relative effect measures; whereas absolute effect measures provide a more direct indication of the public health importance of and effect.

Choose of incidence rate vs cumulative incidence effect measures: As a general guide, incidence rate effect measures are more suitable for occurrence relationships with a long period of elevated risk after exposure; whereas cumulative incidence effect measures are better suited to occurrence relationships where elevated risk is concentrated within a short period after exposure. The reason for this concerns which of the two types of effect measure are most stable over time under different scenarios of chronic and acute causal effects.

3.7 Estimating Effect Measures from 2×2 Tables

The calculation of these effect measures, or at least their conceptualisation, is assisted by the use of 2×2 contingency tables, the interpretation of which differs according to the type of data, which in turn is dependent on the type of study.

Count Data

	E	$\bar{E}$	Total
D	a	b	m_1
$\bar{D}$	c	d	m_0
Total	N_1	N_0	N

CID or PD = $a/N_1 - b/N_0$ CIR or PR = $a/N_1 \div b/N_0$

Odds in the exposed $a/N_1 \div c/N_1 = a/c$
Odds in the non-exposed $b/N_0 \div d/N_0 = b/d$

Odds ratio, OR $= a/c \div b/d = ad \div bc$

Person-Time Data

	E	$\bar{E}$	Total
D	a	b	m
PT at risk	PT_1	PT_0	PT

IRD = $a/PT_1 - b/PT_0$ IRR = $a/PT_1 \div b/PT_0$

3.8 Sources of Error in Epidemiologic Measurement

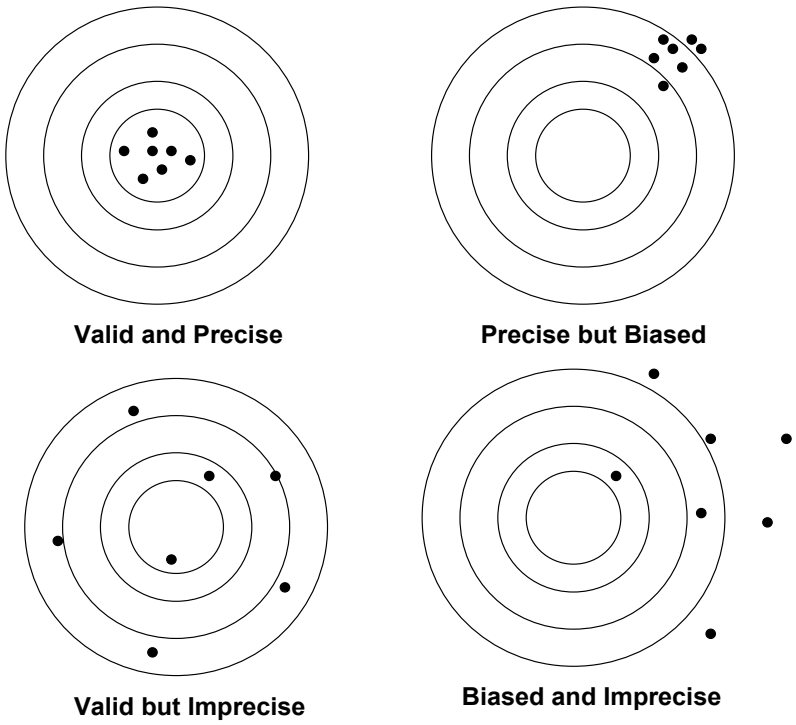
All epidemiological studies, from the calculation of a simple incidence rate to complex research designs such as case-control studies, aim to estimate the true value of an epidemiological measure or effect measure in a population. This true value is known as the theoretical value, because its actual value is unknown and only exists in theory. The best that can be achieved in the real world is to make an ***empirical estimate*** of the ***theoretical value*** by observing the experience of a sample of the population.

Random Error

Random error is a difference between an empirical estimate of an epidemiological measure and its theoretical value that is due to chance variation. Another term for random error is sampling error because it derives from the laws of probability that govern the selection of a sample from a larger population. The lack of potential for random error is known as ***precision***. A precise estimate is one with a small variance, and is associated with a narrow confidence interval. Precision in epidemiological measurement may be improved in either of two ways: (i) increase the sample size; or (ii) use study design parameters that optimise the distributions of exposure or outcome status.

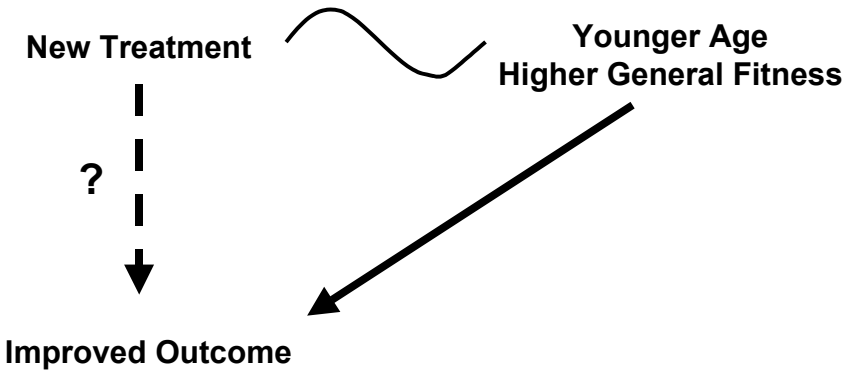
Systematic Error

Systematic error, also known as ***bias***, is present when the theoretical value of an epidemiologic measure cannot be obtained from a study even in the total absence of random error. Bias can be defined as systematic deviation from the truth (theoretical value). The term ***internal validity*** refers to a lack of systematic error; in other words a valid empirical estimate is one that is free from bias. The three common types of systematic error are (i) ***selection bias***; (ii) ***information bias***; and (iii) ***confounding***.



The above diagram helps to distinguish the difference between precision and validity, using an analogy with shots at a marksperson's target.

The form of bias known as confounding is a particular problem in non-experimental studies of health care outcomes, due to the tendency for different clinical decisions, and therefore different treatments, to be administered on the basis of an *a priori* assessment of likely risks and benefits. For example, a new, expensive or relatively more dangerous form of treatment may be administered more often to younger and fitter patients who stand a better chance of survival and a good outcome. This ***confounding by indication*** is depicted below:



3.9 Analysis of Epidemiologic Measures

We will not delve much in this short course into methods of statistical analysis. Suffice to say that there are two general approaches available for the analysis of data, and it is possible to use either one or both of these to summarise the empirical evidence that a study provides. The two approaches are hypothesis testing through a test of significance, and effect measurement through the use of a point estimate and confidence interval.

Industry choice example: To be illustrated in class.

Hypothesis Testing

The procedure for hypothesis testing is as follows:

1. Conceptualise an association.
2. Determine the corresponding null hypothesis (that is, that there is no association).
3. Perform a statistical test to obtain a p-value.
4. If $p<\alpha$ then reject the null hypothesis, and accept the alternative hypothesis that the data support the association, or if $p\geq\alpha$ then on inference can be made.

Effect Measurement

The procedure for effect measurement is as follows:

1. Conceptualise an association.
2. Determine the best effect measure to characterise the strength of association, and the one that can be estimated from available data.
3. Calculate a point estimate and confidence interval for the effect measure.

3.9 More on Measures of Health Care Utilisation

Measures of utilisation of health services are well developed for institutional health care generally and hospital services in particular. Detailed information on terminology and definitions used in the measurement of institutional health care in Australia may be found in the *National Health Data Dictionary* and the *Australian Hospital Statistics* (formerly *Hospital Utilisation and Costs Study*) report series published by the Australian Institute of Health and Welfare. What follows are some of the most important terms and definitions.

Classification of Services from Hospitals by Diagnosis and Procedure

Principal diagnosis *(or* ***principal condition****)*: The condition chiefly responsible for occasioning the admission of the patient to hospital.

Additional diagnoses *(or* ***additional conditions****)*: All conditions, except the principal condition, that arose during the patient's stay, that affected the patient's treatment or length of stay by greater than one day or that existed at the time of admission and for which treatment was given. Conditions having no bearing on the hospital stay should be excluded.

Principal procedure: The procedure which consumed the greatest amount of hospital resources or, if this cannot be determined, that which was the definitive treatment for the principal condition.

Additional procedures: All significant procedures, except the principal procedure, performed on the patient during the episode of care. Significant procedures are surgical in nature, carry a procedural risk, carry an anaesthetic risk, require specialised training or require special facilities or equipment only available in an acute care setting.

External cause: The cause of the principal condition, where that condition is the result of injury, poisoning or violence.

Key Utilisation Concepts and Definitions for Hospital Services

Admission: The process by which an inpatient commences an episode of care. A formal admission is the administrative process whereby the hospital records the commencement of treatment or care and accommodation of a patient. Statistical admissions occur after a patient returns from leave of absence or undergoes a change of type (see separation).

Separation: A health services event that occurs when an inpatient is (1) formally discharged, (2) transferred to another institution, (3) absconds, (4) dies whilst in care, (5) changes status between the categories or nursing home type, rehabilitation, palliation or acute care, or (6) leaves hospital for a period of leave exceeding 7 days (acute hospitals) or 10 days (psychiatric hospitals). Types (1) to (3) are known as formal separations and types (4) and (5) are known are statistical separations.

Length of stay: The number of days obtained by subtracting the admission date from the separation date, excluding any leave days. Same-day patients are deemed to have a length of stay of 1.0 day.

Some Utilisation Concepts and Definitions Used with Linked Hospital Morbidity Data

Index admission/separation/procedure: The admission, separation or procedure for an individual patient that meets a case definition (ie, the particular event that is of interest to the study); eg, an index admission date for an index hospital episode defined by any surgical procedure for urinary stones. Index events are often used to define the beginning of follow-up of a specified patient cohort.

First admission/separation/procedure: The first-time a patient is admitted, separated or undergoes a procedure that meets specified criteria. First admissions can be used as an indication of the incidence rate of certain diseases and serious injuries.

Cumulative admissions/separations/procedures: The total number of admissions, separations or procedures in general, or for a specified condition or of a specified type, occurring within a specified time interval such as one year.

Cumulative length of stay: The sum of all lengths of stay experienced within a specified time interval such as one year. Cumulative LOS may be a general measure or restricted to the sum of LOS for a specific condition or for a specific procedure.

3.10 Health Care Episodes

A health care episode is the period of time during which a specific disease process, illness or health care problem or treatment process is present. It is characterised by an onset and an ending, between which the health problem state applies. The concept of a health care episode has existed for more than 60 years and has achieved increasing attention due to its value as an integrating device to study patterns of care and account for the economic costs of the different components of institutional and ambulatory care that make up the complete treatment of a health problem. There are four types of health care episode, the first three of which pertain to treatment and the fourth to health maintenance:

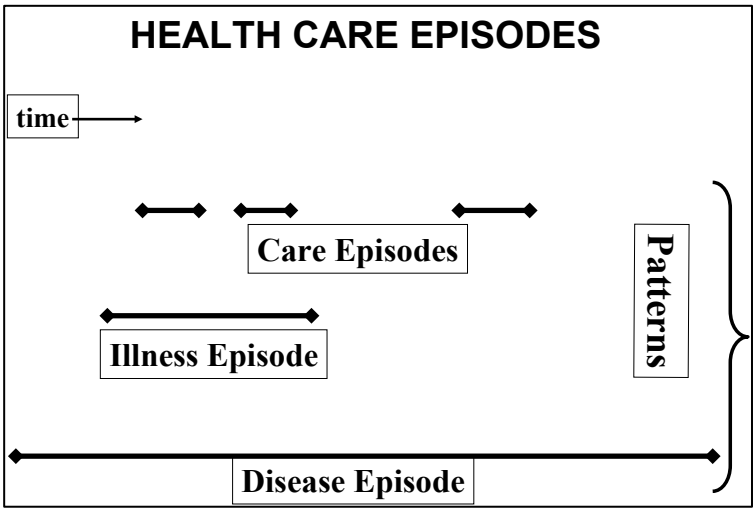
Disease episode: The interval of time during which a given disease process is present, as defined by clinical parameters. This may begin with the onset of subclinical disease and may end with expiration of a specified period of time, return to normal or pre-episode function, cure of the disease or death (including, if the disease is life-long); eg, the life-long disease episode of diabetes mellitus.

Illness episode: The interval of time during which the patient receives a continuous spell of signs or symptoms defined as sickness; eg, the signs and symptoms of acute myocardial infarction.

Care episode: A series of temporally contiguous health care services related to treatment of a given spell of illness; eg, the diagnosis, treatment and aftercare of acute appendicitis.

***Health maintenance episode**:* This type of episode differs from the others. It is an episode of care that relates to preventing disease (eg, immunisation), improving health or appearance (eg, cosmetic procedures), addressing an administrative objective (eg, mandated medical examination), or another non-disease related aim (eg, contraception).

Episodes are characterised by (1) pre-episode risk according to predictive factors, (2) duration, (3) nature of the problem, (4) resource use and (5) outcome.



Returning to the diagram of severe diabetics at risk of renal failure in section 3.2, it will now be apparent that what the diagram represented was a layered series of episodes concerning disease (severe diabetes), illness (chronic renal failure) and health care (eg, renal dialysis and renal transplantation). You should also now be in a position to answer the clinical collaborator's questions, at least conceptually using the diagram. The answers are as follows (the PY estimates are approximations):

1. $CI_{t=1;\ \text{renal failure/diagnosis}} = 1/6 = 0.16$.
2. $IR_{\text{transplant/dialysis}} = 3/6.5PY = 0.46/PY$.
3. $P_{30/6/02} = 1/9 = 0.11$.
4. $P_{30/6/01} = 1/9 = 0.11$;
   $PR = 1.0$.
5. $MR_{\text{dialysis}} = 2/6.5PY = 0.31/PY$;
   $MR_{\text{non-dialysis}} = 3/24PY = 0.12/PY$;
   $MRR = 2/6.5 \div 3/24 = 2.46$.

References

Australian Institute of Health and Welfare. **National Health Data Dictionary – Institutional Health Care. Summary Edition 1993.** Canberra: AGPS, 1993.

Australian Institute of Health and Welfare. **National Health Data Dictionary. Version 7.0.** AIHW Cat. No. HWI 15. Canberra: AIHW, 1998.

Brameld KJ, Holman CDJ, Threlfall TJ, *et al.* **Increasing "active prevalence" of cancer in Western Autralia and its implications for health services.** Submitted.

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'.** Australian and New Zealand Journal of Public Health 1999; 23: 464-467.

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

Cooper-Stanbury M, Solon R, Cook M. **Hospital Utilisation and Costs study 1991-92. Volume 1. A Survey of Public Hospitals and Related Data.** Australian Institute of Health and Welfare: Health Services Series No. 5. Canberra: AGPS, 1994.

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

Elandt-Johnson RC. **Definition of rates: some remarks on their use and misuse.** *Am J Epidemiol* 1975: 102: 267-71.

Gail MH, Benichou J. **Encyclopedia of Epidemiologic Methods.** Chichester: John Wiley & Sons, 2000.

Haggar FA, Preen DB, Pereira G, Holman CDJ, **Einarsdottir K. Cancer incidence and mortality trends in Australian adolescents and young adults, 1982-2007.** *BMC Cancer.* 2012; 12: 151 – 161

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.

Holland WW, Olsen J, Florey CDV (eds). **The Development of Modern Epidemiology.** Oxford University Press, 2007.

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.

Hyndman JCG, Holman CDJ, de Klerk NH. **A comparison of measures of access to child health clinics and the implications for modelling the location of new clinics.** *ANZJPH* 1999; 23: 189-195.

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

Morabia A (ed). **History of Epidemiologic Methods and Concepts.** Basel: Birkhäuser, 2004.

Nedkoff LJ, Briffa TG, Preen DB, Sanfilippo FM, Hung J, Ridout SC, Knuiman M, Hobbs M. **Gender- and Age-Specific Trends in the Incidence of Hospitalized Acute Coronary Syndromes in Western Australia.** *Circulation* 2011; 2011; 4: 557-564.

Olsen J, Saracci R, Trichopoulos D. **Teaching Epidemiology: A Guide for Teachers in Epidemiology, Public Health and Clinical Medicine. Third Edition.** Oxford University Press, 2010.

Porta M (ed). **A Dictionary of Epidemiology. Fifth Edition.** Oxford University Press, 2008.

Rothman KJ. **Epidemiology. An Introduction.** Oxford University Press, 2002.

Rothman KJ, Greenland S, Lash TL. **Modern Epidemiology. Third Edition.** Philadelphia: Lippincott, Williams & Wilkins, 2008.

Shi L. **Health Services Research Methods**. Albany, US: Delmar, 1997.

Snow KJ, Young JT, Preen DB, Lennox NG, Kinner SA. **Incidence, Correlates and Diagnostic Follow Up of Hepatitis C Virus Infection in a Large Cohort of Prisoners Who Have Injected Drugs.** *BMC Public Health* 2014; 14:830 doi 1471-2458/14/830.

TRAINING SESSION 3: MONITORING HEALTH CARE UTILISATION WITH LINKED DATA

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

Training session 3 continues with the use of your linked file of vasectomy and vasovasostomy data. We will be using the file to demonstrate a number of principles concerning the measurement of health care utilisation.

Training Session Resources

In this exercise you will make further use of your vaslinks data file saved from exercise 2:

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

Estimating Incidence Measures from Linked Data

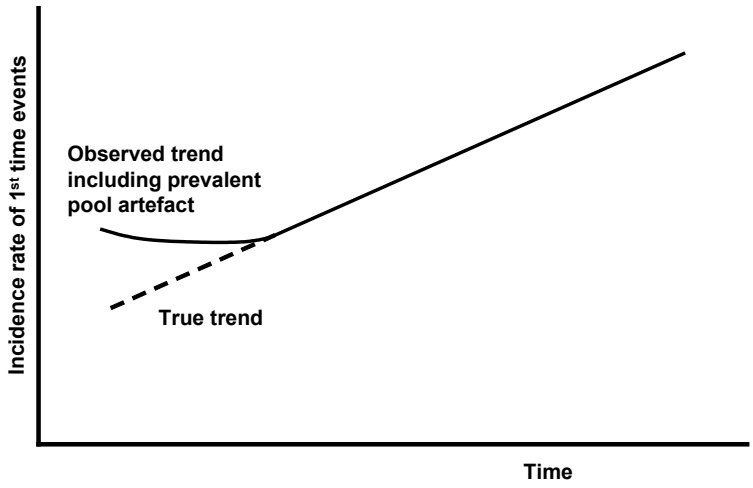
Incidence Measures

Counting incident events, such as disease or service utilisation events, as the numerators of epidemiologic measures using linked data involves the following considerations:

1. Identification of the occurrence of the particular incident event on records in the file and a ***date of occurrence of each incident event***.
2. Whether ***total events*** or only ***first-time events*** are being enumerated. It is possible to construct incidence rate measures using either total or first-time events; but a cumulative incidence measure can be constructed only from first-time events. The choice will usually depend on whether the object of the study is the overall burden of recurrent events borne by the population at risk, or the degree to which those in the population at risk can avoid the occurrence of the event altogether.
3. If only first-time events are to be enumerated, but the event is potentially recurrent, one should consider the problem of the ***prevalent pool effect***.

Prevalent Pool Effect

Linked service utilisation data, such as hospital morbidity data, can be used to estimate the incidence rate of a disease provided that patients with the condition receive the health service (eg, admission to hospital) at least once early in the course of the disease. In the case of incidence rates estimated from first-time hospital admissions, there is likely to be over-estimation of the number of incident cases during the initial years of the data set.



This problem, known as the *prevalent pool effect*, results from the erroneous inclusion of prevalent cases that have had a previous hospital admission prior to commencement of the study observation period. Such cases are then incorrectly enumerated as if they were first-time hospital admissions when in reality they are not. There are three methods of correcting for the prevalent pool effect:

1. Use the primary linked health data in conjunction with a ***linked secondary data source***, which can identify if there is a history of the disease prior to the study period. For example, if linked hospital morbidity data on stroke are available for 1980-2004 and there was a statewide stroke register in place from 1970-1982, which then ceased operation, a linkage to the stroke register could be used to identify and remove from the estimation of incidence trends in 1980-2004 any stroke patients with a registry record prior to 1980.

2. Enforce a ***clearance period*** – essentially, discarding the data from the initial years of observation based on a visual inspection of where the prevalent pool effect appears to end.

3. Apply the Brameld-Holman ***backcasting method***, which implements an actuarial model (in fact, a retrograde survival model) to derive correction factors for the levels of over-ascertainment of first-time cases 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. The method requires advanced computing and statistical skills and is not further considered in this introductory course; however, the abstract and a key figure from the paper are reproduced below.

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

THEORY AND METHODS

Improved methods for estimating incidence from linked hospital morbidity data

Kate J Brameld,[1] C D'Arcy J Holman,[1] David M Lawrence[1,2] and Michael ST Hobbs[1]

Accepted 4 April 2003

Background Linked hospital morbidity data can be used to estimate the incidence of serious chronic disease. However, incidence rates calculated from first-time hospital admissions tend to be overestimated as a result of the erroneous inclusion of prevalent cases that have had previous hospital admissions prior to the study observation period. To address this problem, we have developed the backcasting method.

Method A retrograde survival model was implemented to calculate the level of over-ascertainment of incidence according to the number of years of linked data on which the estimates were based and corresponding correction factors were calculated. The method is illustrated using the example of linked hospital morbidity data on diabetes mellitus and then acute myocardial infarction, which was validated against the Perth MONICA database for cardiovascular disease.

Results Corrected estimates of the incidence of diabetes and acute myocardial infarction were produced. The incidence of diabetes was shown to be lower than in North America in accordance with prevalence estimates, whereas the incidence of acute myocardial infarction was overestimated by approximately 10%.

Conclusion A new method is presented for estimating incidence trends in disease from linked hospital morbidity data. The advantages of this method are its ease of use with routinely collected data and the relatively low cost of applying it in comparison with community surveys or maintaining formal disease registers. The method has other applications using linked data, such as the study of trends in first-time health care procedures and pharmaceutical prescriptions.

Keywords Medical record linkage, incidence, diabetes mellitus, statistical models

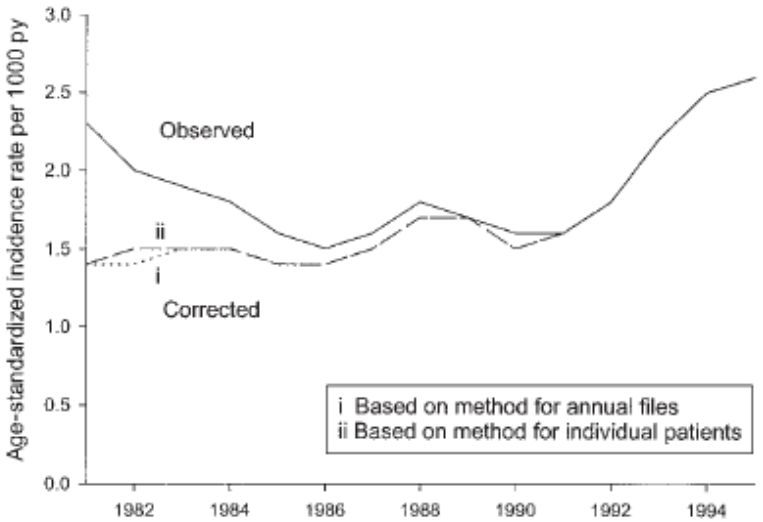
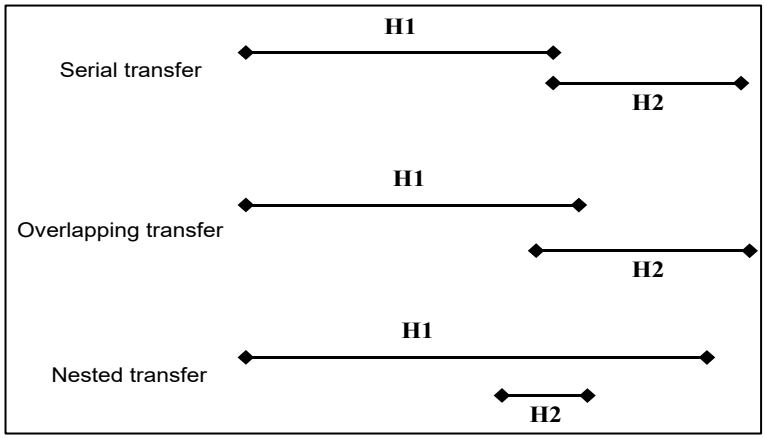


Figure 3 Observed and corrected incidence rate of hospitalized diabetes, Western Australia, 1981–1995

Inter-Hospital Transfers

The occurrence of inter-hospital transfers add considerably to the complexities of analysing linked hospital morbidity data. Transfers may be of several types, the most common of which are shown below:



Other combinations are possible (eg, a *disconnected transfer* with a gap between H1 and H2 due to long-haul travel or an *aligned transfer* where H1 and H2 have identical admission and separation dates) and some health care episodes involving multiple transfers or transfers between three institutions may be very complex. Even the simple serial transfer causes problems, as it may be unclear from the linked data as to whether the second record results from an inter-hospital transfer or a readmission to a different hospital or the same hospital for a new condition or complications that have developed unexpectedly later on the day of discharge.

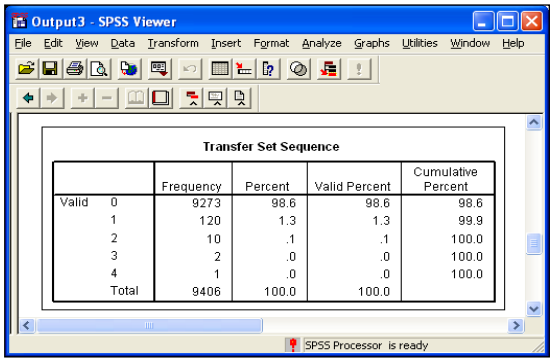
In practice, we define an inpatient record as a transfer record if the admission date falls on any day, except the last day, of another hospital stay for the same patient; or if the admission date falls on the separation date of another hospital stay with a separation type indicating 'inter-hospital transfer.' Even this must be regarded as a working definition as there are exceptions (eg, long distance transfers involving overnight patient transport).

There are several implications of inter-hospital transfers for the analysis of health care utilisation and outcomes. These include:

1. The need to account for transfers when determining the ***length of stay*** of hospital episodes or cumulative LOS over a longer period.
2. The need to account for ***conditions or procedures*** recorded on either the initial or subsequent records of a transfer set.
3. The need to avoid counting transfers as ***readmissions*** and to adjust the ***time-zero*** for commencement of risk of readmission from the last date of separation of a transfer cluster.

A general syntax to account for any number and combination of inter-hospital transfers is beyond the scope of this introductory course. For those who are interested, one effective approach in SPSS is to sort the file by admission date within rootlpno, and use a scratch variable to store the separation date and type of the most 'advanced' separation date known for a block at any one processing point. Any subsequent record found to intersect the most 'advanced' record of separation will be a transfer and can be incrementally numbered using a transfer set sequence variable. This approach can cope with multiple overlapping and non-lapping nested transfers.

The vaslinks file contains 133 serial and overlapping transfers (there are no nested transfers) with the highest order being a 4th transfer (see frequency distribution opposite). Thus an approach based on the following syntax will suffice to flag and number these transfers in a data file such as this, with a preferred sequence that is sorted by separation date within rootlpno. Note that this simplified approach will be insufficient in many other circumstances.



Transfer Set Sequence

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	9273	98.6	98.6	98.6
	1	120	1.3	1.3	99.9
	2	10	.1	.1	100.0
	3	2	.0	.0	100.0
	4	1	.0	.0	100.0
	Total	9406	100.0	100.0	

SPSS

```
COMPUTE transseq = 0.
IF (morbseq GE 2 AND LAG(sepdate) GT admdate) transseq = 1.
IF (morbseq GE 2 AND LAG(septype) EQ 2 AND LAG(sepdate) EQ admdate) transseq = 1.
IF (morbseq GE 2 AND transseq EQ 1 AND LAG(transseq) GE 1) transseq = LAG(transseq) + 1.
EXECUTE.
```

SAS

```
by rootlpno;
retain transseq;
lagsep = lag(sepdate);
lagstype = lag(septype);
if first.rootlpno then transseq = 0;
else do;
  if admdate < lagsep then transseq = transseq + 1;
  else if (admdate = lagsep) and (lagstype = 2) then transseq = transseq + 1;
        else transseq = 0;
end;
run;
```

Stata

```
generate transseq=0
replace transseq=1 if morbseq>=2 & sepdate[_n-1]> admdate
replace transseq=1 if morbseq>=2 & septype[_n-1]==2 & sepdate[_n-1]==admdate
replace transseq=transseq[_n-1]+1 if morbseq>=2 & transseq==1 & transseq[_n-1]>=1
```

R

```
vaslinks4$transseq <- 0
vaslinks4$transseq <- ifelse(vaslinks4$morbseq >= 2 &
                             Lag(vaslinks4$sepdate) > vaslinks4$admdate,
                             1,vaslinks4$transseq)
vaslinks4$transseq <- ifelse(vaslinks4$morbseq >= 2 & Lag(vaslinks4$septype) == 2 &
                             Lag(vaslinks4$sepdate) == vaslinks4$admdate,
                             1,vaslinks4$transseq)
attach(vaslinks4)
for(i in 1:nrow(vaslinks4)){
 if(morbseq[i] >= 2 && transseq[i] == 1 &&
   transseq[i-1] >= 1) vaslinks4$transseq[i] <- vaslinks4$transseq[i-1] + 1
}
 detach(vaslinks4)
```

SPSS, Stata and R users: Your syntax is straightforward. From an assigned default value of 0, transseq is replaced by the value 1 when the first overlapping transfer (2nd line of syntax) or serial transfer (3rd line of syntax) in a block is encountered. The 4th line performs a further replacement of transseq by an incremented value if the previous record in the block was also a transfer.

SAS users: Your syntax is made more complex by the need to ensure that the variables sepdate and septype are examined by the computer for every record in the file so that the lag functions applied to these variables predictably return values from the record immediately before. This is achieved in the 3rd and 4th lines of syntax, where the lagged values are assigned to two temporary file variables, lagsep and lagstype. The purpose of the retain statement in the 2nd line is to carry over a default value of transseq from each record to the next. These carried over values are the starting points from which the syntax in the 7th – 8th lines performs correctly.

Important note for Western Australian analysts: While this course uses septype = 2 to denote an inter-hospital transfer, actual coded values and their meanings in the WA Hospital Morbidity System are subject to change. For example, in 2007, an inter-hospital transfer was coded to 1, 3 or 5. You should therefore check what codes apply to your data set.

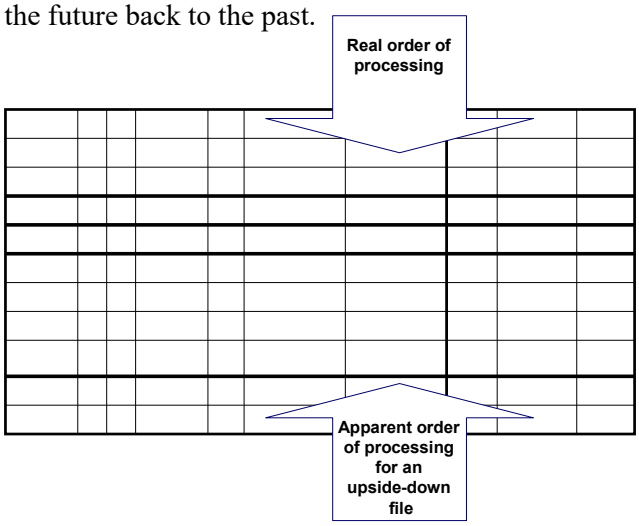
File Processing Order and Information Backflow

Many statistical software packages process a data file moving in one direction from the first to the last record in the file. In the analysis of linked data files, this ***file processing order*** is appropriate when information derived from earlier record(s) pertaining to an individual needs to be carried forward and placed on or made to interact with information on subsequent records. Often, however, our interest is in collecting or deriving outcomes information from the later records pertaining to an individual and loading this information back onto the EOR loading area of the index record. You may have noticed a problem like this in exercise 1, where we loaded information about a first cancer record forward onto a second cancer record; whereas it would have been more useful to have been able to load the information in the opposite direction; ie, from the second cancer record back to the first. A similar need arises in dealing with inter-hospital transfers, where it is usually desirable to load information from H2-H*X* back onto H1.

Different statistical software packages offer a range of different solutions to solving the problem of information backflow, and because these facilities are so important to the analysis of linked health data, we shall consider them separately in SPSS, SAS, Stata and R.

General - upside-down files: This approach can be implemented with virtually any package. Because the file processing order of the software package cannot be reversed, we achieve the same result by turning the sequence of records in the file upside-down; hence an 'upside-down file'. This is achieved by sorting the records by fileseq in descending order. **Useful tip:** Care must be taken when dealing with upside-down files to restore the preferred sequence of the file prior to performing other runs that require the records to be in the original sequence. In general, it is unwise to save a linked data file in an upside-down state or even to leave it unattended, thereby running the risk that you might forget that the file is upside-down. Writing syntax for upside-down files is a skill that requires practice, because epidemiological concepts, involving a longitudinal view that proceeds from the past to the future, must be translated into computing concepts that move from the future back to the past.

Useful tip: Most analysts find it easier to write syntax for upside-down files if they do not try to imagine a reflection of the file around a horizontal axis, but imagine rather that the file stays the ***right-way-up*** with the computing processing order reversed so as to proceed from the bottom-up (see opposite).



SPSS users: Below is an example of syntax that if performed on an upside-down file that would load a final separation date in a transfer set back onto other records in the same transfer set.

SPSS (upside-down file)
```
COMPUTE findate = sepdate.
FORMATS findate (EDATE10).
IF (LAG(transseq) GT 0) findate = LAG(findate).
EXECUTE.
```

SAS users: Use the lag function on an upside-down file as with SPSS. However, in SAS the procedure must avoid placing a lag function within a conditional statement.

SAS (upside-down file)
```
data file2 (drop = lag_transseq);
set file1;
by rootlpno;
length lag_transseq findate 8.;
format findate ddmmyy10.;
retain findate;
lag_transseq = lag(transseq);
if first.rootlpno or lag_transseq = 0 then findate = sepdate;
```

Stata and R users: A similar approach to that in SPSS may used with an upside-down file:

Stata (upside-down file)
```
generate findate=sepdate
format findate %d
replace findate=findate[_n-1] if transseq[_n-1]>0
replace findate=sepdate if [_n]==1
```

R (upside-down file)
```
vaslinks4$findate <- vaslinks4$sepdate
for(i in 2:nrow(vaslinks4)){
   if(vaslinks4$transseq[i-1] > 0) vaslinks4$findate[i] <- vaslinks4$findate[i-1]
}
```

Alternatively, to move the value of a variable back a specified number of records in a right-way-up file, use can be made of ***explicit subscripting***. For example, the following syntax will work with up to a maximum of four transfers in a transfer set. For each record, it loops through the next four records to find the most distal record where transseq equates to the numbered iteration of the loop. Thus, if there are three transfers lying ahead of the first admission in the transfer set (for which transseq = 0), the replacement of findate on the first admission with the sepdate on the third transfer will occur, because for the third transfer record the value of transseq (= 3) continues to be equal to the number of the iteration of the loop (i = 3). However, for the next record, being the first one after the transfer set, transseq $\neq$ 4, and thus the conditional in the 4th line of syntax fails and no further replacement of findate occurs.

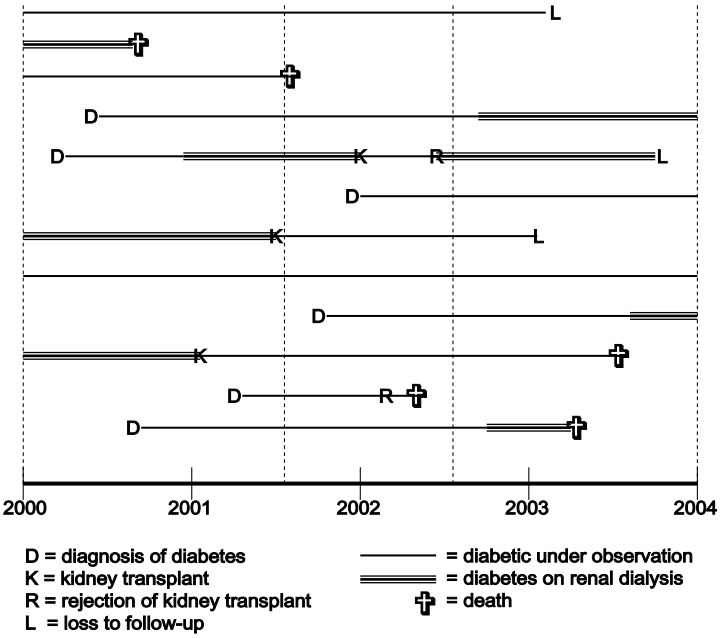
Stata (explicit subscripting; right-way-up file; maximum four transfers)
```
generate findate=sepdate if transseq==0
format findate %d
forvalues i=1(1)4 {
        replace findate=sepdate[_n+`i'] if transseq[_n+`i']==`i'
        }
```

Schematic and Conceptual Tools to Assist in Writing Syntax

There are a number of schematic and conceptual tools that can assist you in deciding on the best order in which to write the syntax needed for an analysis; knowing what syntax to write; and checking in advance that the syntax is likely to yield the desired result. You have come across some of these tools already and it is now a convenient time to summarise them.

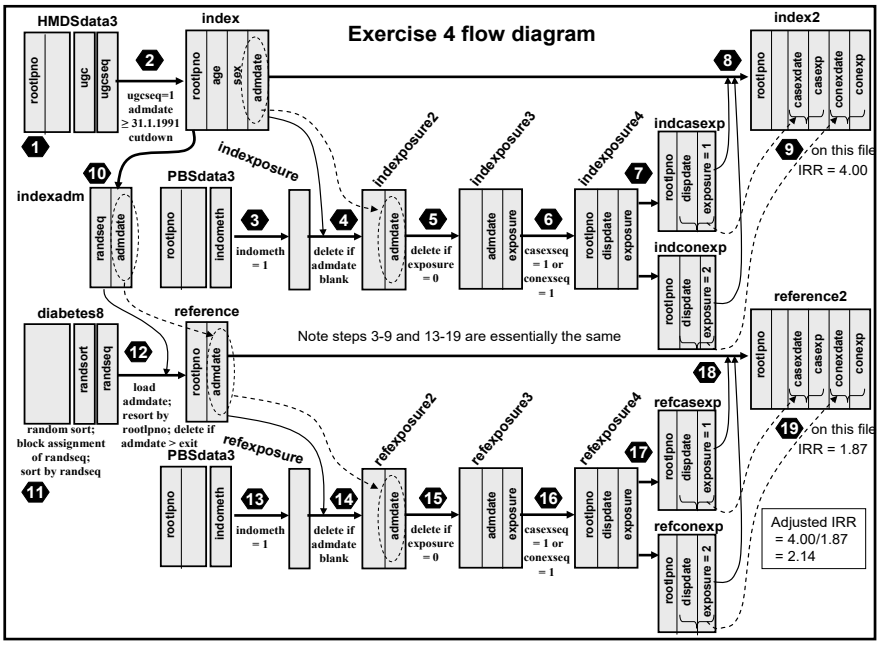
Time Traveller Traces

These diagrams are based on heath care episodes and, as we saw in the theory class, are useful in thinking through possible contingencies or relevant measurement theory prior to their implementation in syntax. Shown opposite is the example of renal failure in severe diabetics, but now with two cross-sectional lines marked at the mid-points of 2001 and 2002, being the dates when point prevalence estimate were requested for the proportions of severe diabetics on renal dialysis. Both the denominators and numerators of these prevalence measures are defined by which time traces straddle the cross- sections. In the advanced course we learn how to prepare relevant syntax based on these concepts/



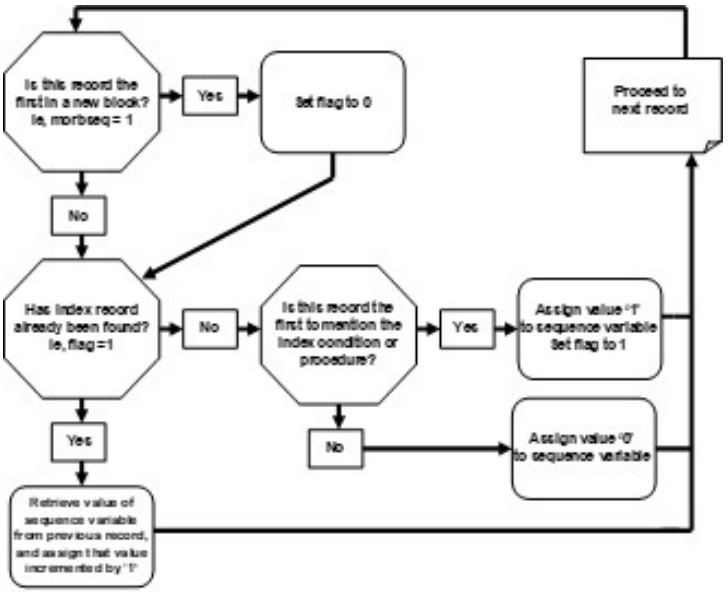
File Flow Diagrams

These enable you to plan out your syntax in terms of what files are to be used or created, how they are to be merged or developed and in what order the steps should be taken. Usually, the desired end result is a single file that contains all information needed for final analysis loaded onto a single record for each individual. You have seen two of these already at the end of exercises 1 and 2. However, file flow diagrams can become much more detailed than those, and the more complex the file management processes become, the more essential it is to map out in advance how you plan to proceed. For example, the file flow diagram below, taken from the advanced course, demonstrates a relatively more complex series of file preparations prior to the analysis of a case-time-control study.



Syntax Flow Diagrams

These diagrams can assist you to think through the logic behind a complex piece of syntax before you attempt to write it out in a computing language. They are especially helpful for planning syntax that will require the use of multiple conditional statements, whereby different processes need to be applied to records in a file that satisfy different conditions. For example, the syntax flow diagram opposite is how an SPSS programmer might thinkthrough the logic of computing an index sequence variable. Note that syntax flow diagrams have a tendency to become language-specific. The example diagram here is less suitable for the SAS or Stata programmer who intends to use by-group processing to generate the index sequence variable.



Pseudo-Syntax

Pseudo-syntax is a convenient way to jot down an initial draft of a piece of syntax without taking the time required to observe the language formalities required by your statistical package. It can also be a way for analysts using different languages to communicate syntax ideas to each other. There are no conventions or grammatical rules for pseudo-syntax. Rather, the writer merely uses easily understood words and arithmetic symbols to express the basic logic of the syntax. An example of pseudo-syntax given earlier in this course was the following approach to creating a tag variable with positional attribution. It should be possible for a SPSS, SAS, Stata or R programmer to work from this to prepare correct syntax in their respective language.

Pseudo-syntax

```
Set tag to 0.
Set counter to 0.
REPEAT for dummy = diag 1 TO diag6.
  IF dummy = 1st target value THEN tag = tag + 1 × 10^counter.
  IF dummy = 2nd target value THEN tag = tag + 2 × 10^counter.
  IF dummy = any target value THEN counter = counter + 1.
END REPEAT.
```

File Block Diagram

Drawing a diagram of one or more illustrative blocks of records for certain types of individuals in a file can be helpful either in guiding the preparation of syntax or in testing syntax 'by hand' to see if it produces the intended end results under different circumstances. An example of use of file block diagram to guide the preparation of syntax was anticipated in step 15 in exercise 1, where is was suggested that you sketch out such a diagram to help you identify the best way to count the numbers of 1st–time vasovasostomies occurring before and after a 1st–time vasectomy in the file. Below is an example of what your assistive diagram may have looked like:

rootlpno		morbseq	vasseq	revseq
		1	0	0
		2	0	0
		1	0	0
		2	1	0
		3	2	0
		4	3	1
		5	4	2
		1	0	0
		2	1	0
		3	2	0
		4	3	0
		5	4	0
		1	0	0
		2	0	1
		3	0	2
		4	0	3
		5	1	4
		6	2	5

This type of index vasosvasostomy record uniquely identifies men with 1st–time reversal after a 1st-time vasectomy; whereas the index vasectomy record is unhelpful, being indistinguishable from the pattern in the block below, for example, where no subsequent reversal occurred.

Criteria: revseq = 1 and vasseq > 1.

This type of index vasectomy record uniquely identifies men with 1st–time reversal before a 1st-time vasectomy.

Criteria: vasseq = 1 and revseq >1.

Overview of Exercise 3

Step 1. Open vaslinks file.

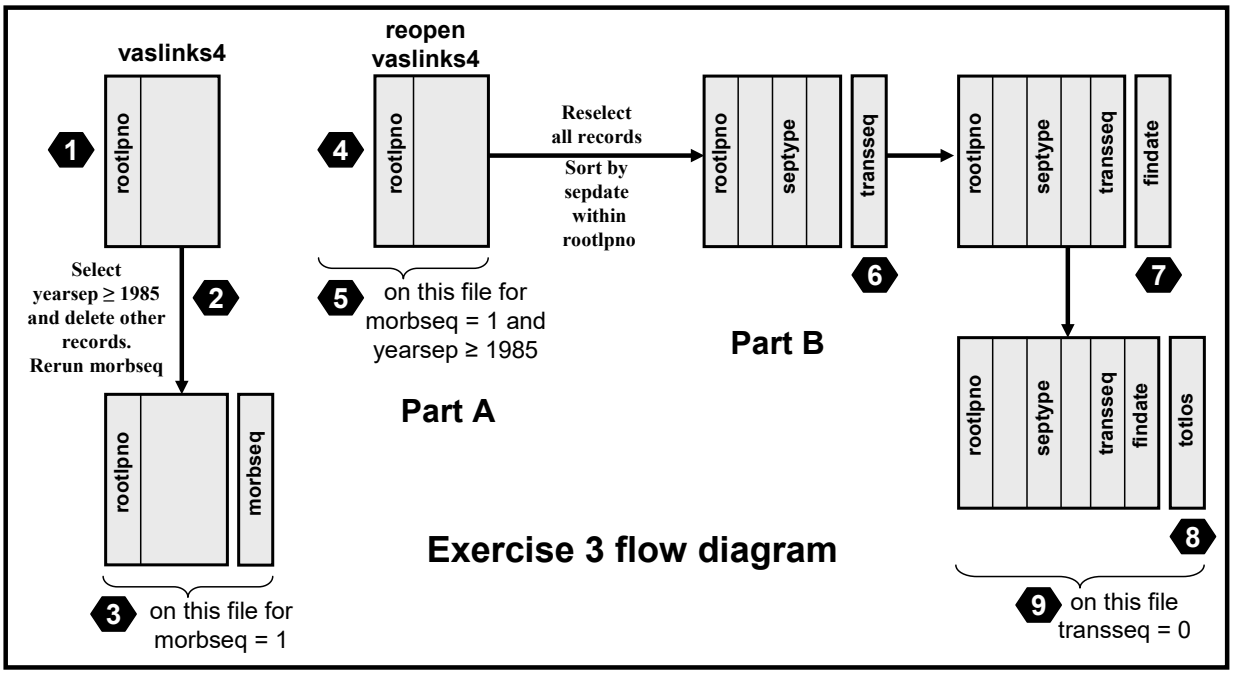
Step 2-3. Determine apparent 1st time admission post-1985

Step 3-5. Apply clearance period to correct prevalence pool effect.

Step 6. Create transseq variable to identify inter-hospital transfers (septype=2 & admdate<(LAG(sepdate)

Step 7. Backflow final separation date to index record.

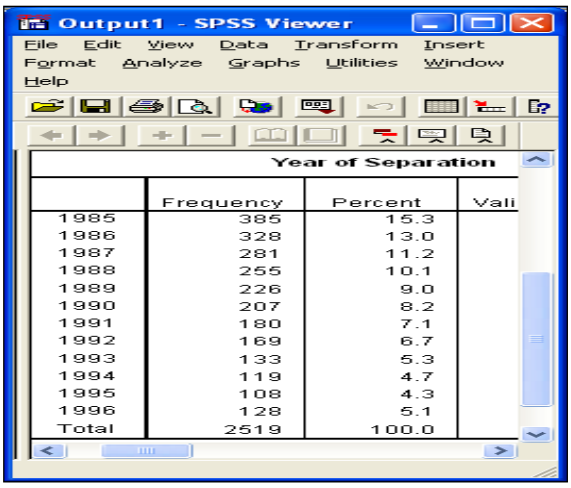
Step 5-6. Calculate total length of stay to include transfers.



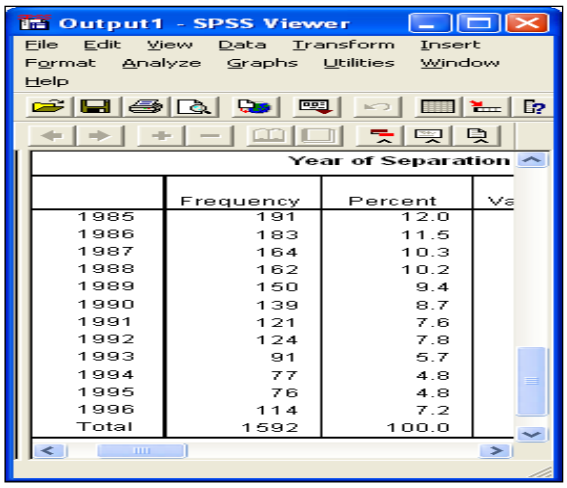
Exercise Instructions Part A: Prevalent Pool Effect

1. Open your vaslinks4 file as further developed and saved from exercise 2.
2. Part A provides a simple demonstration of the prevalent pool effect. Select the records where yearsep ≥ 1985 and delete the remaining records. Now rerun your syntax to create the morbseq variable and select the records where morbseq = 1. These are the apparent first-time hospital admissions from 1985 onwards, in men who at some time had an in-hospital vasectomy or vasovasostomy, in the absence of any earlier historical data. **Important: Do not save this file.**
3. Generate a frequency distribution of these records by yearsep. These numbers (shown on the left below) would be the numerators used to calculate incidence rates of first-time hospital admissions in this population of men.
4. Discard the above data without saving them and re-open your vaslinks4 file covering all the years of separation from 1980 onwards. Select the records where morbseq = 1 and yearsep ≥ 1985 but do not delete any records this time.
5. Generate a frequency distribution of these records by yearsep. These numbers (shown on the right below) would be the numerators used to calculate incidence rates of first-time hospital admissions where the prevalent pool identifiable from the past five years of hospital admissions (1980-84) has been cleared. Note how the numbers are vastly different and how the relative overestimation of first-time hospital morbidity without the clearance period is greatest in the early years of the time series. Reselect all the records.

SPSS

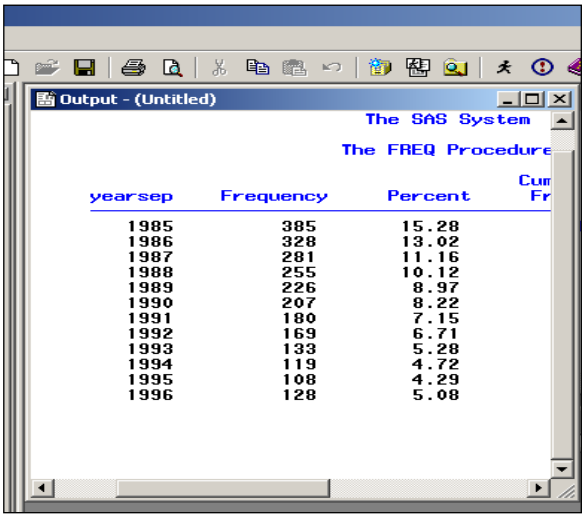


	Frequency	Percent	Vali
1985	385	15.3	
1986	328	13.0	
1987	281	11.2	
1988	255	10.1	
1989	226	9.0	
1990	207	8.2	
1991	180	7.1	
1992	169	6.7	
1993	133	5.3	
1994	119	4.7	
1995	108	4.3	
1996	128	5.1	
Total	2519	100.0	

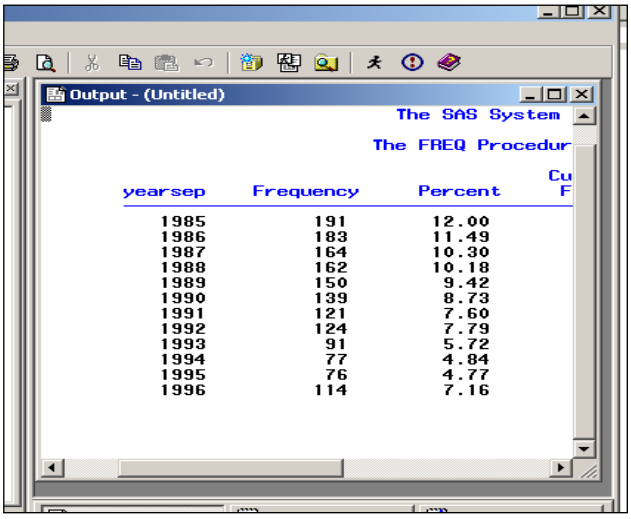


	Frequency	Percent	Va
1985	191	12.0	
1986	183	11.5	
1987	164	10.3	
1988	162	10.2	
1989	150	9.4	
1990	139	8.7	
1991	121	7.6	
1992	124	7.8	
1993	91	5.7	
1994	77	4.8	
1995	76	4.8	
1996	114	7.2	
Total	1592	100.0	

SAS

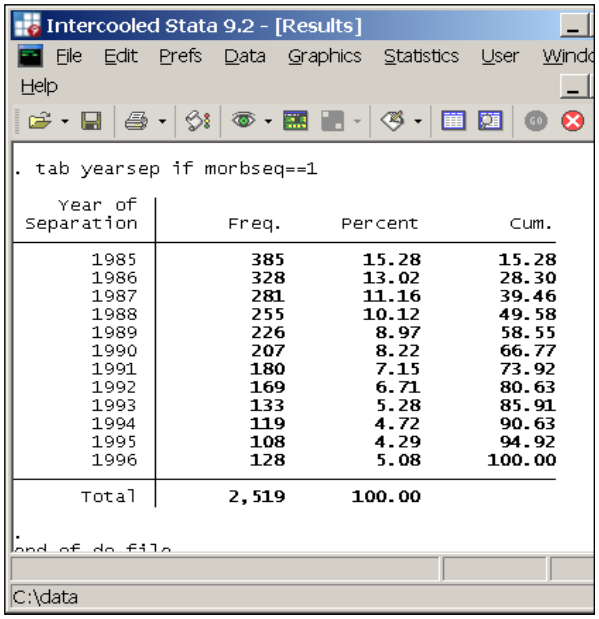


yearsep	Frequency	Percent	Cum Fr
1985	385	15.28	
1986	328	13.02	
1987	281	11.16	
1988	255	10.12	
1989	226	8.97	
1990	207	8.22	
1991	180	7.15	
1992	169	6.71	
1993	133	5.28	
1994	119	4.72	
1995	108	4.29	
1996	128	5.08	



yearsep	Frequency	Percent	Cu F
1985	191	12.00	
1986	183	11.49	
1987	164	10.30	
1988	162	10.18	
1989	150	9.42	
1990	139	8.73	
1991	121	7.60	
1992	124	7.79	
1993	91	5.72	
1994	77	4.84	
1995	76	4.77	
1996	114	7.16	

Stata

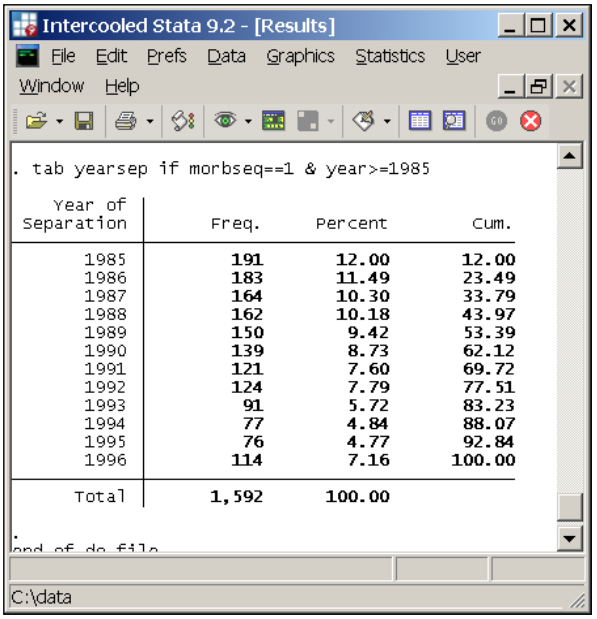


Intercooled Stata 9.2 - [Results]

File Edit Prefs Data Graphics Statistics User Windo
Help

```
. tab yearsep if morbseq==1

   Year of
Separation |      Freq.     Percent        Cum.
-----------+-----------------------------------
      1985 |        385       15.28       15.28
      1986 |        328       13.02       28.30
      1987 |        281       11.16       39.46
      1988 |        255       10.12       49.58
      1989 |        226        8.97       58.55
      1990 |        207        8.22       66.77
      1991 |        180        7.15       73.92
      1992 |        169        6.71       80.63
      1993 |        133        5.28       85.91
      1994 |        119        4.72       90.63
      1995 |        108        4.29       94.92
      1996 |        128        5.08      100.00
-----------+-----------------------------------
     Total |      2,519      100.00
```

C:\data



Intercooled Stata 9.2 - [Results]

File Edit Prefs Data Graphics Statistics User
Window Help

```
. tab yearsep if morbseq==1 & year>=1985

   Year of
Separation |      Freq.     Percent        Cum.
-----------+-----------------------------------
      1985 |        191       12.00       12.00
      1986 |        183       11.49       23.49
      1987 |        164       10.30       33.79
      1988 |        162       10.18       43.97
      1989 |        150        9.42       53.39
      1990 |        139        8.73       62.12
      1991 |        121        7.60       69.72
      1992 |        124        7.79       77.51
      1993 |         91        5.72       83.23
      1994 |         77        4.84       88.07
      1995 |         76        4.77       92.84
      1996 |        114        7.16      100.00
-----------+-----------------------------------
     Total |      1,592      100.00
```

C:\data

Exercise Instructions Part B: Inter-Hospital Transfers

6. Create a ***transfer sequence variable***, transseq, to number the records that are transfers using the simplified approach covered in the training session. Identify transfers as records with an admission date that occurs before the separation date of the previous record for the same patient, or on the same date provided that the separation type of the previous record indicates there was a transfer (ie, septype = 2). Note that to use the simplified approach to this problem recommended above, <u>your file must be sorted by ascending sepdate within rootlpno</u>. Obtain a frequency distribution for transseq to confirm there are 133 transfers distributed as 120 1st transfers, 10 2nd transfers, two 3rd transfers and one 4th transfer. Note that the 133 transfers occurred in 120 transfer sets.

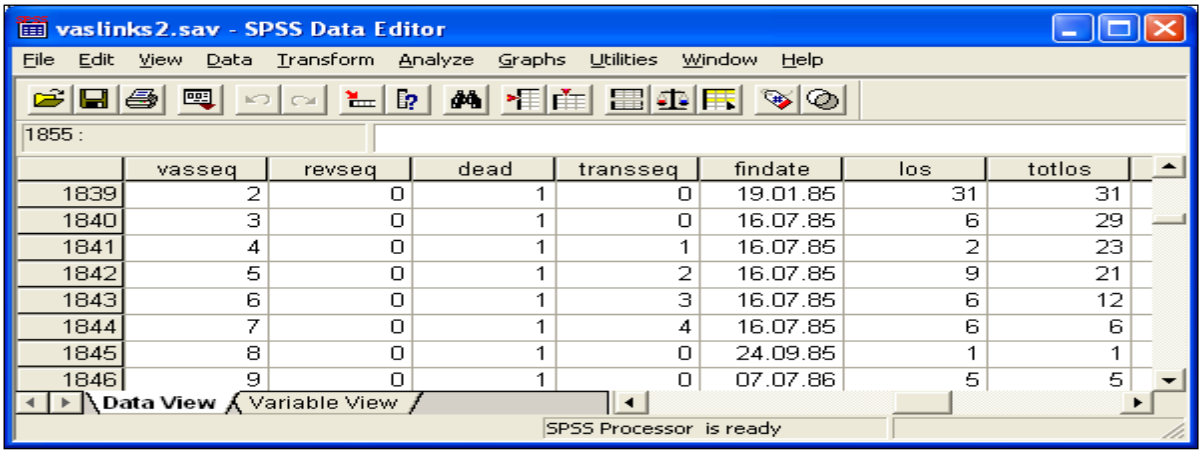
7. You should now compose syntax that will load the last date of separation in a transfer set back on to the earlier records in the set, including H1. The aim is to create a new variables: findate, which is the final separation date if the record is part of a transfer set, or simply just the current separation date if there is no transfer set involved. Use an information backflow method applicable to your package to assign the correct values to these three variables.

   HINT for SPSS, Stata and R users: It is best to turn the file upside-down by sorting in descending order of fileseq. Create findate and set its default value to be the same as sepdate for every record. Then use a lag function to ***propagate*** the last separation date within a transfer set backwards and use it to update findate on earlier records in the same transfer set.

   HINT for Stata users: Or, as an alternative approach to that above, if you prefer to use the file the right-way-up, you can generate findate and set its default value to sepdate for every record where transseq=0. You then create a loop that uses explicit subscripting to look forward at each of the next four records in turn [because max(transseq)=4] and to replace findate with the potentially later value of sepdate for each record looking forward where the transseq sequence number is equal to the loop number. See the text earlier in the training section concerning this method in Stata.

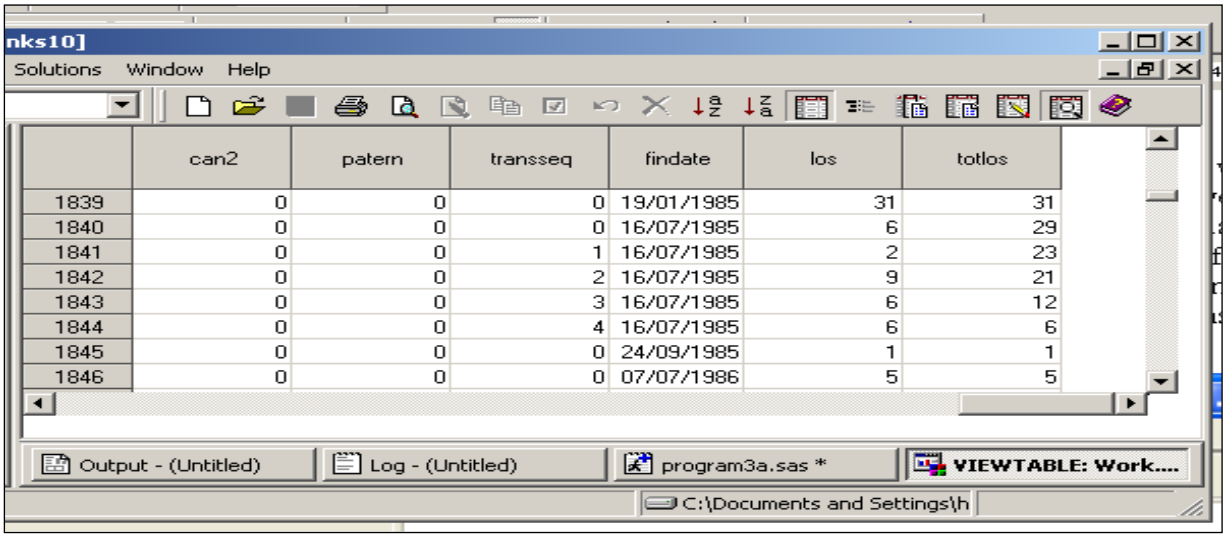
8. For every record in the file use date functions to create a length of stay variable, los, which is the difference in days between the date of separation and the date of admission. We will assume that by convention a los of '0' (***same-day admission***) is recoded to '1' (different countries have different rules about this). Now create a total length of stay variable, totlos, which is the difference in days between the final date of separation and the date of admission. Again, a value of '0' is recoded to '1'. At this stage, your data file should look something like the screen below, where record numbers 1840-44 are the transfer set with four transfers. Note how the totlos accumulates to 29 days on the H1 record at number 1840.

SPSS



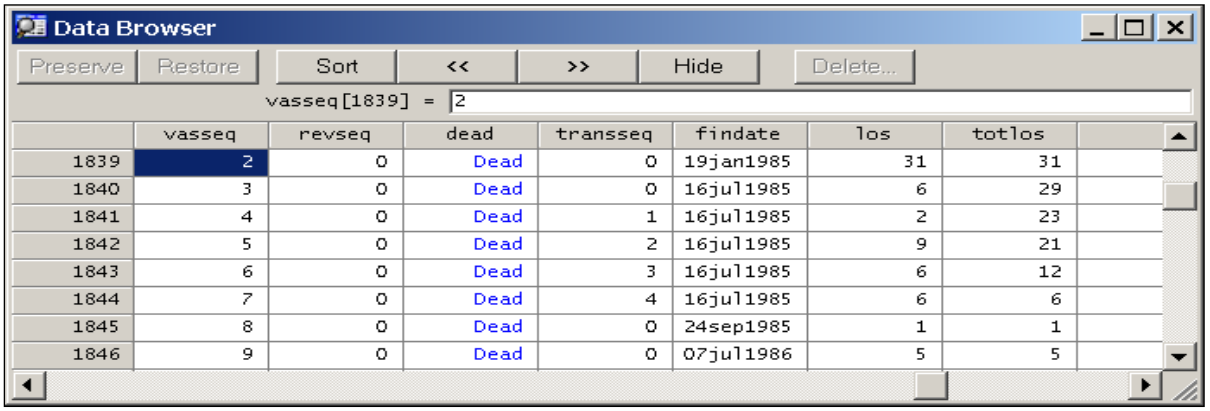
	vasseq	revseq	dead	transseq	findate	los	totlos
1839	2	0	1	0	19.01.85	31	31
1840	3	0	1	0	16.07.85	6	29
1841	4	0	1	1	16.07.85	2	23
1842	5	0	1	2	16.07.85	9	21
1843	6	0	1	3	16.07.85	6	12
1844	7	0	1	4	16.07.85	6	6
1845	8	0	1	0	24.09.85	1	1
1846	9	0	1	0	07.07.86	5	5

SAS



	can2	patern	transseq	findate	los	totlos
1839	0	0	0	19/01/1985	31	31
1840	0	0	0	16/07/1985	6	29
1841	0	0	1	16/07/1985	2	23
1842	0	0	2	16/07/1985	9	21
1843	0	0	3	16/07/1985	6	12
1844	0	0	4	16/07/1985	6	6
1845	0	0	0	24/09/1985	1	1
1846	0	0	0	07/07/1986	5	5

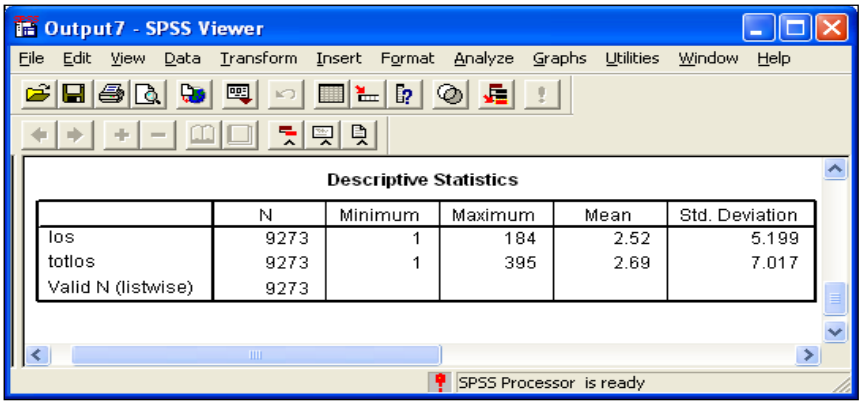
Stata



	vasseq	revseq	dead	transseq	findate	los	totlos
1839	2	0	Dead	0	19jan1985	31	31
1840	3	0	Dead	0	16jul1985	6	29
1841	4	0	Dead	1	16jul1985	2	23
1842	5	0	Dead	2	16jul1985	9	21
1843	6	0	Dead	3	16jul1985	6	12
1844	7	0	Dead	4	16jul1985	6	6
1845	8	0	Dead	0	24sep1985	1	1
1846	9	0	Dead	0	07jul1986	5	5

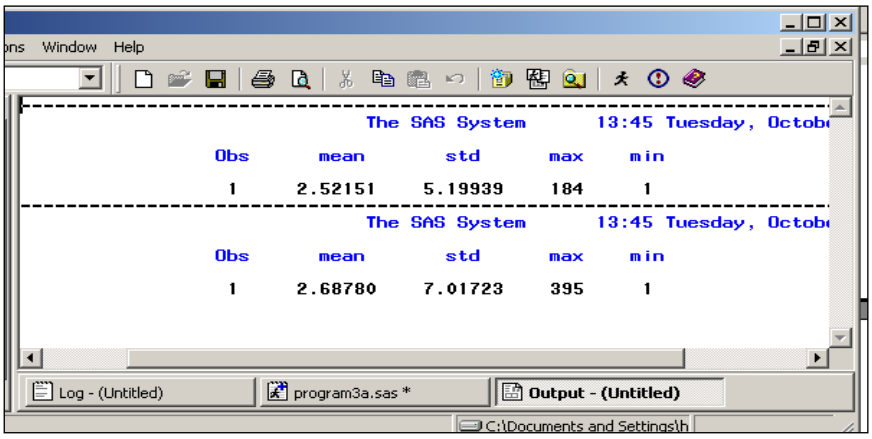
9. Now select the non-transfer records (ie, transseq = 0) and obtain the mean, SD and range for los and totlos. Your results should be as below. Note how taking account of the inter-hospital transfers reveals the true episodic length of stay and how the range is increased.

SPSS

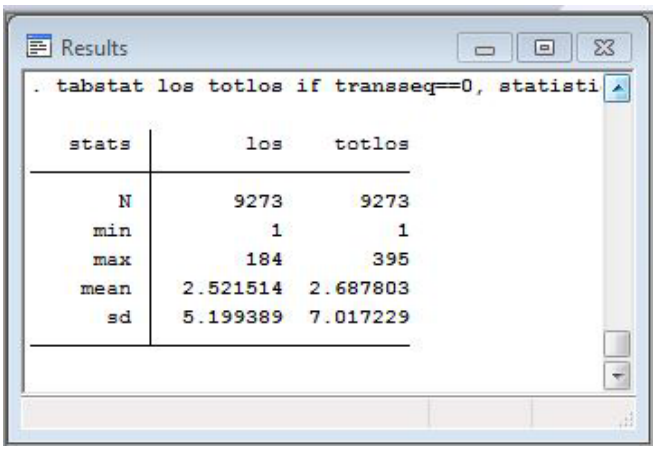


	N	Minimum	Maximum	Mean	Std. Deviation
los	9273	1	184	2.52	5.199
totlos	9273	1	395	2.69	7.017
Valid N (listwise)	9273				

SAS



Stata



stats	los	totlos
N	9273	9273
min	1	1
max	184	395
mean	2.521514	2.687803
sd	5.199389	7.017229

References

Holman CDJ, Wisniewski ZS, Semmens IL, Rouse IL, Bass AJ. **Population-based outcomes after 28246 in-hospital vasectomies and 1902 vasovasostomies in Western Australia.** *BJU International* 2000; 86: 1043-1049.

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

REVIEW 3: 'MEASURING HEALTH AND SERVICES USING LINKED DATA'

Please check that you are familiar with the following topics:

- Conceptualisation and calculation of basic epidemiologic measures ☐
- Rationale for and method of actuarial survival analysis ☐
- Relationships between basic measures ☐
- Conceptualisation and calculation of effect measures ☐
- Estimating effect measures from 2x2 tables ☐
- Sources of random and systematic error ☐
- Hypothesis testing and effect measurement ☐
- Index, first-time and cumulative admissions/separations/procedures/LOS ☐
- Health care episodes (disease/illness/care episodes) ☐
- The prevalent pool effect ☐
- Types of inter-hospital transfers ☐

Please check that you understand the meaning 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 validity ☐
- p-value and confidence interval ☐
- Principal and additional diagnoses and procedures, and external causes ☐
- Admission, separation, length of stay ☐
- Time traveller trace, file flow diagram, syntax flow diagram, pseudo-syntax and file bock diagram ☐

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

- Implement a clearance period to reduce the prevalent pool effect ☐
- Know how to work on an upside-down file (if applicable) ☐
- Count first-time incident events ☐
- Identify and number the sequence of inter-hospital transfers ☐
- Derive total length of stay for a hospital transfer set ☐

4. THEORY OF MEASUREMENT OF HEALTH OUTCOMES

4.1 Health Outcomes and Their Measurement

Some Historical Origins of the Health Outcomes Movement

Florence Nightingale: In 1863 she published ***Notes on Hospitals***, reporting deaths at 106 principal hospitals of England and noting a 'mortality percent on inmates' of 90.8% in 24 London hospitals. Some quotations: '*Facts such as these raise grave doubts as the advantages to be derived from hospitals at all, and have led many to think that a poor sufferer would have a much better chance of recovery if treated at home.' 'We have known incurable cases discharged from one hospital, to which the deaths ought to have been accounted, and received in another hospital, to die there in a day or two, thereby lowering the mortality rate of the first at the expense of the second.'*

Ernest Codman: In 1917 he published ***Case Records and Their Value***, promoting his 'end-results' idea at Massachusetts General Hospital. Some quotations: *'So I am called eccentric for saying in public: that Hospitals, if they wish to be sure of improvement, - must find out what their results are; must analyse their results; must compare their results with those of other hospitals; and must welcome publicity not only of their successes, but for their errors.' 'This volume is dedicated to Richard C. Cabot, because I respect his motives, admire his courage and energy, but heartily disapprove of some of his opinions and methods, for he seems to want to reform the bottom of the profession, while I think the blame belongs to the top.'*

Contemporary Origins of the Health Outcomes Movement

Contemporary US origins: The health outcomes movement first started in the US in the mid-1980s due to the increasing proportion of expenditures going to health care and the need for cost containment, recognition of the serious limitations of available information about the effects of many services and the perception of large variations in the use of medical procedures between geographical areas and providers. In 1986, the US Government mandated the Institute of Medicine to assess the US Medicare for quality and cost-effectiveness improvements, and a federal Agency for Health Care Policy and Research, now the *Agency for Healthcare Research and Quality*, was created.

Contemporary UK origins: In the UK, the outcomes movement began in the late 1980s, with 'health gain' initiatives commencing in Wales and with the NHS reforms. The UK perspective, as well focussing on cost and practice variation, also identified the importance of systematic review of evidence from randomised controlled trials as promoted through the *Cochrane Collaboration*.

Contemporary Australian origins: In Australia, the outcomes movement arguably started in 1990 with the *National Health Strategy* and specifically with a Background Paper entitled *Making it Better*. This led to the NHMRC establishing a Standing Committee on Quality of Health Care, from which NHMRC Clinical Practice Guidelines were developed.

Some Definitions Used in Outcomes Research

Health outcome: A change in the health of an individual, a group of people or population which is attributable to an intervention or a series of interventions.

Effectiveness: The extent to which an intervention produces a desired outcome. ***Efficacy*** is sometimes regarded as the potential for an intervention to be effective under ideal conditions of use.

Adverse event: An unintended injury or complication which results in disability, death or prolongation of hospital stay and is caused by health care management rather than the patient's disease. It may involve technical complications, diagnostic error, failure to implement known strategies for injury minimisation or medication error. In various countries, hospital chart reviews have revealed that 2.9–16.6% of patients in acute care hospitals experienced one or more adverse events (the *Australian Quality in Health Care* study and *Canadian Adverse Events Study* are examples of these surveys, as well as the seminal *Harvard Medical Practice Study*).

Safety: The extent to which the probability of preventable unintended injury or complications which may result in disability, death or prolongation of hospital stay, caused by health care management rather than the patient's disease, is minimised.

Accessibility: The extent of an individual's or population's ability to obtain health services. Related concepts are those of ***horizontal equity*** (equal treatment for those with equal need) and ***vertical equity*** (different treatment for those with unequal needs).

Appropriateness: The extent to which an intervention maximises potential health benefits valued by informed individuals or populations after considering likely outcomes, their probabilities and costs.

Taxonomy of Outcome Measures

The following is a compilation of different systems, including those of the *UK Clearing House for Information on the Assessment of Health Outcomes*, the *US Medical Outcomes Study* and the *Australian National Hospital Outcomes Program*. An important effect of the health outcomes movement has been a broader perspective on what are the evaluable end-results of health services, and the especially greater awareness of the importance of non-clinical patient-based outcomes.

A. ***Measures of effectiveness***:

   (1) ***Clinical endpoints*** ('traditional outcomes'): survival, person-years of life saved, symptoms (including pain), signs and laboratory markers of disease severity or recurrence.

   (2) ***Functional health status***: physical function (eg, impairment measures), mental health function (eg, depression, social adjustment measures), social function (eg, social support measures), measures of specific disability and handicap.

   (3) ***General well being***: (i) *generic instruments*: multidimensional health status profiles (eg, SF-36, Sickness Impact Profile) and indices (eg, Quality of Wellbeing Scale), which cover many of the dimensions of health-related quality of life in a single instrument; (ii) *disease-specific instruments* (eg, Quality of life Questionnaire for Cancer Patients).

B. ***Measures of safety***: eg, risks of complications, unplanned readmissions, revisional procedures and less direct iatrogenic diseases including cancer and birth defects.

C. ***Measures of economic outcome***: These include measure of:

   (1) ***cost-effectiveness***; eg, cost per case mix-adjusted separation;
   (2) ***cost-benefit***; eg, cost-benefit ratio; and
   (3) ***cost-utility***; eg, cost per quality-adjusted life year.

D. ***Measures of accessibility***: eg, waiting times for elective surgery, accident and emergency services and outpatients.

E. ***Measures of acceptability***; eg, patient satisfaction with care, patient perceptions.

F. ***Measures of appropriateness***; eg, adherence to best practice guidelines, variations in intervention rates.

Quality of Care

Quality of care is an overarching term which includes efficiency, effectiveness, accessibility, patient satisfaction and appropriateness of care and patient safety is that of quality of care, first championed by Donabedian in 1966. The determinants of quality of care were classified by Donabedian into the elements of structure (resources), process (tasks) and outcomes. The key idea in Donabedian's model is that a fundamental functional relationship exists between these components of health services: ie, structure $\Rightarrow$ process $\Rightarrow$ outcome.

4.2 Miettinen's Eight Elements of Study Design

To measure a change in health phenomena which might be attributable to an intervention it is necessary to have an adequate grasp of study design. The famous epidemiologist, Olli Miettinen, has identified the essential elements of study design as the domain, determinants, outcome, occurrence relation, confounders, effect modifiers, study base and sampling strategies.

Domain (1st element)

The domain of a study is concerned with the type of people for whom the results will be most immediately applicable in the sense of ***external validity*** or ***scientific generalisation***. Thus, if a study of the efficacy of a new chemotherapy agent is undertaken in post-surgical patients with stage II colorectal cancer, then the result may be immediately generalised (taken to be externally valid) for all patients with stage II colorectal cancer and not merely those who were participants in the study. This means that the study domain consists of patients with stage II colorectal cancer. Note that patients with stage I or stage III colorectal cancer are not part of the study domain and hence the results may not be valid for those types of patients.

The choice of the study domain is a major issue because it sets boundaries on the contributions to knowledge arising from the research. Often the domain is not so much a matter of 'choice' but rather a matter what type of patients or research participants are readily available to the researcher, For example, a study of a new method of treating diabetic ketoacidosis could be undertaken on patients with juvenile diabetes in a paediatric hospital, on patients with mostly gestational diabetes in a women's hospital, or on patients with mostly mature-onset diabetes in a geriatric medicine unit. Although the new treatment under investigation is the same in each instance, the study domains are all completely different and the scientific contributions in each instance will be limited to knowledge about the treatment of keto-acidosis only in the particular type of patient included in the research.

Determinants (2nd element)

These are often one or more new health service interventions that the researcher wishes to evaluate. However, it is also possible to have determinants that are not interventions, but rather characteristics of patients, such as their type of histopathology to be used as a prognostic indicator, or non-medical exposures that patients receive such as whether or not the patient drinks a lot of green tea.

It is important to distinguish the determinant(s), as the main focus of effect estimation, from potential confounders and effect modifiers (see below), which are better viewed as nuisance variables that are essential to measure, to control and to investigate, but which are not the motivation for performing the research.

The researcher needs to have the most valid and reliable method possible for measuring the study determinants. Many determinants are straightforward to measure; eg, whether or not the patients were offered or underwent a certain surgical procedure. Others are not easy at all; eg, whether or not a high risk patient was adequately heparinised against venous thrombosis prior to a period of prolonged post-operative immobility. The issues involved in obtaining valid and reliable measures of study variables, including determinants, are beyond the scope of this short course.

Outcome (3rd element)

The outcome is the variable that is the object of the research in terms of the purported benefits or harm arising from exposure to the determinant(s). Some studies have more than one outcome, in which instance it may be important to decide in advance which one will be the principal outcome variable. We have seen already a taxonomy of types of outcome measures that may be relevant to clinical and health services research.

Occurrence Relation (4th element)

The occurrence relation is the relationship between the outcome and its determinants. In outcomes research we are especially concerned with the strength of this relationship rather than whether or not a relationship exists. This is because very weak relationships between interventions and outcomes are of little practical importance. It is therefore essential to devise an ***effect measure*** to parameterise the direction and strength of the occurrence relation. We have seen already a classification of binary outcome effect measures divided into 'absolute' and 'relative'. These are revised below with the addition of absolute and relative effect measures of continuous outcomes such as blood pressure:

Type of outcome	Absolute effect measures	Relative effect measures
Binary outcome (yes/no) Eg, death, recurrence, disease complications, a test result above a set criterion.	• Incidence rate difference (IRD) • Cumulative incidence difference (CID) • Prevalence difference (PD) (all often called a risk difference, RD)	• Incidence rate ratio (IRR) • Cumulative incidence ratio (CIR) • Prevalence ratio (PR) (all often called a risk ratio, RR)
Continuous outcome Eg, blood pressure, exact test results, pain score, quality of life score.	Difference between means or medians Regression coefficient(s).	Ratio of means or medians

All of these effect measures are built from ***basic measures***, which are observed separately in persons exposed to different levels of the determinant(s). These simple measures are entities like incidence rates, cumulative incidences, prevalences and simple means:

Confounders (5th element)

The most important consideration of confounders in research is to identify them in advance so that steps can be taken to control the bias they cause through careful study design or adjustment in analysis. The most common variables acting as confounders (and effect modifiers) in health outcomes research are sociodemographic factors such as age and sex, severity of illness including complications and comorbidity from other illnesses.

Effect Modifiers (6th element)

We will cover the concept of effect modification and its distinction from confounding on the last day of this introductory course. As with confounders, the most important consideration of effect modifiers in health outcomes research is to identify them in advance so that steps can be taken to take account of the modification in the way that the study is designed or analysed.

Study Base (7th element)

This is the 'sampling frame' for the study from which the actual research subjects are selected. The most common type of study base used in health outcomes research is a cohort of patients who are enrolled in the study at a particular point in calendar time or, more commonly, a point in the history of their illness such as a date of diagnosis or date of treatment. This is what is known as a ***primary study base***, because the entire population of patients at risk of end points is enumerated from the outset, independent of which patients sustain an end point. There is also a **secondary study base**, where first of all a set of end points (= the ***cases***) is observed and then the researcher conceptualised what study base (= the ***population at risk***) must have existed to give rise to those cases.

An example of a primary study base would be a cohort of Hangzhou residents first admitted to a Hangzhou teaching hospital with a diagnosis of diabetes mellitus in 2003-2007. An example of a secondary study base would be the population of diabetic patients who in the event of developing diabetic renal failure would undergo first-time renal dialysis in a Hangzhou teaching hospital in 2003-2007. A specification of the study base (population at risk) then follows from a careful conception of the patients who have at least the potential to incur this outcome of renal dialysis, taking into account the nature of the research question to be posed. Specification of the study base has especially important implications for the proper design of case-control studies.

Sampling Strategies (8th element)

It is common in many health outcome research projects for all study measurements to be taken on all patients in the study base, in which case the sampling strategy is known as a ***census***. It need not be always necessarily to achieve a 100% sampling fraction to obtain valid and sufficiently precise results. For example, in a nested case-control design, the researcher take measurements on all patients with end points and a sample (<100%) of all of the patients in the cohort at time-zero (t_0 = the commencement of observation).

4.3 Experimental vs Non-Experimental Assignment of Determinant

Motivation for Experimental Studies

An ***experiment*** is a study in which the conditions are under the direct control of the researcher. Most importantly, in the context of health services research, in an experiment the researcher controls the conditions under which research participants are assigned to the different levels of the determinant(s). Thus an experiment, such as a ***clinical trial***, is used to assess the outcome of one or more interventions in quantitative terms.

A distinction is sometimes made between an ***explanatory trial***, which assesses the effect of the active component of intervention under ideal conditions ($\therefore$efficacy), and the ***pragmatic trial***, which assesses the effect of the intervention in practice ($\therefore$effectiveness). Experiments in clinical research are essentially a post-WWII phenomenon, their widespread adoption being first promoted by Sir Austin Bradford-Hill in 1948. While the traditional model of the experiment is a clinical trial of the intended effects of treatment on disease, the same principles of study design are applicable to many other research questions, including provider and consumer behaviour, preventive interventions and the effects of health service policies, diagnostic technology and delivery systems. Contrary to common misconceptions, linked administrative health data can and have been used in the context of experimental studies – you will see an example later in this session.

Theory of Experimental Studies: Implications for Non-Experimental Research

Control of systematic error: The technical objective of an experiment is to produce a valid estimate of the intended effects of intervention. The apparent effect of intervention (E) depends on –

- therapeutic effect of the treatment under investigation: TE
- natural history of disease: NH
- therapeutic effect of other medical care: OE
- patient subjectivity: PS
- doctor subjectivity: DS

In the intervention group: $E_i = TE_i + NH_i + OE_i + PS_i + DS_i$
In the control group: $E_r = TE_r + NH_r + OE_r + PS_r + DS_r$
The apparent effect of treatment is:
$E_i - E_r = (TE_i - TE_r) + (NH_i - NH_r) + (OE_i - OE_r) + (PS_i - PS_r) + (DS_i - DS_r)$

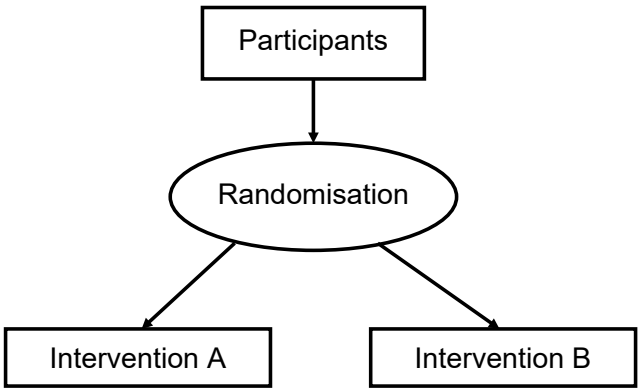
According to Miettinen, the paradigm of the ideal clinical trial is intended to produce:

1. ***comparability of populations***, $NH_i = NH_r$, through randomisation;
2. ***comparability of effects***, $OE_i = OE_r$, through randomisation and the use of placebo; and
3. ***comparability of information***, $PS_i = PS_r$ and $DS_i = DS_r$, through use of randomisation, placebo and double-blind conditions;

thereby making the apparent effect, E, equal to the therapeutic effects of the treatment, $TE_i - TE_r$.

Implication for non-experimental studies: To the extent that non-experimental studies are affected by sources of systematic error such as selection bias, confounding and information bias it is because these studies fail to achieve comparability of populations, effects and information as occurs in the experimental paradigm when perfectly implemented.

Determinant contrast: The simplest and most common form of experiment is the ***parallel group design***. Here subjects are allocated in parallel to two or more experimental conditions (eg, intervention vs placebo or new treatment vs established practice).



Implication for non-experimental studies: The classical form of clinical trial has its analogy in non-experimental health services research as the cohort study, the most important difference being that because the cohort study is non-experimental, the researcher has no influence over which research participants are exposed to which levels of the determinant(s). A large proportion of research based on linked health data takes the form of cohort studies.

Analysis of experiments: An important issue in the analysis of experimental trials is the distinction between ***intention to treat analysis*** and ***on-protocol analysis***. Intention to treat means analysis according to the original determinant allocations and not according to what interventions subjects eventually received, which may have differed due to non-compliance, protocol-deviation or loss to follow-up. One helpful way to think about this is that determinant is not ***actual exposure*** to the trialled intervention, but a strong ***opportunity to be exposed***. On-protocol analysis means analysis in which participants who did not actually receive their randomly assigned intervention (or lack of intervention) are excluded. The intention to treat analysis is internally valid, but tends to underestimate the effect of actual exposure because it is diluted by those in whom the opportunity to be exposed does not materialise. Both analyses should be reported.

Implication for non-experimental studies: Intention to treat analysis has analogies in the analysis of linked health data. These take the form of quasi-experimental studies (covered below) and propensity score and instrumental variable analyses (covered in the advanced course).

Motivation for Non-Experimental Studies

The double-blind, placebo'ed, randomised controlled trial enjoys a high degree of internal validity (lack of systematic error). However, it is not always possible nor even desirable to perform an RCT or other experiment to measure the effects of a health services intervention. This may be due to any of the following reasons:

Ethics and Practicability:

- The intervention is established practice.
- The intervention shows so much promise that ***clinical equipoise*** does not exist.
- A funding authority is unwilling to wait for evidence from an experiment or believes that a less demanding evaluation will be good enough.
- The cost of an experiment is prohibitive.
- The outcomes are too rare or long-term to be detected in an experiment.

External Validity and System-Wide Perspectives:

- We wish to evaluate the intervention as used by average providers on average patients.
- We wish to evaluate the performance of a provider of an intervention.

- We wish to assess the broader effects of the intervention within a health system.
- We wish to evaluate the patterns of utilisation of the intervention.

Under these circumstances, non-experimental methods of epidemiologic research may be used to estimate the effects of interventions and other exposures in populations whose exposure status has been determined by the complex interplays of real life.

4.4 Taxonomy of Non-Experimental Study Designs

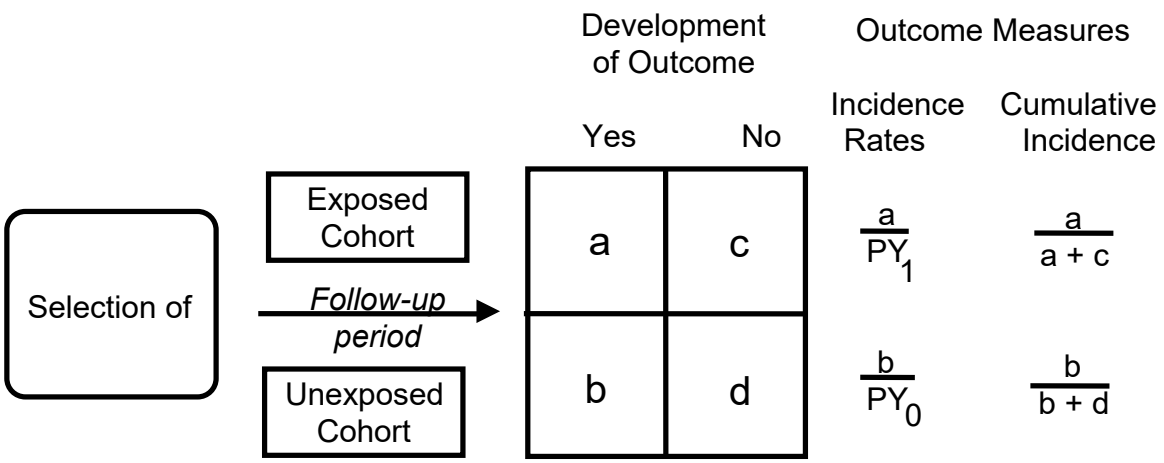
The following non-experimental study designs are available using a simplified system of classification:

A. ***Aggregational studies*** (ecological studies).
B. ***Person-based studies***:
   (1) ***Cross-sectional studies***.
   (2) ***Cohort studies***.
   (3) ***Case-control studies***.
   (4) ***Case distribution studies***.

In this theory session, we are concerned with the design and applications particularly of cohort and case-control studies. You should note that leading epidemiologic theorists, including Miettinen and Maclure, have pointed out that the taxonomy of study designs is a multi-dimensional problem; and uni-dimensional categories like those above are a simplification. There is not a singular formula to describe the features of a cohort study, for example, and the case-control sampling strategy may be applied to many different design situations.

4.5 Cohort Studies

In a cohort study, naturally occurring groups of exposed and non-exposed subjects, who have not yet experienced the outcome of interest, are identified from a source population and are followed over time to observe the occurrence of the outcome. Apart from the absence of experimental allocation, randomisation and a placebo and blinding of subjects, a cohort study is similar in design to a clinical trial. The methods of analysis of cohort studies and clinical trials are essentially the same. Note that cohort studies are always performed on cohort populations, where membership is defined by one or more events (eg, an index procedure) often (but not always) at the start of follow-up, known as time-zero (t_0).



Design Options for Cohort Studies

Type of reference group for the determinant(s): There are three main options – (1) internal comparison within the same cohort; (2) external comparison with a similar but non-exposed cohort and (3) general population comparison.

Prospective vs retrospective: The prospective cohort study is initiated at around the time of identification of exposure status and commencement of follow-up. It may take several years to complete. In the retrospective cohort study, identification of exposure status and follow-up is ascertained retrospectively using records of exposure and outcomes. This produces a much faster result.

Matching: For each exposed subject, one or more non-exposed subjects are selected having the same value of some matching factor which is a potential confounder. Matching in cohort studies controls for confounding and may improve study efficiency.

Analysis of Cohort Studies

Data from a cohort study may be presented in the form of a comparison of incidence rates or cumulative incidence measures between exposed and non-exposed subjects, or a comparison of means in the instance of a continuous outcome. The effects measures may be of the absolute or relative type. Depending on which type of measure is being used, there are different formulae applied for hypothesis testing and effect measurement. Time-related events are handled with a life-table analysis. Modelling techniques usually involve the use of Cox regression for person-time (time to failure) data; Poisson regression for person-time data with repeat outcomes; logistic regression for count data on binary outcomes and multiple linear regression for continuous outcomes.

4.6 Uses of a Data Linkage System in Longitudinal Research

'Longitudinal research' refers to either a clinical trial (experimental form) or cohort study (non-experimental form). There are three possibilities for the use of a pre-planned data linkage system in longitudinal health outcomes research:

1. ***Encapsulated linked longitudinal study***: This is conducted entirely within the information resources covered by the data linkage system. The study subjects are identified, followed up from time-zero and classified according to their determinant(s) and outcome(s) using the databases covered by the data linkage system as the only source of study data. Moreover, it should be possible to perform such studies using anonymised linked health data, thus conserving privacy to a very high degree.
2. ***Linked longitudinal study with supplementary data***: This occurs when the patients to be studied are identified and followed up from time-zero for subsequent events of interest (eg, death or out-migration as censoring events) using databases covered by the data linkage system, but there is at least one variable on which supplementary information must be sought from outside the system using a database that is not pre-linked or data collection in the field (eg, hospital chart reviews, patient or doctor questionnaires). The variable in question may be a missing determinant or outcome measure, confounder or effect modifier.
3. ***Linked follow-up of an independently enrolled cohort***: This occurs when the study participants have already been identified and enrolled into an experimental or non-experimental cohort population independently of the data linkage system, usually with recording of baseline data on exposures, confounders and effect modifiers. An *ad hoc* linkage to data linkage system is then undertaken to enable follow-up and ascertainment of outcomes recorded in the health databases covered by the system. An example of this would be a randomised clinical trial requiring long-term follow-up of hard clinical end-points such as cancer, stroke or death.

The following pages give examples of each of the three types of use of a data linkage system.

Encapsulated Linked Longitudinal Study

Kemp A, Preen DB, Morlet N, Clark A, McAllister IL, Briffa T, Sanfilippo F, Ng JQ, McKnight C, Reynolds W, Gillies MC. **Myocardial infarction after intravitreal vascular endothelial growth factor inhibitors: a whole-population study**. *Retina* 2013; 33: 920-927

MYOCARDIAL INFARCTION AFTER INTRAVITREAL VASCULAR ENDOTHELIAL GROWTH FACTOR INHIBITORS

A Whole Population Study

ANNA KEMP, PHD,*† DAVID B. PREEN, PHD,* NIGEL MORLET, FRANZCO,‡
ANTONY CLARK, MBBS,‡§ IAN L. MCALLISTER, FRANZCO,‡¶ TOM BRIFFA, PHD,**
FRANK M. SANFILIPPO, PHD,** JONATHON Q. NG, MBBS, PHD,†† CHARLOTTE MCKNIGHT, MBBS,‡¶
WAYNE REYNOLDS, MBBS,‡ MARK C. GILLES, FRANZCO, PHD‡‡

Purpose: To determine the risk of thromboembolic and gastrointestinal bleeding events in the 12 months after injections of bevacizumab or ranibizumab compared with photodynamic therapy and a nontreated community sample.

Methods: Hospital and death records were examined for 1,267 patients treated with vascular endothelial growth factor inhibitor and 399 patients treated with photodynamic therapy attending Western Australian eye clinics from 2002 to 2008, and 1,763 community controls, aged ≥50 years. Hospital records from 1995 to 2009 were analyzed for history of myocardial infarction (MI), stroke, and gastrointestinal bleeding before treatment. Records were searched for evidence of these events in the 12 months after treatment.

Results: The 12-month MI rate was higher for vascular endothelial growth factor inhibitor patients than photodynamic therapy patientsand the community group (1.9/100 vs. 0.8 and 0.7, respectively). No differences were observed between patients treated with bevacizumab and ranibizumab. The adjusted MI rate was 2.3 times greater than the community group (95% confidence interval, 1.2–4.5) and photodynamic therapy rate (95% confidence interval, 0.7–7.7). The 12-month MI risk did not increase with the number of injections administered (hazard ratio, 0.9; 95% confidence interval, 0.5–1.5). Stroke and gastrointestinal bleeding did not differ between any exposure groups.

Conclusion: Although all the adverse events examined were rare, patients treated with vascular endothelial growth factor inhibitors were significantly more likely to experience fatal or nonfatal MI than the community group. This increased risk may be related to the underlying age-related macular degeneration or vascular endothelial growth factor inhibitor use itself.

RETINA 33:920–927, 2013

Table 3. Results from Cox Proportional Hazards Regression Comparing Risk (and 95% CIs) of MI in the 12 Months After Index Treatment and Number Needed to Harm for Bevacizumab and Ranibizumab Patients, PDT Patients, and the Community Group

Exposure Groups	Unadjusted HRs	Adjusted HRs*	Number Needed to Harm
VEGF inhibitors vs. community	2.47 (1.25–4.87) $P = 0.01$	2.29 (1.16–4.54) $P = 0.02$	93 (52–414)
VEGF inhibitors vs. PDT	2.44 (0.73–8.14) $P = 0.15$	2.32 (0.70–7.74) $P = 0.17$	97 (52–∞)†
PDT vs. community	1.02 (0.29–3.58) $P = 0.89$	0.97 (0.28–3.43) $P = 0.96$	6,896 (142–∞)

*Models included age, sex, history of MI, and history of diabetes.
†The CIs spanning infinity indicate nonsignificant estimates for number needed to harm.[25]

Linked Longitudinal Study with Supplementary Data

Kemp A, Preen DB, Saunders C, Boyle F, Bulsara M, Malacova E, Roughead EE. **Early discontinuation of endocrine therapy for breast cancer: who is at risk in clinical practice?** *SpringerPlus Oncology* 2014; 3: 282
doi:10.1186/2193-1801-3-282

RESEARCH **Open Access**

Early discontinuation of endocrine therapy for breast cancer: who is at risk in clinical practice?

Anna Kemp[1*], David B Preen[1], Christobel Saunders[2], Frances Boyle[3], Max Bulsara[4], Eva Malacova[1] and Elizabeth E Roughead[5]

Abstract

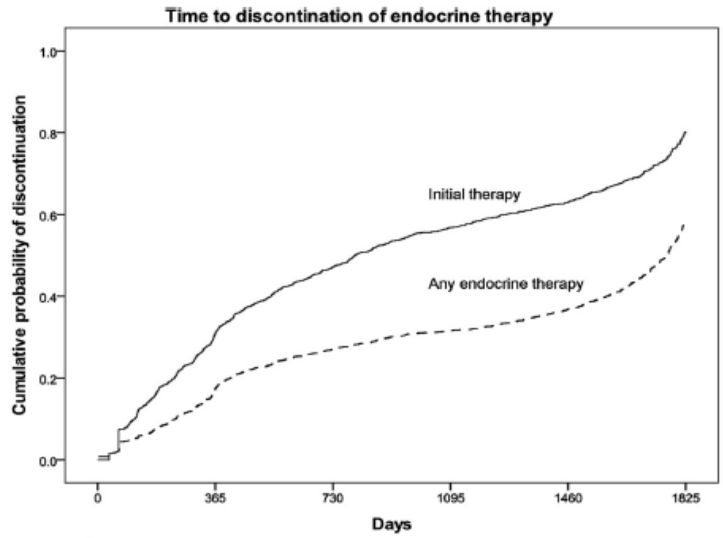
Purpose: Despite evidence supporting at least five years of endocrine therapy for early breast cancer, many women discontinue therapy early. We investigated the impact of initial therapy type and specific comorbidities on discontinuation of endocrine therapy in clinical practice.

Methods: We identified women in a population-based cohort with a diagnosis of early breast cancer and an incident dispensing of anastrozole, letrozole or tamoxifen from 2003–2008 ($N = 1531$). Pharmacy and health service data were used to determine therapy duration, treatment for pre-existing and post-initiation comorbidities (anxiety, depression, hot flashes, musculoskeletal pain, osteoporosis, vaginal atrophy), demographic and other clinical characteristics. Time to discontinuation of initial, and any, endocrine therapy was calculated. Cox regression determined the association of different characteristics on early discontinuation.

Results: Initial endocrine therapy continued for a median of 2.2 years and any endocrine therapy for 4.8 years. Cumulative probability of discontinuing any therapy was 17% after one year and 58% by five years. Initial tamoxifen, pre-existing musculoskeletal pain and newly-treated anxiety predicted shorter initial therapy but not discontinuation of any therapy. Early discontinuation of any therapy was associated with newly-treated hot flashes (HR = 2.1, 95% CI = 1.3–3.3), not undergoing chemotherapy (HR = 1.4, 95% CI = 1.1–1.8) and not undergoing mastectomy (HR = 1.5, 95% CI = 1.2–1.8).

Conclusions: Less than half of women completed five years of endocrine therapy. Women at greatest risk of stopping any therapy early were those with newly-treated hot flashes, no initial chemotherapy, or no initial mastectomy. This suboptimal use means that the reductions in recurrence demonstrated in clinical trials may not be realised in practice.

Keywords: Persistence; Adherence; Tamoxifen; Aromatase inhibitors; 45 and up study



Initial therapy	No. at risk	1027	757	581	368	115
	Discontinued	32%	48%	57%	64%	81%
Any therapy	No. at risk	1254	1069	932	646	268
	Discontinued	17%	27%	31%	37%	58%

Figure 1 Kaplan-Meier graph showing time to discontinuation of initial endocrine therapy, and any endocrine therapy.

Linked Longitudinal Follow-up of an Independently Enrolled Cohort

Calver J, Wiltshire A, Holman CDJ, *et al*. **Does health assessment improve health outcome in indigenous people? A RCT with 13 years of follow up.** *ANZ J Public Health* 2005; 29: 107-111.

Does health assessment improve health outcomes in Indigenous people? An RCT with 13 years of follow-up

Janine Calver, Amy Wiltshire, C. D'Arcy J. Holman
School of Population Health, The University of Western Australia

Ernest Hunter
School of Population Health, University of Queensland

Carol Garfield, Diana L. Rosman
Data Linkage Unit, Department of Health, Western Australia

Abstract

Objective: To examine the impact of a multi-component health assessment on mortality and morbidity in Kimberley Aboriginal residents during a 13-year follow-up.
Method: A population-based randomised controlled trial using linked hospital, cancer and death records to evaluate outcomes in 620 intervention and 6,736 control subjects.
Results: The intervention group had a higher rate of first-time hospitalisation for any reason (IRR=1.37; 95% CI 1.25-1.50), a higher rate of injury-related hospital episodes (IRR=1.31; 95% CI 1.15-1.48) and a higher notification rate of alcohol-related cancers. There was a smaller difference in the rates of multiple hospitalisations (IRR=1.14; 95% CI 0.75-1.74) and no improvement in overall mortality compared with controls (IRR=1.08; 95% CI 0.91-1.29).
Conclusions: There was no overall mortality benefit despite increased health service contact associated with the intervention.
Implications: Although not influencing mortality rates, multi-component health assessment may result in a period of increased health service use in Aboriginal and Torres Strait Islander populations, thus constituting an 'intervention'. However, this should not be confused with systematic and sustained interventions and investment in community development to achieve better health outcomes.

Table 2: Age-adjusted incidence rate ratios (IRRs) of mortality and cancer outcomes in the intervention groups compared with control group.

Mortality/cancer outcome measure	Intention to treat analysis[a]			On protocol analysis[a]		
	IRR	95% CI	*p*-value	IRR	95% CI	*p*-value
Mortality						
All-cause mortality						
All ages	1.08	0.91-1.29	0.373	1.12	0.92-1.36	0.281

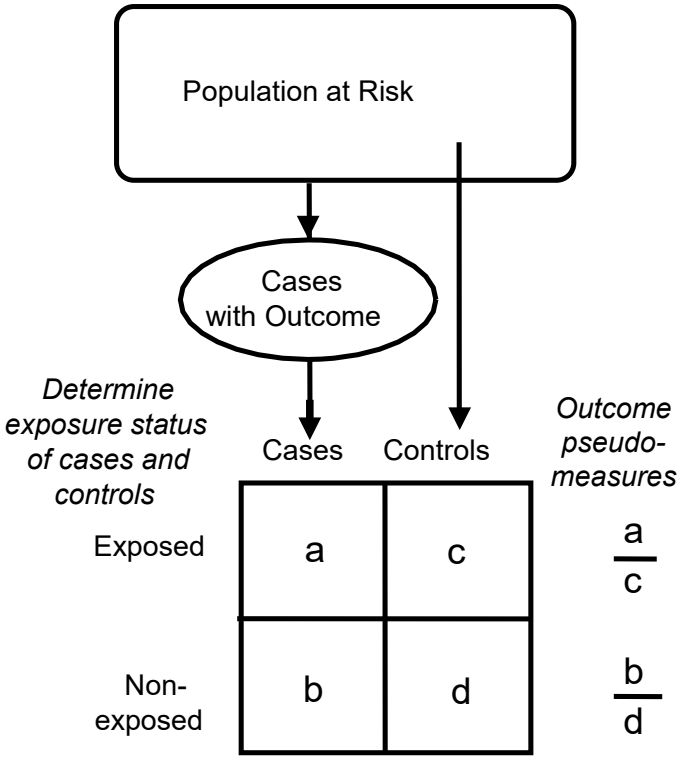
Table 4: Age-adjusted incidence rate ratios (IRRs) of first hospital separation in the intervention groups compared with the control group.

First hospital separation outcome measure	Intention to treat analysis[a]			On protocol analysis[a]		
	IRR	95% CI	*p*-value	IRR	95% CI	*p*-value
All first hospital separations	1.37	1.25-1.50	<0.001	1.54	1.39-1.70	<0.001

It is important that you understand the cohort study, as it is the design most frequently implemented using linked administrative data sets. It is the only design that you need to understand to complete the practical exercises in this introductory course. However, do not be misled into thinking that case-control studies are barely relevant to the analysis of linked health data sets (see below).

4.7 Case-Control Studies

In the case-control study, the subjects are selected according to the presence of the outcome of interest and not according to exposure levels of the determinant(s). Persons with a given outcome (***cases***) and persons at risk of the outcome (***controls***) are selected from either a dynamic population or a cohort population. Case-control studies require less subjects to provide adequate precision, and in the absence of a data linkage system can often be completed within a shorter time frame and are often the only practical study design to use in the case of a rare outcome. However, because information on exposure to the determinant is often (but not always) collected retrospectively, they can be subject to a greater level of information bias.



Case-Control Study on a Dynamic Population

When the case-control study examines cases from a dynamic population, the controls should be selected at random in such a way as to validly represent the distribution of the person-time in that population according to exposure status. Exposure status may refer to an exposure at the point of control recruitment and/or a history of past exposure that the control subject carries with them as a risk indicator. Exposure status is determined in exactly the same way in the cases, either using a questionnaire survey, laboratory tests, patient records or administrative data.

The ratio of exposed to non-exposed controls provides an estimate of the ratio of exposed to non-exposed person-time in the dynamic population, ie, $c/d \approx PT_1/PT_0$. This relationship is used to estimate the incidence rate ratio comparing the rate in the exposed with the rate in the non-exposed.

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

Case-Control Study on a Cohort Population (Hybrid or Nested Case-Control Study)

When the population at risk is a cohort rather than a dynamic population, a case-control sampling strategy may be used as an alternative to a cohort study as a means to perform the study more efficiently. Here the cases are the same as in the equivalent cohort study, but the controls are a sample of the population at risk rather than a census of the entire cohort.

The OR from a hybrid study has different interpretations according to how the controls were selected. If the controls are selected at random from the entire cohort population at time-zero (the *case-base study*), then $c/d \approx N_1/N_0$ and the OR is an estimate of the cumulative incidence ratio. If the controls are selected from the residual of the cohort at risk each time a case occurs (*incidence density sampling* in the *case-cohort study*), then $c/d \approx PT_1/PT_0$ and the OR is an estimate of the IRR.

Design Options for Case-Control Studies

Control/case ratio: The number of cases is often limited whereas controls are plentiful. Study precision can be increased by multiple controls per case up to a point. Usually more than 4-5 controls per case provides insufficient further improvement in precision to warrant the additional effort.

Population-based vs clinic-based controls: Whether the controls are best selected from the general population or from among patients presenting to the same clinic or hospital for unrelated reasons should depend on the way cases are defined and identified. If everyone in a demographically defined population is at risk of becoming a case in the study, it is population-based and the controls should be selected from that population. If only patients attending a specific clinic can be cases, the study is clinic-based and controls should be selected accordingly.

***Matching*:** Controls are selected so as to match their corresponding cases with respect to confounding factors such as age and sex. Matching in case-control studies does not control confounding, but when performed it may increase study efficiency provided that the factors matched are strong confounders.

Analysis of Case-Control Studies

Data from a case-control study are analysed to obtain the odds ratio, which may then be interpreted as either an IRR, CIR (or CIOR), PR (or POR) depending of the details of the design. Modeling techniques usually involve the use of logistic regression.

Applications of Case-Control Studies

The case-control design is used extensively in aetiologic research. In the field of health services research, its most common use is in clinical safety research and the quality of care (eg, cerebral palsy and the quality of obstetric care; vaginal cancer and DES exposure in utero). It is less often used in clinical efficacy research (eg, prevention of myocardial infarction and antihypertensives), except in the case of the evaluation of screening and early detection programs (eg, breast cancer and BSE; cervix cancer and cervical cytology). It is rarely used in the evaluation of diagnostic technology.

Because linked health data sets are readily construed in terms of events that define cohort populations rather than dynamic populations, it has been relatively uncommon to perform case-control studies on them. The situation is changing, however, and the important role of case-control concepts has become increasingly apparent. A ***nested case-control study*** 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 strategies such as chart review. An illustrative example of this is outlined below where the supplementary data concerned the type of chemoprophylaxis used in cataract surgery patients. It may also be preferred for ethical reasons to reduce the volume of health data released from administrative data collections. Case-control concepts have also led to the new case distribution stable of research designs (see 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.

Reducing the risk for endophthalmitis after cataract surgery: Population-based nested case-control study

Endophthalmitis Population Study of Western Australia sixth report

Jonathon Q. Ng, MBBS, PhD, Nigel Morlet, FRACS, FRANZCO, Max K. Bulsara, MSc, James B. Semmens, MSc, PhD

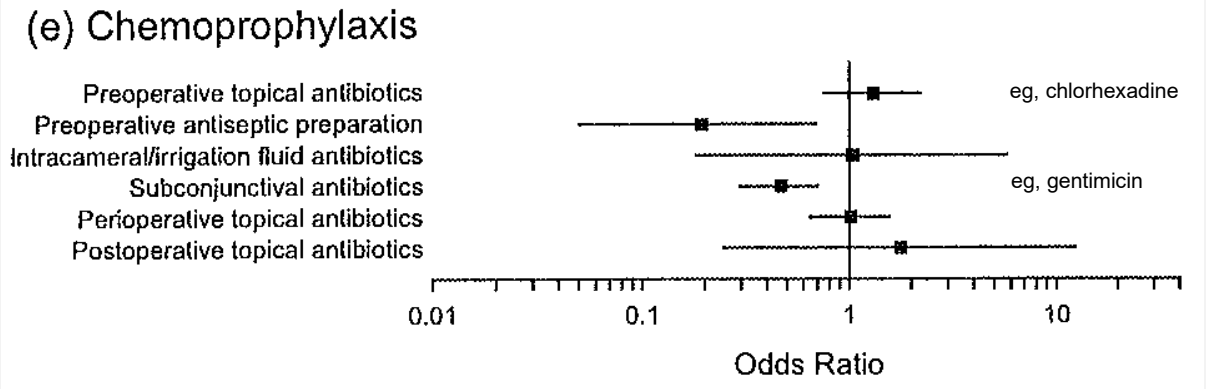
PURPOSE: To characterize operative and nonoperative risks for the potentially blinding complication of endophthalmitis after cataract surgery.

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%.



4.8 Differences Between Cohort and Case-Control Studies

Until the 1970s, case-control and cohort studies were viewed as mirror image study designs. The cohort study started with subjects in different exposure categories and looked forward in time to measure the experience of outcomes. The case-control study was viewed as the same process in reverse: cases and control groups being followed back in time to determine their exposure experience. This way of visualising cohort and case-control studies is known as the *classical viewpoint*.

In fact, the classical concepts of moving forward from exposure to disease in the cohort study, and backward from disease to exposure in the case-control study are popular simplifications which have no bearing on the order in which information is recorded or identified by the researcher. There is no reason why information on exposure status at the beginning of follow-up in a cohort study cannot be collected after follow-up is complete (as in the retrospective cohort study), apart from the fact it may achieve greater accuracy to collect the data earlier. Similarly, in a case-control study, exposure information on cases and controls may be obtained from records made prior to the diagnosis of disease in the cases. For this and several other reasons, the classical viewpoint is confusing.

An alternative *contemporary viewpoint*, advanced by Miettinen, is based on the recognition that the two types of study designs are not mirror images (the *trohoc* fallacy). Often (but not always) they are performed on different population types and are always distinguished by the extent of enrolment of the population at risk. A cohort study is always performed on a cohort population and the whole population at risk of becoming a case in the study is enrolled. The case-control study may be performed on either a cohort or dynamic population and enrols a sample of the population at risk of becoming a case as the controls.

4.9 Case-Distribution Studies

Commencing with Maclure's case-crossover study in 1991, a new stable of epidemiologic study designs has been instituted, known as *case-distribution studies*. They are a form of case-control study, which have the distinguishing feature of not using separate control subjects to estimate the distribution of the population at risk according to exposure status. Initially four different case distribution designs were described: (1) case-crossover study; (2) case-time-control study; (3) case-genotype study; and (4) case-specular study. The first two types of case distribution study are especially valuable to the field of health services research.

Case-distribution studies use 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 provides the basis for effect measurement. These studies use alternative methods to a control group to serve this function.

Several forms of case-distribution design are well-suited to the use of linked administrative health data to study short-term adverse events, such as in medication safety or post-procedural risks.

<u>Example</u>: *McEvoy et al (2005) used a case-crossover design to examine driver's use of a mobile phone within 10 minutes of a road crash and during a 10-minute period at the same time of day one week earlier. Use of a mobile phone whilst driving was associated with a four-fold increase in the incidence rate of road crashes within the following 10 minutes. The study required ad hoc linkage to mobile phone account records.*

Case-distribution studies are covered in some detail in the advanced course, especially their use in pharmaco-epidemiology.

References

Baker GR, Norton PG, Flintoff V, *et al*. **The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada.** *CMAJ* 2004; 170 (11). doi:10.1503/cmaj.1040498.

Brennan TA, Leape LL, Laird NM, *et al*. **Incidence of adverse events and negligence in hospitalized patients. results of the Harvard Medical Practice Study I.** *NEJM* 1991; 324:370-377.

Buring JE, Hennekens CH. **Methodology of intervention trials in individuals.** In: Detels R, Holland WW, McEwen J, Omenn GS, (eds). *Oxford Textbook of Public Health. Vol 2*, pp. 585-595. New York, OUP, 1997.

Calver J, Wiltshire A, Holman CDJ, *et al*. **Does health assessment improve health outcome in indigenous people? A RCT with 13 years of follow up.** *ANZ J Public Health* 2005; 29: 107-111.

Campbell D, Stanley J. **Experimental and Quasi-experimental Designs for Research.** Chicago:McNally, 1966.

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

Clark A, Morlet N, Ng JQ, Preen DB, Semmens J. **Long term trends and outcomes of anterior vitrectomy in Western Australia.** *Acta Opthalmologica* 2014; doi: 10.1111/aos.12453.

Finn JC, Jacobs IG, Holman CDJ, *et al*. **Outcomes of out-of-hospital cardiac arrest patients in Perth, Western Australia, 1996-1999.** *Resuscitation* 2001; 51: 247-255.

Greenland S, Morgenstern H. **Classification schemes for epidemiologic Research designs.** *J Clin Epidemiol* 1988; 41: 715-716.

Greenland S. **A unified theory of case-distribution analyses.** *Am J Epidemiol* 1997; 145: A225.

Greenland S. **Randomisation, statistics, and causal inference.** *Epidemiology* 1990; 1: 421-429.

Guyatt G, Sackett D, Taylor DW, *et al*. **Determining optimal therapy – randomised trials in individual patients.** *NEJM* 1986; 314: 889-892.

Haggar F, Pereira G, Holman CDJH, Einarsdottir K, Preen DB. **Adverse obstetric and perinatal outcomes following treatment of adolescent and young adult cancer: a population-based cohort study.** *PLOS ONE* 2014; 9(12): e113292 doi:10.1371/journal.pone.0113292.

Haggar FA, Preen DB, Pereira G, Einarsdottir K, Woods J, Boushey R, Mamazza R. **Maternal and neonatal outcomes in pregnancies following colorectal cancer.** *Surgical Endoscopy* 2013; 27: 2327 – 2336.

Hennekens CH, Buring JE. **Epidemiology in Medicine.** Boston: Little Brown & Co, 1987.

Hoffmeister H, Mensink GBM **Community based intervention trials.** In: Detels R, Holland WW, McEwen J, Omenn GS, (eds) . *Oxford Textbook of Public Health. Vol 2*, pp. 571-584. New York, OUP, 1997.

Holman CDJH, Wisniewski ZS, Semmens JB, *et al*. **Population-based outcomes following 28 246 in-hospital vasectomies and 1 902 vasovasostomies in Western Australia.** *BJU Int* 2000; 86: 1043-1049.

Kemp A, Preen DB, Morlet N, Clark A, McAllister IL, Briffa T, Sanfilippo F, Ng JQ, McKnight C, Reynolds W, Gillies MC. **Myocardial infarction after intravitreal vascular endothelial growth factor inhibitors: a whole population study.** *Retina* 2013; 33: 920-927.

Kemp A, Preen DB, Saunders C, Boyle F, Bulsara M, Malacova E, Roughead EE. **Early discontinuation of endocrine therapy for breast cancer: who is at risk in clinical practice?** *SpringerPlus Oncology* 2014; 3: 282 doi:10.1186/2193-1801-3-282.

Khoury MJ, Flanders WD. **Nontraditional epidemiologic approaches in the analysis of gene-environment interaction** *Am J Epidemiol* 1996; 144: 207-213.

McEvoy SP, Stevenson MR, McCartt AT, *et al*. **Role of mobile phones in motor vehicle crashes resulting in hospital attendance: a case-crossover study.** *BMJ* 2005; 331: 428-430.

Maclure M. **Taxonomic axes of epidemiologic study designs: A refutationist perspective.** *J Clin Epidemiol* 1991; 44: 1045-1053.

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.

Meinert CL. **Clinical Trials. Design, Conduct, and Analysis.** New York: Oxford University Press, 1986.

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

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.

Pocock SJ. **Clinical Trials. A Practical Approach.** Chichester: John Wiley & Sons, 1983.

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

Weis NS. **Clinical Epidemiology: The Study of the Outcome of Illness.** New York: Oxford Press, 1996.

TRAINING SESSION 4:
HEALTH OUTCOMES RESEARCH WITH LINKED DATA

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

Training session 4 continues with the use of your linked file of vasectomy and vasovasostomy data. We will be using the file to illustrate the use of non-experimental data linkage research methods to study both adverse outcomes and treatment benefits.

Training Session Resources

In this exercise you will make further use of your vaslinks data file saved from exercise 2:

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

Survival Analysis Variables

Survival analysis is the general term given to a number of different methods of analysis of longitudinal follow-up data taken from cohort studies. These forms of analysis are very commonly employed with linked administrative health data, because such data sets lend themselves to the study of time-dependent outcome events in cohorts of patients.

The two most common forms of survival analysis are the actuarial (or life table) survival analysis and Cox regression. In both instances, two survival variables are required to perform the analysis:

1. A ***status variable*** (sometimes called *outcome variable*), which is set to 1 if the study event is experienced by the patient during the observation period and otherwise set to 0. It is thus also what is termed a ***binary indicator***.
2. A ***time variable***, which is the length of time that the patient is followed before the study event is experienced or, if no study event occurs, the length of time to the end of the available observation period.

There are a number of important practical considerations in the calculation of these survival variables, most of which have a theoretical basis.

Time-Zero (t_0)

This is the date when patients or study participants enter their period of follow-up for the purpose of observing if the study event occurs. It is conceptually the point in time at which risk of the event begins. Typically, risk of outcome events (good and bad) begins from the time of an index health event, and for hospital morbidity data, the two readily available dates are the index admission date or index separation date. Sometimes other dates, such a date of cancer diagnosis, are available and are the more ideal t_0. Many analyses involve a theoretical t_0 that commences from the date of a procedure performed during hospital stay. Usually such procedure dates are absent from administrative data sets and the admission date is used as an approximation. More often than not, procedures are performed early during a hospital stay, with the remainder of the stay providing for a period of recovery. Thus the admission date is usually the best approximation to a true procedure date. However, if the study event is hospital re-admission or a revisional procedure after discharge, t_0 should be taken as the index separation date extended to include any contiguous transfers. This is because there is no risk of readmission until the patient is discharged from the index stay.

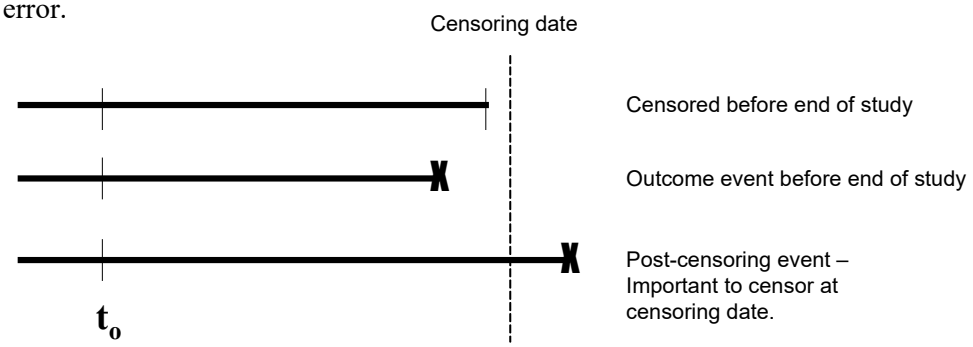
Note that there can be instances where t_0 commences, for at least some cohort members, at an arbitrary point in time, dependent on the date when outcome event data first become available rather than the date when the index health event occurred. For example, in a study of cancer incidence in men undergoing inpatient vasectomy from 1980, if the cancer registry commenced full population coverage from 1 January 1982, then this will become t_0 for men vasectomised in 1980-81; whereas t_0 in men vasectomised from 1982 onwards will be their date of vasectomy.

Censoring Date

The date at the end of the period of follow-up is the date of the outcome event for those who experience it. For those who survive without experiencing the outcome event, there must still be a date to define the end of follow-up of that individual and this is known as the *censoring date*. For subjects who are observed until the end of the study period, the censoring date is set to the final day of the study period. Sometimes subjects who do not experience the outcome event are nevertheless withdrawn from follow-up before the end of the study for other reasons. A common reason in clinical populations is the death of the patient, where death is not the outcome event. For example, if the outcome event is readmission to hospital, a patient who dies in the community without being readmitted before the end of the study period would be censored at their date of death. A second reason for premature censoring is known, permanent migration out of the geographic area covered by outcome information system. Under these circumstances the date of censoring would be taken as the best approximation available to the date of migration.

Post-Censoring Events

It is an important principle that outcome events only count for the purpose of analysis if they occur within the study period. For example, if the study period ends on 31 December 2004, a subject who is known to have experienced the outcome event on 1 January 2005, for the purposes of this study, must be recorded as having a '0' outcome status and a survival time censored on 31 December 2004. To do otherwise would be to treat this subject differently from others in the study and this would introduce a source of error.



Once-Off and Repeated Outcome Events

Survival analysis is suitable for what are known as ***time to failure data***. A feature of data of this type is that only the first outcome event is the object of study and any subsequent repeat outcome events (assuming they are possible) are ignored. If the density of repeated outcome events is of particular research interest, it is best to use incidence rates with person-time denominators or Poison regression.

Creating a Default Follow-up Date

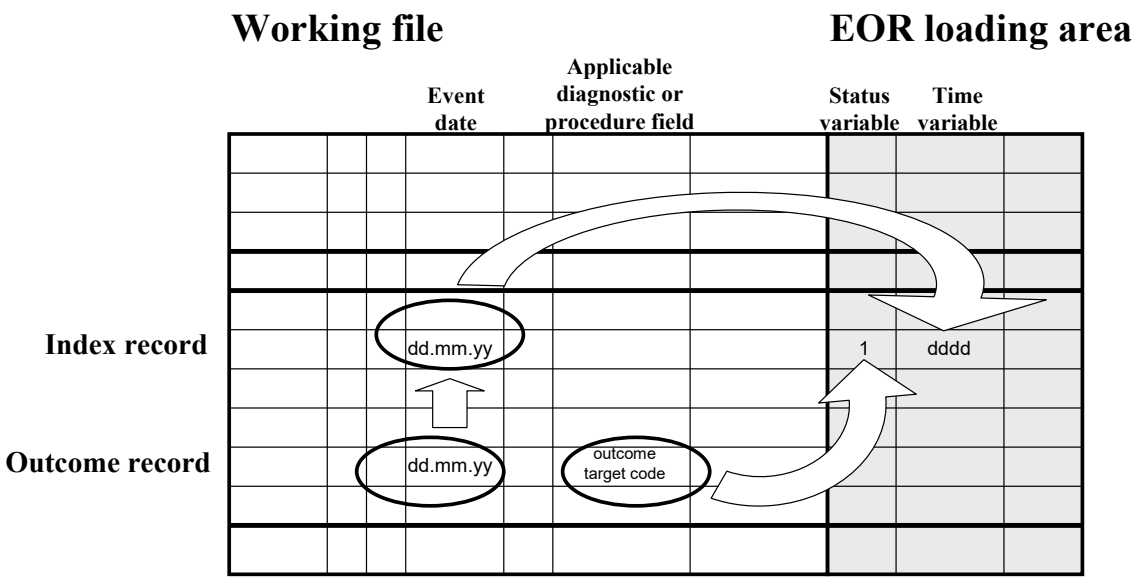
A default follow-up date is when a person's observation period is censored in the absence of the outcome event. It is convenient to create a value of this date variable, call it followup, for every record in the file, especially where survival from more than one outcome is to be analysed. The word 'default' signifies that it is the censoring date used in the absence of occurrence of each outcome.

In practice, the values of followup are determined by hierarchical replacement. At first, the actual date representing the end of the study is assigned to followup for every record in the file. Then for those who died before the study end date the value of followup is replaced by the person's date of death. Finally, if information on outward migration is available, or there is another basis on which to remove a person from the population at risk, then a further replacement of the value of followup is made. Some analysts prefer to go even further and replace followup with an earlier outcome event date where applicable, but then followup is no longer a 'default' follow-up date, but an actual follow-up date *per se*.

If the outcome event is death itself, it does not matter that the default follow-up date is also the date of death, because a deceased person is indeed at risk of death up to that point. However, the guidelines above would not work for a post-mortem cohort study concerning organ harvesting in trauma victims who are brain dead. This extreme example serves to make the point that there may be uncommon scenarios where the general guidelines set out above require principled modification.

Information Backflow for Survival Variables

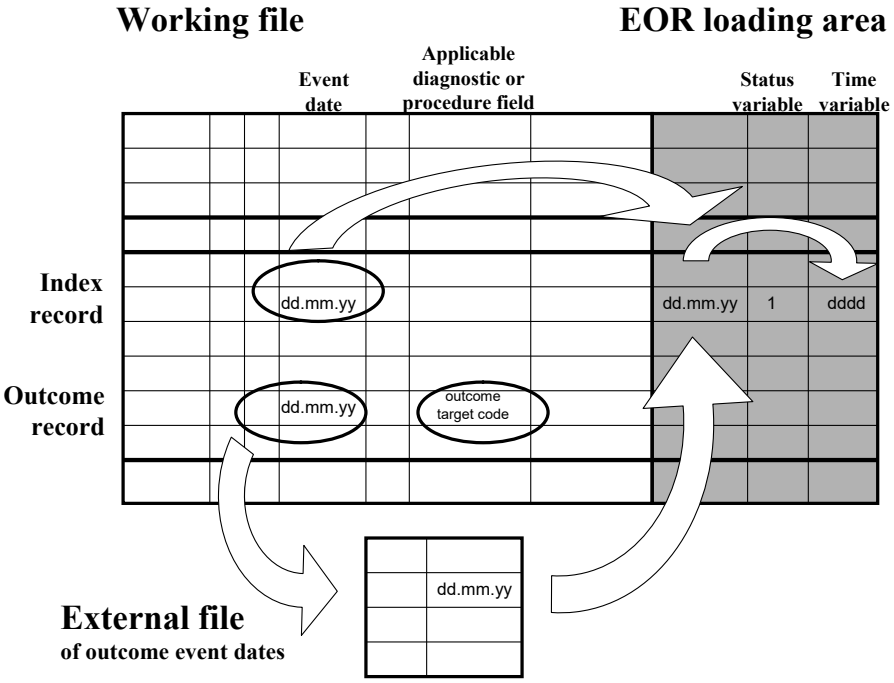
In Training Session 3 we considered the problem of how to find relevant information on later records for an individual and to load this information back onto the EOR loading area of the index record. This is highly relevant to survival analysis of encapsulated longitudinal data, where the outcome event (if it is present) must be detected on a later record in the individual's block. Once detected, the survival time between t_0 and the date of the outcome event must be calculated and that information also must be loaded back onto the index record. An illustration of the desired result appears below.



There are at least three methods available to achieve the desired result: (i) using a ***vector file***; (ii) using an ***upside-down file***; or (iii) use of ***explicit subscripting*** in Stata. Of these approaches, the vector file is the simplest to implement *de novo* in all languages and thus we will consider it first.

Information Backflow Using a Vector File

A vector file is a temporary external file, usually with just a small number of variables, that is created for the purpose of transmitting information from one part of a block to the end-of-record loading area of another part of the block. In practice, the transmitted information is often appended to all records in the block, albeit that it may only be intended for use on the index record.



The diagram above illustrates the use of a temporary external file as a vector file to transmit information needed to calculate survival variables from a first-time outcome record to a preceding index record. Note that the vector file will need to contain rootlpno identifiers to enable its cargo to be merged with the correct blocks in the working file. The steps involved are as follows:

Step 1 – Create an external file of outcome records: The first record for each individual to mention the outcome event is selected and saved as a separate file.

Step 2 – Cut-down external file of outcome records to rootlpno ***and outcome date***: All variables in the external file of outcome records are deleted save for rootlpno and the outcome event date variable, which is renamed. This is now the vector file.

Step 3 – External loading of outcome dates: The outcome event dates are loaded back onto the original working data file, using a merge procedure, so that the outcome event date is loaded onto every relevant record in the working file keyed by rootlpno.

Step 4 – Create a status ***binary indicator variable***: The status binary indicator variable is set to '1' if an outcome event has been loaded from within the relevant risk period; otherwise it is '0'. IMPORTANT: Ignore any post-censoring outcome event that occurs after followup.

Step 5 – Create a time ***variable***: The time variable is assigned the duration in days between the relevant date on each record (eg, admission date) and the date on which observation of the individual ceased, either because they sustained a valid outcome event or they ran out of follow-up time.

The vector file method is readily implemented in SPSS, SAS, Stata and R. Below is an illustration of what form the basic vector file syntax could take in SAS for steps 1-3:

SAS (vector file)

```
* Create vector file;
data vasvector (drop = revseq);
  set vaslinks4 (keep = rootlpno admdate revseq rename = (admdate = vasodate));
  where revseq = 1;
run;

* Merge vector file information onto working file;
data vaslinks5;
merge vaslinks4 (in=wf)
   vasvector (in=ef);
by rootlpno;
if wf then output;
run;
```

R (vector file)

```
# Create vector file (vasodates)
vasosub <- vaslinks4[vaslinks4$treat==2, ]
vasodates <- aggregate(vasosub$admdate,list("rootlpno"=vasosub$rootlpno),min)

# Merge vasodates information to working file
vaslinks4 <- left_join(vaslinks4, vasodates, by=c("rootlpno"="rootlpno"),
               keep = FALSE)
names(vaslinks4)[which(names(vaslinks4)=="x")] <- "vasodate"

rm(vasosub, vasodates)
```

Information Backflow Using an Upside-Down File

This method can be implemented in all languages, but it is most suited to SPSS due to the availability of scratch variables. To implement a similar approach in SAS or Stata requires the use of temporary variables and the SAS implementation has additional challenges. The following illustrative syntax and narrative outlines one possible approach to achieving the desired result in SPSS.

SPSS (upside-down file)

```
COMPUTE reverse = 0.
IF ($casenum EQ 1 or rootlpno NE lag(rootlpno)) #reverse = 0.
IF ($casenum GT 1 and rootlpno EQ lag(rootlpno)
         AND #reverse EQ 1) reverse = 1.
DO IF (treat EQ 2 and admdate LE followup).
  COMPUTE reverse = 1.
  COMPUTE #reverse = 1.
END IF.
EXECUTE.
IF ($casenum EQ 1 or rootlpno NE lag(rootlpno)) #date = followup.
IF (treat EQ 2 and admdate LE followup) #date = admdate.
COMPUTE timerev = xdate.tday(#date) - xdate.tday(admdate).
EXECUTE.
```

Step 1 – Create and propagate a* status *binary indicator variable: This is done using both a status data file variable and a #flag scratch variable. While it is possible to write syntax that will only set status to '1' on the index record if a subsequent outcome event occurs, there are advantages (eg, in analyses where 2nd, 3rd, etc, outcomes are to have separate analyses) in adopting an approach whereby going backwards through an individual's block, status is set to '0' until a record with an outcome event is discovered. For this record and all earlier records status is set to '1'. Thus the interpretation of a status value of '1' is that an outcome event has occurred on the same or later record (but ignoring any post-censoring outcome event).

Step 3 – Create a* time *variable: This is also done using both a time data file variable and a #date scratch variable. The time variable has a value assigned for every record in the file, which is the number of days between the relevant date on each record (eg, admission date) and the date that follow-up of the individual ceased, either because they sustained the outcome event, the end of the study period was reached or the individual was censored due to death or some other reason before the end of the study period. Post-censoring records are thus assigned a negative time value. While it is possible to write syntax that will only record a value for time on the index record, there are advantages (see above) in assigning a time value to every record. An effective approach is, going backwards through an individual's block, #date is assigned the value of the followup date until a record with an outcome event is discovered. For this record and all earlier records #date is set to the date of the event. For each record, time is calculated as the number of days elapsed between the relevant date on each record and #date. Thus time on the outcome record has the value of zero; and time on the index record has the value needed for a survival analysis.

Note that if there are multiple outcome records in an individual's block, the recommended approach will 'rewind' the survival time to zero in each instance where an outcome record is encountered. This results in time on the index record being assigned the survival time to the date of the first outcome event, which is exactly what a time to failure analysis requires.

Information Backflow Using Explicit Subscripting in Stata

The final method peculiar to Stata is illustrated below. Here the analyst has observed in advanced that the maximum number of records in a block starting from the index vasectomy record is 47; ie, max(vasseq) = 47. The explicit subscripting loop starting in the 2nd line of syntax searches ahead to find an index reversal record (revseq = 1), if it exists, and after checking that it does not represent a post-censoring event (at the end of the 4th line), it replaces the default 0 value of reverse with a value of 1 on the index vasectomy record (vasseq = 1). The last line of syntax then replaces followup with the admission date of the reversal record. Thus this particular syntax is not using followup as a 'default' follow-up date, but rather as a follow-up date *per se*.

Stata (explicit subscripting)

```
generate reverse=0
forvalues i = 1(1)47 {  /* 47 is the maximum number of all vasseq*/
        set more off
        by rootlpno: replace reverse=1 if vasseq==1 & revseq[_n+`i']==1 & admdate[_n+`i']<=followup
        by rootlpno: replace followup=admdate[_n+`i'] if reverse==1 & revseq[_n+`i']==1
}
```

Which method should I use? If you are a SAS user and still coming to terms with syntax, the vector method is best as upside-down syntax in SAS is difficult. Stata users may wish to go straight to explicit subscripting because it is an excellent technique to master. SPSS users have a ready choice: the vector file method is very easy, but the upside down method develops a deeper level of skill that can be very useful in more complex situations.

Overview of Exercise 4

Step 1. Open vaslinks file.

Step 2. Create status and time for vasectomy.

Step 3. Identify and assign censoring dates

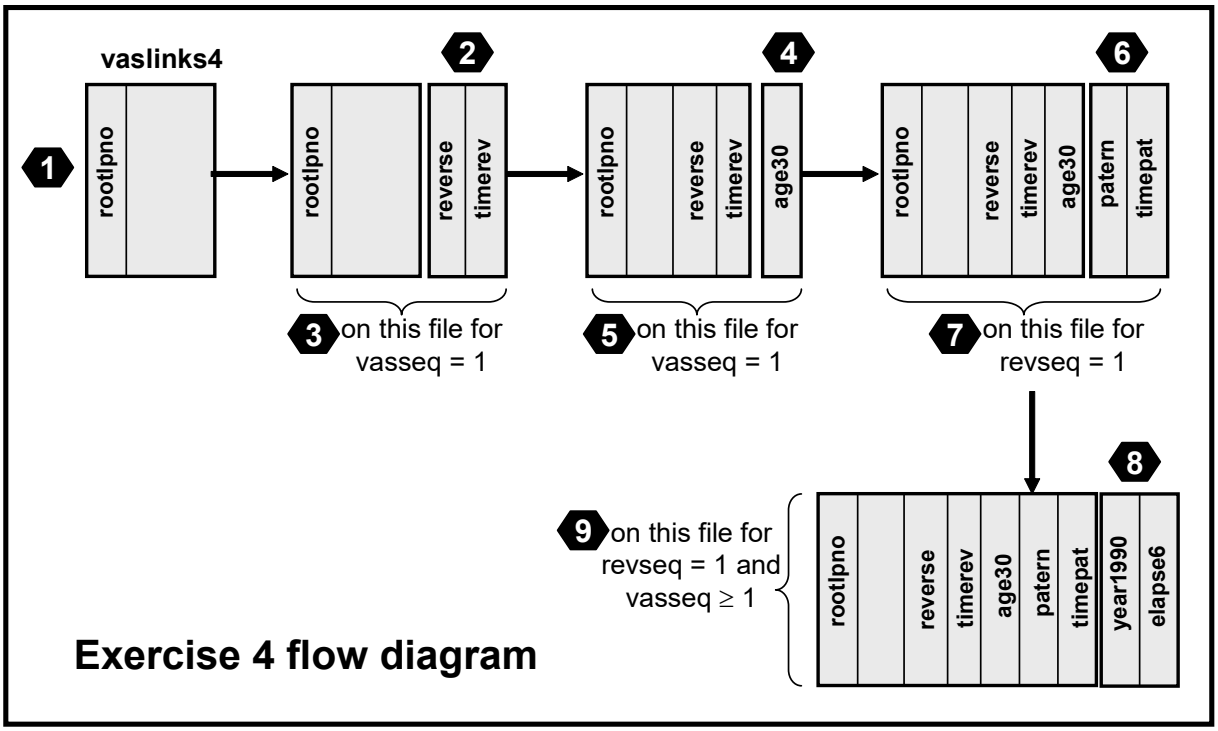
Step 4. Actuarial survival analysis for reversal for men <30 vs 30+ years.

Step 5. Cox regression analysis for reversal by age.

Step 6. Create status and time fields for paternity after reversal analysis

Step 7. Actuarial survival analysis for success fraction of paternity following reversal.

Step 8. Multivariate Cox regression for paternity (by age and time).



Exercise Instructions

In this exercise, you will be using the previously created vaslinks practice data set to measure the cumulative incidence of different outcomes of health care.

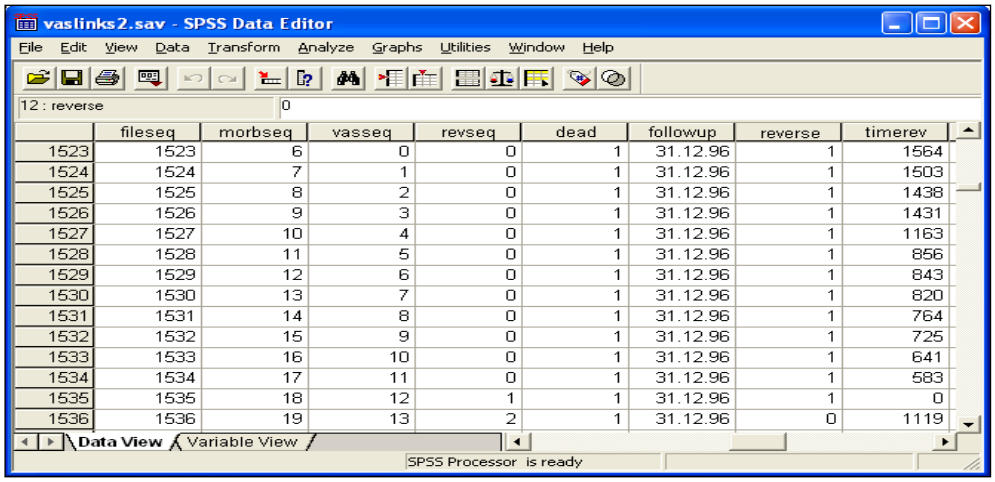
1. Open the vaslinks4 data file as your working file and re-familiarise yourself with its structure. Note that due to the nature of the research questions addressed in this exercise, inter-hospital transfers are not of consequence and may be ignored.

2. The first health care outcome you are required to measure is the cumulative incidence of vasovasostomy in men who had an in-hospital vasectomy. To perform this analysis, you need to create the following two survival variables with values loaded onto each vasectomy index record:
   (i) a 0-1 binary indicator, reverse, set to 1 if the man had a subsequent vasovasostomy by 31 December 1996; and
   (ii) the time in days, timerev, from the date of admission for vasectomy to one of three dates as set out below:

Circumstances	End date to calculate **timerev**	
Men with vasovasostomy on or before 31/12/1996	Admission date for vasovasostomy	
Men with no vasovasostomy by 31/12/1996, who died on or before 31/12/1996	Date of death	(ie, the default **followup** date)
Men with no vasovasostomy or death by 31/12/1996	31/12/1996	

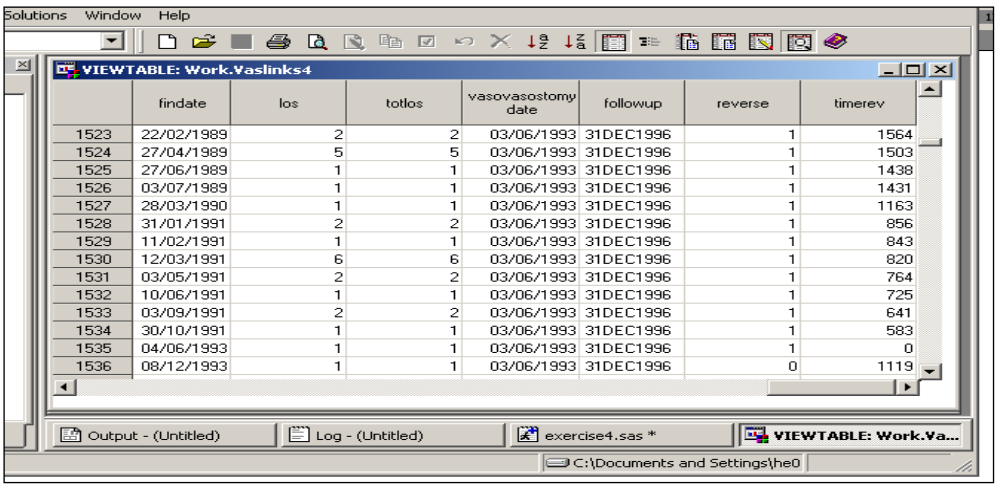
If your syntax works, your file could look like the following, although note that these screenshots are based on an upside-down method developed in SPSS and replicated in SAS and Stata:

SPSS



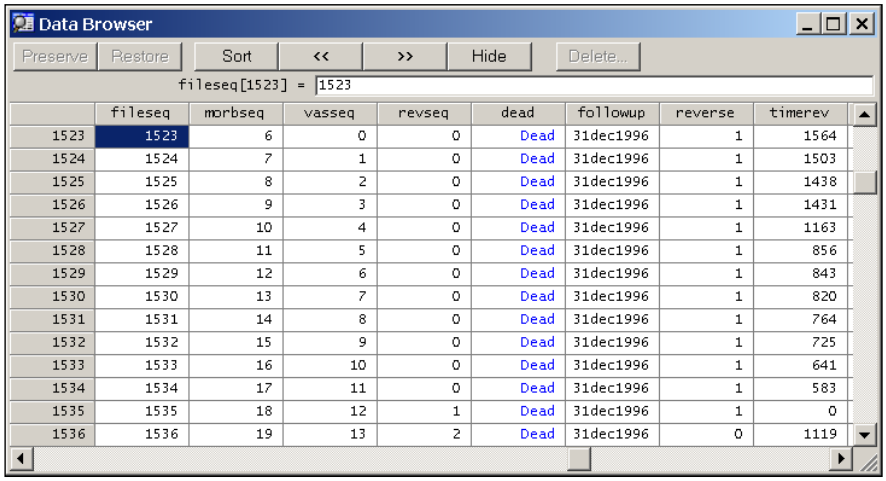
	fileseq	morbseq	vasseq	revseq	dead	followup	reverse	timerev
1523	1523	6	0	0	1	31.12.96	1	1564
1524	1524	7	1	0	1	31.12.96	1	1503
1525	1525	8	2	0	1	31.12.96	1	1438
1526	1526	9	3	0	1	31.12.96	1	1431
1527	1527	10	4	0	1	31.12.96	1	1163
1528	1528	11	5	0	1	31.12.96	1	856
1529	1529	12	6	0	1	31.12.96	1	843
1530	1530	13	7	0	1	31.12.96	1	820
1531	1531	14	8	0	1	31.12.96	1	764
1532	1532	15	9	0	1	31.12.96	1	725
1533	1533	16	10	0	1	31.12.96	1	641
1534	1534	17	11	0	1	31.12.96	1	583
1535	1535	18	12	1	1	31.12.96	1	0
1536	1536	19	13	2	1	31.12.96	0	1119

SAS



	findate	los	totlos	vasovasostomy date	followup	reverse	timerev
1523	22/02/1989	2	2	03/06/1993	31DEC1996	1	1564
1524	27/04/1989	5	5	03/06/1993	31DEC1996	1	1503
1525	27/06/1989	1	1	03/06/1993	31DEC1996	1	1438
1526	03/07/1989	1	1	03/06/1993	31DEC1996	1	1431
1527	28/03/1990	1	1	03/06/1993	31DEC1996	1	1163
1528	31/01/1991	2	2	03/06/1993	31DEC1996	1	856
1529	11/02/1991	1	1	03/06/1993	31DEC1996	1	843
1530	12/03/1991	6	6	03/06/1993	31DEC1996	1	820
1531	03/05/1991	2	2	03/06/1993	31DEC1996	1	764
1532	10/06/1991	1	1	03/06/1993	31DEC1996	1	725
1533	03/09/1991	2	2	03/06/1993	31DEC1996	1	641
1534	30/10/1991	1	1	03/06/1993	31DEC1996	1	583
1535	04/06/1993	1	1	03/06/1993	31DEC1996	1	0
1536	08/12/1993	1	1	03/06/1993	31DEC1996	0	1119

Stata



Data Browser

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

fileseq[1523] = 1523

	fileseq	morbseq	vasseq	revseq	dead	followup	reverse	timerev
1523	1523	6	0	0	Dead	31dec1996	1	1564
1524	1524	7	1	0	Dead	31dec1996	1	1503
1525	1525	8	2	0	Dead	31dec1996	1	1438
1526	1526	9	3	0	Dead	31dec1996	1	1431
1527	1527	10	4	0	Dead	31dec1996	1	1163
1528	1528	11	5	0	Dead	31dec1996	1	856
1529	1529	12	6	0	Dead	31dec1996	1	843
1530	1530	13	7	0	Dead	31dec1996	1	820
1531	1531	14	8	0	Dead	31dec1996	1	764
1532	1532	15	9	0	Dead	31dec1996	1	725
1533	1533	16	10	0	Dead	31dec1996	1	641
1534	1534	17	11	0	Dead	31dec1996	1	583
1535	1535	18	12	1	Dead	31dec1996	1	0
1536	1536	19	13	2	Dead	31dec1996	0	1119

Which method should I use? Regardless of which method you use, you should first create a followup variable as the applicable censoring date in the absence of an outcome event. Then …

HINT for SPSS users: You have a ready choice: the vector file method is easy, but the upside down method develops a deeper level of skill that can be useful in more complex situations. If you go for the hard method, follow the logic set out in the training session and divide your syntax into two distinct parts for the reverse status variable and the timerev time variable.

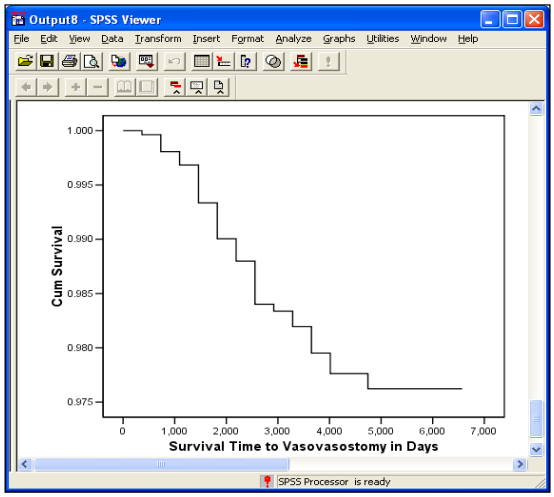
HINT for SAS users: If you are still coming to terms with syntax, the vector method is best as upside-down syntax in SAS is difficult. You create a vector file containing the rootlpno and the date of the first vasovasostomy. These data items can then be loaded back onto the working file using merge and are thus propagated onto all records. Then derive the reverse status variable, and the timerev time variable for the index record.

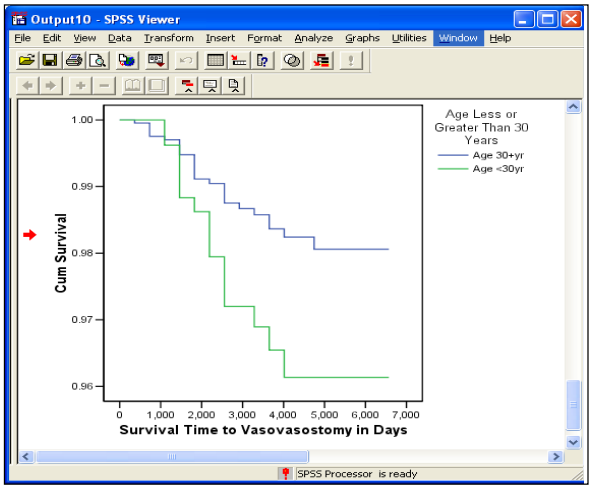
HINT for Stata users: Stata users may wish to go straight to explicit subscripting because it is an excellent technique to master. Alternatively, take the easy path and use a vector file.

Hint for R users: Vector files are particularly useful for calculations of linked health variables in R because alternative methods that automatically iterate through row-by-row calculations according to some native file processing order are not inherent within R nor its default set of functions. This method therefore avoids the inefficient looping through rows.

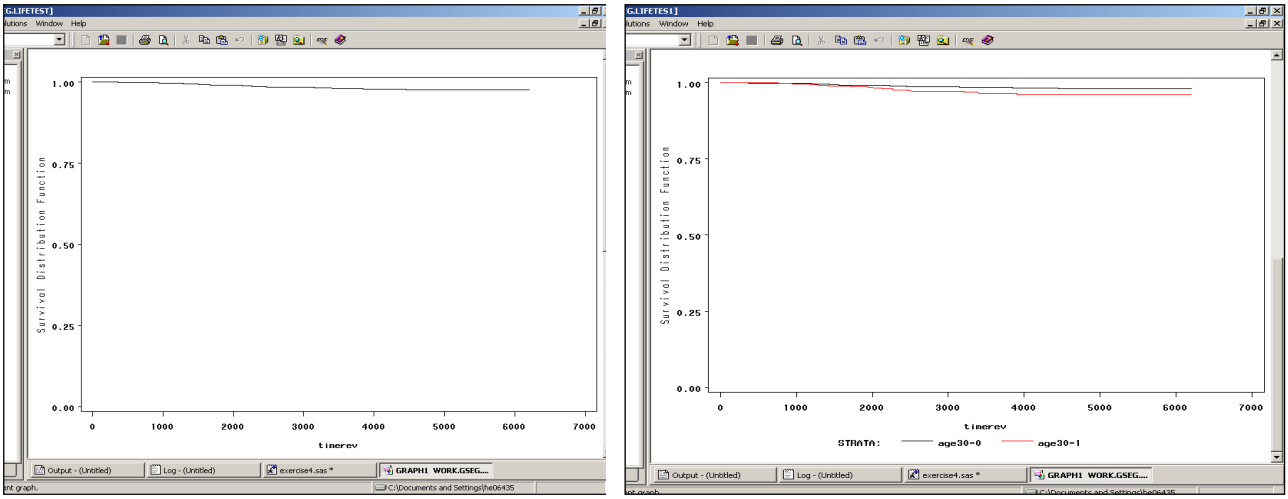
3. Select the vasectomy index cases and perform an actuarial survival (life table) analysis with reverse as the status variable and timerev as the time variable, divided into 365 day intervals. The survival curve should be like the one below on the left. What is the estimate of the risk of a reversal by 10 years after vasectomy?

SPSS

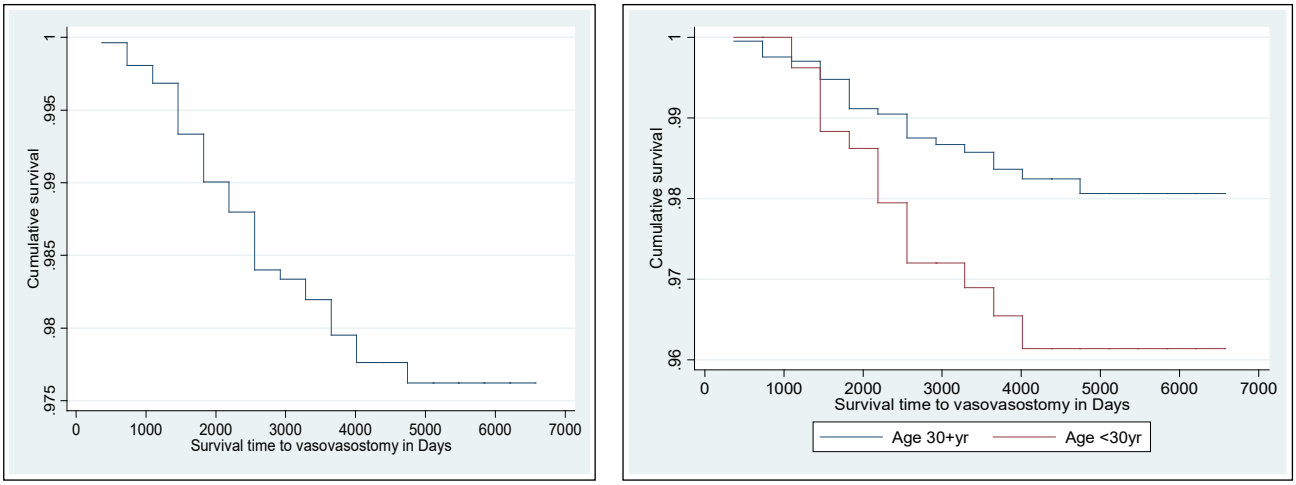




SAS



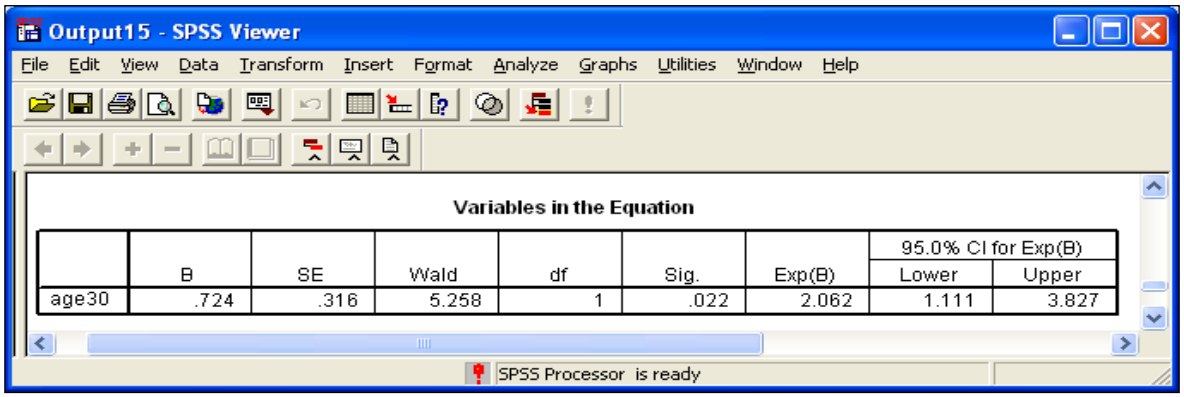
Stata



4. Create an age30 binary indicator for age group, set to 0 for age ≥ 30 years and 1 for ages < 30 years.

5. Perform the life table analysis again, using age30 as a factor to compare the risk of reversal by 10 years between younger and older men (see results above on the right). Now repeat the analysis, this time using a Cox regression model with timerev as the time variable, reverse as the status variable and age30 as the single covariate. Request 95% CL's for the relative risk if not already a default in your package. How do you interpret this result (which should be the one shown below)? Note that the Wald statistic in Stata is reported as a z-score (normal distribution), which has been squared in the SPSS output to follow a chi-square distribution.

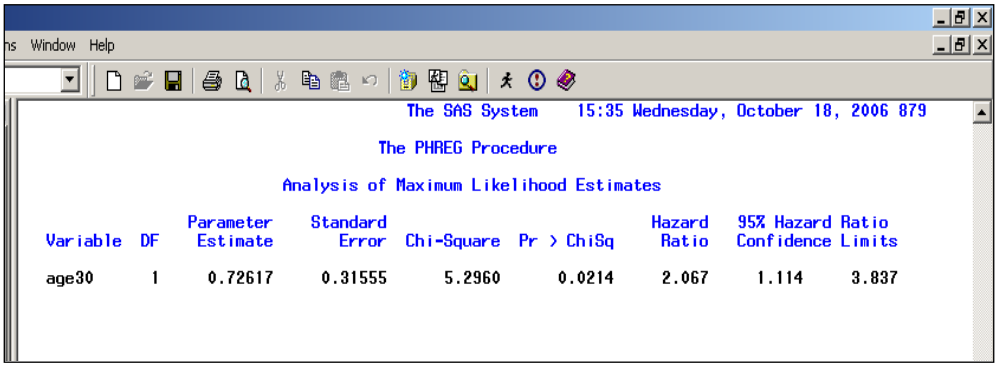
SPSS



Variables in the Equation

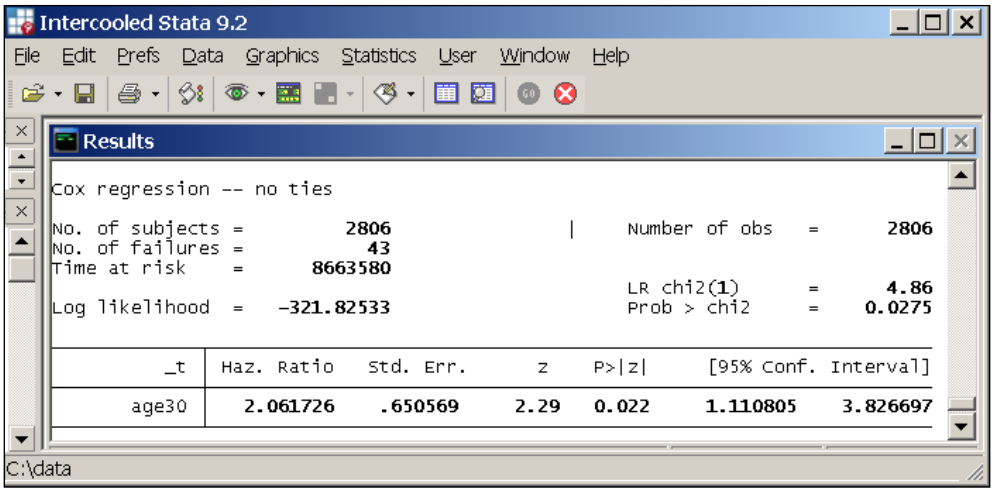
	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B) Lower	Upper
age30	.724	.316	5.258	1	.022	2.062	1.111	3.827

SAS



```
The SAS System    15:35 Wednesday, October 18, 2006 879

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

age30      1        0.72617     0.31555      5.2960      0.0214   2.067      1.114     3.837
```

Stata



```
Cox regression -- no ties

No. of subjects =        2806                 |    Number of obs    =      2806
No. of failures =          43
Time at risk    =     8663580
                                                   LR chi2(1)       =      4.86
Log likelihood  =   -321.82533                     Prob > chi2      =    0.0275

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age30 |   2.061726    .650569     2.29   0.022     1.110805    3.826697
------------------------------------------------------------------------------
```

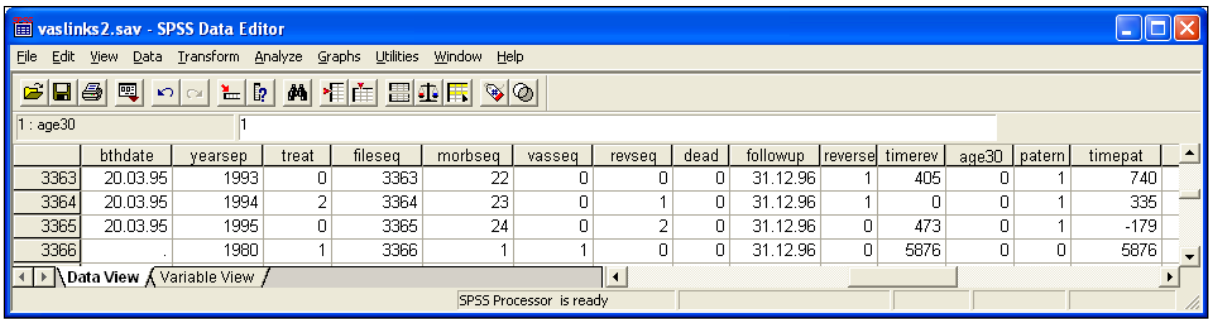
6. The second health care outcome for study in this exercise is the cumulative incidence of paternity on a birth certificate in men who had a vasovasostomy. The analysis is similar to that in question 3. You need to create two survival variables with values loaded onto each vasovasostomy index record:

   (i) a 0-1 binary indicator, patern, set to 1 if the man had fathered a child by 31 December 1996; and

   (ii) the time in days, timepat, from the date of admission for vasovasostomy to either the birth date, date of death or 31/12/96, using the same principles outlined in question (3).

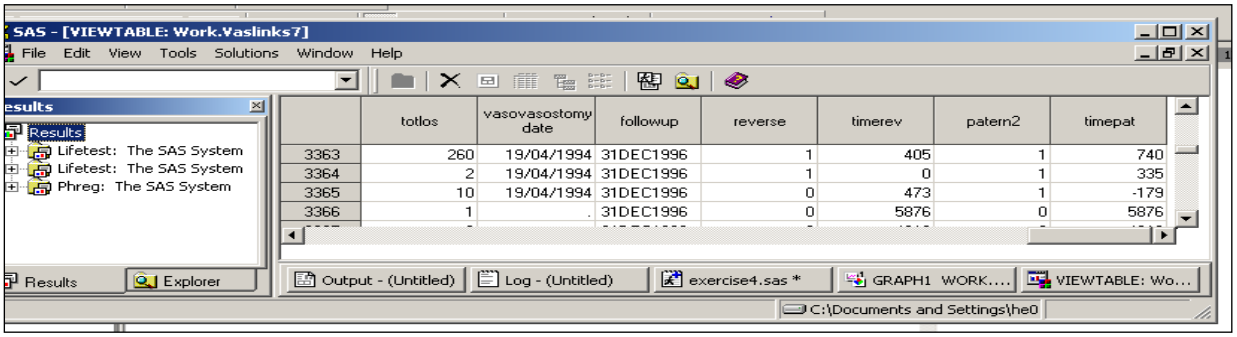
   Note that your task is made easy by the merging of birth information, where applicable, onto the EOR area of every record of men who fathered children post-vasovasostomy. Complex syntax and upside-down analysis are not required. You must, however, exclude any post-censoring birth of children after the end of the study period. Your file should look like the screen below.

SPSS



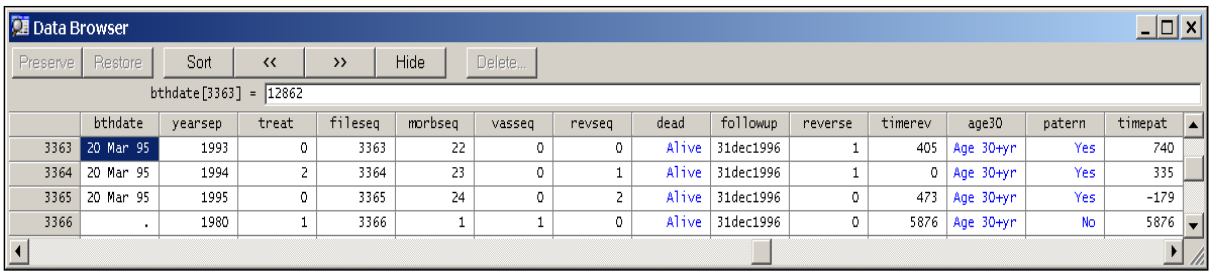
	bthdate	yearsep	treat	fileseq	morbseq	vasseq	revseq	dead	followup	reverse	timerev	age30	patern	timepat
3363	20.03.95	1993	0	3363	22	0	0	0	31.12.96	1	405	0	1	740
3364	20.03.95	1994	2	3364	23	0	1	0	31.12.96	1	0	0	1	335
3365	20.03.95	1995	0	3365	24	0	2	0	31.12.96	0	473	0	1	-179
3366	.	1980	1	3366	1	1	0	0	31.12.96	0	5876	0	0	5876

SAS



	totlos	vasovasostomy date	followup	reverse	timerev	patern2	timepat
3363	260	19/04/1994	31DEC1996	1	405	1	740
3364	2	19/04/1994	31DEC1996	1	0	1	335
3365	10	19/04/1994	31DEC1996	0	473	1	-179
3366	1	.	31DEC1996	0	5876	0	5876

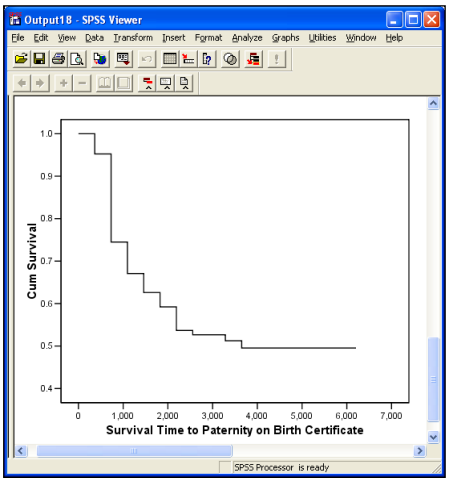
Stata

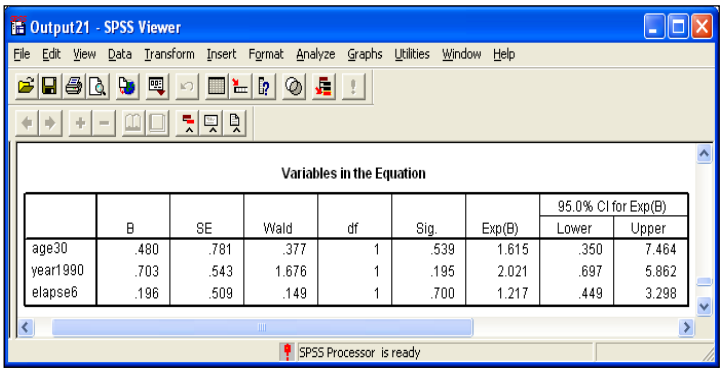


	bthdate	yearsep	treat	fileseq	morbseq	vasseq	revseq	dead	followup	reverse	timerev	age30	patern	timepat
3363	20 Mar 95	1993	0	3363	22	0	0	Alive	31dec1996	1	405	Age 30+yr	Yes	740
3364	20 Mar 95	1994	2	3364	23	0	1	Alive	31dec1996	1	0	Age 30+yr	Yes	335
3365	20 Mar 95	1995	0	3365	24	0	2	Alive	31dec1996	0	473	Age 30+yr	Yes	-179
3366	.	1980	1	3366	1	1	0	Alive	31dec1996	0	5876	Age 30+yr	No	5876

7. Select the vasovasostomy (not vasectomy) index records and perform a life table analysis to estimate the success fraction for paternity within 10 years of a vasovasostomy (see below on left).

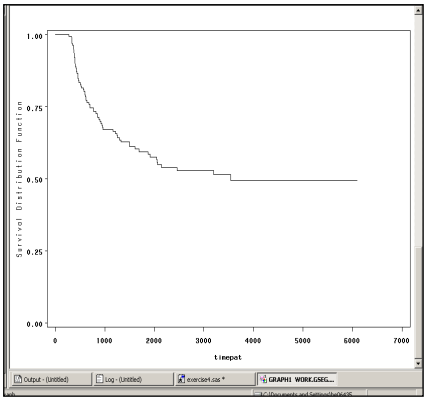
SPSS

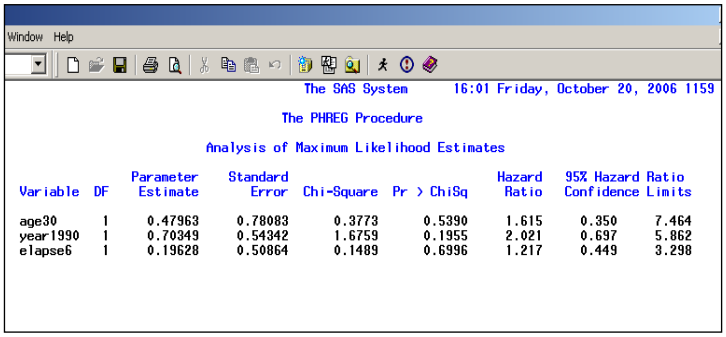




	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B) Lower	Upper
age30	.480	.781	.377	1	.539	1.615	.350	7.464
year1990	.703	.543	1.676	1	.195	2.021	.697	5.862
elapse6	.196	.509	.149	1	.700	1.217	.449	3.298

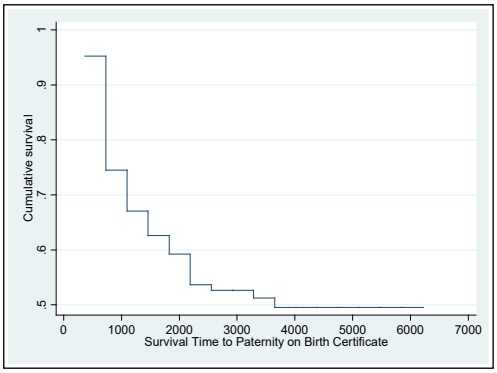
SAS

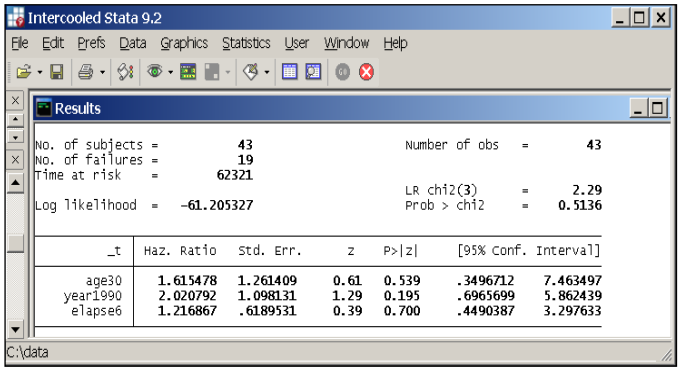




Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
age30	1	0.47963	0.78083	0.3773	0.5390	1.615	0.350	7.464
year1990	1	0.70349	0.54342	1.6759	0.1955	2.021	0.697	5.862
elapse6	1	0.19628	0.50864	0.1489	0.6996	1.217	0.449	3.298

Stata





8. Now create two new covariates: year1990 is set to 0 for separation years <1990 and set to 1 for separation years ≥ 1990. elapse6 is set to 1 for time since vasectomy <6 years and set to 0 for time since vasectomy ≥6 years (it should not matter exactly how you deal with missing data for vasovasostomies with no record of a previous vasectomy in the data set, because you will apply a filter to deal with this issue at step 9). The latter variable is easily calculated by propagating the index vasectomy admission date onto subsequent records and assessing the time elapsed between the vasectomy admission date and the vasovasostomy admission date.

9. Apply a filter to select on the 43 vasovasostomies with a record of a previous vasectomy in the data set (ie, revseq =1 and vasseq ≥1). Run a multivariate Cox regression model to show the independent effects of age at vasovasostomy, elapsed time since vasectomy and calendar year of vasovasostomy on the relative risk of paternity (see above on right). Also run the three covariates in three separate univariate models and comment on whether or not there is an empirical suggestion of confounding between the covariates.

Reference

Holman CDJ, Wisniewski ZS, Semmens JB, Rouse IL, Bass AJ. **Population-based outcomes after 28246 in-hospital vasectomies and 1902 vasovasostomies in Western Australia.** *BJU International* 2000; 86: 1043-1049.

REVIEW 4:
'OUTCOMES RESEARCH USING LINKED DATA'

Please check that you are familiar with the following topics:

- The health outcomes movement ☐
- Taxonomy of health outcome measures ☐
- Miettinen's elements of study design ☐
- Theory of experimental studies ☐
- When to perform non-experimental rather than experimental studies ☐
- Options for longitudinal research using a data linkage system ☐
- Theory of the cohort study ☐
- Theory of the case-control study (including nested case-control study) ☐
- Types of quasi-experimental study design ☐

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

- Health outcome ☐
- Effectiveness and efficacy ☐
- Safety and adverse event ☐
- Accessibility and appropriateness ☐
- Study domain, determinant, occurrence relation and study base ☐
- Time zero ☐
- Censoring date ☐
- Post-censoring event ☐

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

- Creation of a follow-up date ☐
- Creation of survival time and status variables for outcome data already on an index record ☐
- Calculation of survival time and status variables for an outcome on a later record ☐
- Correct application of censoring principles ☐
- Performance of a actuarial survival analysis, with interpretation of the life table and curve ☐
- Performance of a Cox regression analysis, with interpretation of the relative risks ☐

5. THEORY OF RISK ADJUSTMENT

5.1 Confounding and Effect Modification

Confounding

Confounding is ***mixing of effects***; it is the situation in which the effects of two processes are not separated. The distortion of the effect of an exposure on risk is brought about by the association with other factors that influence the outcome.

Example: The following data are concerned with the 18-month cumulative mortality following myocardial infarction according to treatment with a new drug and congestive heart failure status.

	Congestive Cardiac Failure					
	None		*Mild*		*Total*	
New Drug Prescribed	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*
Cardiac deaths	*11*	*5*	*4*	*10*	*15*	*15*
MI patients	*98*	*41*	*18*	*42*	*116*	*83*
Cumulative mortality	*11.2%*	*12.2%*	*22.2%*	*23.8%*	*12.9%*	*18.1%*
CMR	*0.92*		*0.93*		*0.72*	
CMD	*-1.0%*		*-1.6%*		*-5.2%*	

In the example, the absolute reduction in cumulative mortality in patients taking the new medication is considerably less than in those not taking it. However, the drug is prescribed more often in patients with no congestive cardiac failure and these patients have a much better outcome regardless of whether or not the drug is prescribed. Thus, the effect of the new drug is mixed up with (ie, confounded by) the prognostic effect of congestive cardiac failure. This is probably an example of ***confounding by indication***, as it is likely that patients without congestive cardiac failure were considered on clinical grounds to be better candidates for the new drug. The fact that confounding exists here is made obvious empirically by the much smaller reduction in risk associated with the prescription of the medication in either of those with or without congestive cardiac failure.

Effect Modification

Effect modification occurs when the effect of a putative causal factor is modified according to the level of exposure to another factor. However, it should be noted that the presence or absence of effect modification is dependent on the choice of a relative or absolute measure of effect. In the example above, there is essentially no modification of the CMR by the presence of mild congestive cardiac failure, whereas the CMD is distinctly different (modified) across the two strata of mild or no congestive cardiac failure.

Example: Here is an extreme example concerning the safety of a new treatment, where the principal concern is that the treatment relieves symptoms but causes early death. It appears innocuous until the results are stratified by age group:

	Age <50yr		*Age ≥50yr*		*Total*	
New Treatment	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*
Deaths	*90*	*20*	*110*	*380*	*200*	*400*
Patients	*1410*	*980*	*390*	*2620*	*1800*	*3600*
Cumulative mortality	*6.4%*	*2.0%*	*28.2%*	*14.5%*	*11.1%*	*11.1%*
CMR	*3.13*		*1.94*		*1.00*	
CMD	*+4.4%*		*+13.7%*		*0.0%*	

It then becomes clear that the new treatment is associated with an increased risk of death in both age groups and that this effect is modified according to whether the patient is <50yr or ≥50yr. This example

also illustrates the arbitrary nature of effect modification, because its presence and degree vary according to whether relative or absolute effect measures are being studied. Here, age <50yr is the strongest modifier on a relative scale, whereas age ≥50yr is the strongest modifier on an absolute scale.

The most common factors acting as either confounders or effect modifiers in health services research are sociodemographic factors, severity of illness (including complications) and comorbidity. The aim of ***risk adjustment*** is to control for the confounding effects of these variables on the associations between interventions and outcomes.

5.2 Risk Profiling and Risk Modelling in Health Services Research

Risk Profiling

Risk profiling is a description of the overall distribution of a population of patients; eg, those attending a given institution or selected for a particular treatment, according to one or more factors that are predictive of resource utilisation or health outcomes. Reasons for risk profiling include:

1. To demonstrate differences between patient populations.
2. To provide a basis for reimbursement by third party payers.

Risk Modelling

When comparing resource utilisation and clinical outcomes between groups of patients undergoing different interventions, the groups should be comparable with respect to *a priori* risk. Failure to assure homogeneity of patient populations can lead to confounding that may obscure the true relationship between treatment and outcome. With uncommon exceptions, health services researchers lack the luxury of selecting and allocating patients to policy or treatment protocols or to different providers, and therefore must compare groups with disparate and complex pre-treatment characteristics. Risk adjustment is a statistical procedure that adjusts for the association between one or more predictive factors and a resource utilisation or outcomes measure. Reasons for risk adjustment include:

1. To control for confounding, especially confounding by indication (= ***risk adjustment***).
2. To provide pre-procedural predictions of risk based on multiple factors (= ***predictive risk***).

Credibility with providers is an important motivation for risk adjustment. Providers are typically unresponsive to unadjusted results concerning health service performance and serious problems arise if results unadjusted for risk are acted upon as if they mean precisely what they seem to mean.

Objects of Risk Profiling or Modelling in Health Services Research

The object of risk profiling or modelling typically includes any of the following:

Unidimensional risk factors:

- ***Sociodemographic factors*** (eg, age, sex, race, social disadvantage, locational disadvantage, use of private health insurance).
- ***Severity of illness*** (eg, clinical staging, intensity of physiological disorder).
- ***Comorbidity*** (eg, Charlson index, Kaplan Index, Multipurpose Australian Comorbidity Scoring System).

Complex risk factors:

- ***Case mix*** (eg, diagnosis related groups, other prospective payment systems).
- ***Composite clinical indices*** (which combine two or more risk factors).

In this course we consider only the three unidimensional risk factors.

5.3 Sociodemographic Factors

The most important sociodemographic factors to consider in risk profiling and risk adjustment are age, sex, race, social disadvantage (also known as ***socio-economic status*** or SES), locational disadvantage (also known as ***remoteness***) and use of private health insurance (also known as ***payment classification***). Other sociodemographic factors that may be important in some studies include marital status, country of birth and occupation. Social and locational disadvantage are generally the most challenging of the sociodemographic factors to measure.

Areal Assignment of Social and Locational Disadvantage Measures

This is the most commonly used system of assigning social and locational disadvantage indices to administrative data sources. ***Postcodes*** can be transformed into indices or groupings of social disadvantage or grouping of locational disadvantage, which are usually published by national statistics bureaus or health departments. As shown by Hyndman and others, the use of postcode to assign disadvantage indices causes significant misclassification error relative to the use of the much smaller clusters of about 200 households. More accurate small area-based social or locatioinal disadvantage indices are sometimes available can be linked to health records if the residential addresses on records are available. The latter can then be geocoded into latitudes and longitudes, and mapped to small areas, which in turn can be used to attribute social and locational disadvantage indices at the small area level to each record. As we have seen, it is possible to pre-assign geocodes from residential addresses using computer-automated linkage procedures.

5.4 Severity of Illness

Severity of illness (also known as ***case severity***) is a measure of the intensity or gravity of a patient's illness, for a given condition or diagnosis. It can be defined from several different perspectives, including whether the severity is patient-assessed or provider-assessed; and whether severity is defined in terms of clinical stage of disease or the objective need for health care resources. For example, a patient with a terminal cancer may be at an advanced stage in terms of clinical severity of illness, but not require as many health care resources as a recently diagnosed patient with cancer.

Case severity measures based on primary data collection: When research is carried out with primary data collection in the health care setting, there are good prospects of recording valid measures of case severity. These are usually one of two types:

(i) ***Physiological case severity systems*** which tend to be fairly generic and are applied in particular health care setting such as intensive care (eg, APACHE = *Acute Physiology and Chronic Health Evaluation*), surgical care (eg, POSSUM = *Physiological and Operative Severity Score for enUmeration of Mortality and morbidity*) or obstetric delivery care (eg, APGAR = *Activity, Pulse, Grimace, Appearance, Respiration* score).

(ii) ***Clinical case severity systems*** which tend to be disease-specific. Examples including the different clinical staging systems for different cancers and a number of trauma severity scoring systems (eg, TRISS = *Trauma Injury Severity Score*).

Case severity measures based on administrative health data: In general, lack of adequate information on severity of illness is a limitation imposed on researchers who work with linked administrative health data. This is an evolving area and some of the possibilities include:

(i) ***Use of diagnostic and procedure codes to identify complications or advanced disease***: Several algorithms have been devised in an attempt to grade the severity of illness using diagnostic and sometimes procedure codes. Examples include the *Disease Staging* system, which uses ICD-9-CM codes and a computerised staging algorithm to classify hospital admissions into three stages:
Stage 1: conditions with no complications or problems of minimal severity.
Stage 2: conditions limited to an organ or system with increased risk of complications.
Stage 3: conditions with multiple site involvement, systemic involvement and poor prognosis.
Other systems using diagnostic and procedure codes in administrative data are the *Medisgroups* system and *Patient Management Categories* (PMCs).

(ii) ***Use of linked utilisation data as a surrogate for severity***: For example, using linked inpatient and outpatient hospital morbidity data, people with mental illness could be grouped into three severity categories: those ever admitted as an involuntary patient; those ever admitted but not as an involuntary patient; and those receiving outpatient care only.

(iii) ***Use of linked pharmaceutical data as a surrogate for severity***: For example, using linked pharmaceutical data, patients with diabetes mellitus could be grouped into three severity categories: those who use insulin; those who use oral hypoglaecaemic agents but not insulin; and those who use neither insulin nor oral hypoglycaemics.

Complications

Complications are not comorbidity in that they arise from the natural history of the index condition or its treatment rather than from a separate pathogenesis. When a complication arises from the natural history of disease it should usually be regarded as part of the continuum of severity of illness and is thus an object of risk adjustment. Complications arising from treatment are a form of adverse event and therefore an outcome event to do with clinical safety. Thus it is important to distinguish prior complications of disease, already in existence at the start of an episode of care to be evaluated, from those that arise in the course of an episode of care. Risk adjustment is generally required for prior complications, but not for prospective complications, which should be construed as a part of the outcome of the episode of care. In practice, the distinction can be sometimes difficult, especially when analysis is based on administrative data. Systems exist for identification of complications, such as the *Complications Screening Program* of Iezzoni and others.

5.5 Comorbidity

Comorbidity refers to one or more diseases that coexist in a study participant in addition to the index condition that is the object of study. Comorbidity is part of the overall burden of illness for a patient, which consists of the following:

1. The primary reason for care (principal diagnosis).
2. The severity of the principal diagnosis.
3. Complications that result from the diagnosis or treatment.
4. Unimportant comorbidities that have a trivial impact on resource use and outcomes.
5. Important comorbidities that increase the intensity of resource use and affect outcomes.

Categories (2) and (5) are the ones that typically necessitate risk adjustment and comorbidity scoring systems are particularly designed to measure the burden of illness arising from category (5).

The existence of more than one illness in the same individual has important effects on the use of the health care system. ***Berkson's bias*** is the tendency for individuals in the community with more than one symptom or disorder to seek treatment at a higher rate than those with just one symptom or disorder. As a result, cases defined by disease status or some aspect of their service use are more likely to have

comorbidity with other disorders than are cases which might be identified through a population-based survey. For example, Hobbs and others in the Oxford Record Linkage Study found that patients with colon cancer had been treated recently for diverticular disease in excess of the rate in the general population; likewise, patients with brain tumours had been treated recently for certain conditions of the central nervous system at higher rates than the general population. This is one of the bases by which the presence of comorbidity may confound comparisons of utilisation rates between different patient groups. Comorbidity may bias comparisons of outcomes measures between different patient groups through confounding by indication.

Charlson's Index

While several comorbidity scales have proven useful for both patient classification in clinical research and prognostication in medical care, Charlson's comorbidity index, and its Deyo and Dartmouth-Manitoba adaptations for use with administrative data, are the most widely adopted systems. Charlson's method consists of a weighted index based on the following 17 comorbid conditions that predict one-year survival.

1. Myocardial infarction.
2. Congestive cardiac failure.
3. Peripheral vascular disease.
4. Cerebrovascular disease.
5. Dementia.
6. Chronic pulmonary disease.
7. Rheumatologic disease.
8. Peptic ulcer disease.
9. Mild liver disease.
10. Diabetes (mild to moderate).
11. Diabetes with chronic complications.
12. Hemiplegia or paraplegia.
13. Renal disease.
14. Any malignancy, including lymphoma and leukaemia.
15. Moderate or severe liver disease.
16. Metastatic solid tumour.
17. Acquired immune deficiency syndrome.

The index was derived from a training population of 559 medical inpatients at a New York hospital, and was validated on a testing population of 685 breast cancer patients at another hospital. Subsequently, other researchers have refined the Charlson index with:

- additional comorbid categories;
- a clearer distinction between complications and comorbid conditions in the index admission; and
- replacement of the original fixed integer scores with empirically-derived, study-specific weights;

but essentially the Charlson index has remained the same conceptual entity that it was when first published in 1987.

Multipurpose Australian Comorbidity Scoring System (MACSS)

Following a process of extensive screening for comorbid conditions that had effects on one-year mortality, 30-day readmissions or LOS in a large training population of 693,037 inpatients in WA, MACSS was developed based on the most frequent 102 comorbid conditions associated with a RR≥1.1 of death or readmission or a LOS difference ≥0.5 days.

The performance of the MACSS and the Charlson index in predicting mortality, readmission and LOS, and in controlling confounding of effect measures by comorbidity, was compared by WA researchers in five test scenarios involving asthma, myocardial infarction, mastectomy, transurethral prostatectomy and major depressive illness. The MACSS performed better than the Charlson index on all three outcomes in all five clinical groups. It reduced the failure of the Charlson index to discriminate on mortality and re-admission outcomes by 5-40%, improved R^2 in LOS models by up to four-fold and often doubled the correction of originally confounded effect measures. The use of the MACSS and similar alternatives to the Charlson index are a new methodological standard for adjustment of comorbidity risk. The list of 102 MACSS conditions is shown overleaf.

Comorbid conditions included in MACSS

Tuberculosis Septicaemia Other bacterial infections HIV or AIDS Candidiasis Systemic mycosis **Any infectious or parasitic disease** Lung cancer Breast cancer Other primary cancer Lymphoma or leukaemia Secondary cancer of lymph nodes Other secondary cancer Benign neoplasms of the colon or rectum Uterine leiomyoma **Any neoplasms** Hypothyroidism Diabetes mellitus Complications of diabetes Malnutrition or cachexia Avitaminosis Gout Fluid or electrolyte disorders Obesity Other metabolic disorders Immune system disorders **Any endocrine, metabolic or immune disease** Anaemias Coagulopathies **Any blood disease** Dementia or senility Alcoholism Illicit drug use Schizophrenia	Affective psychosis Other non-transient psychoses Non-psychotic depression Neuroses or personality disorders Mental retardation Incontinence **Any mental disorder** Inflammatory CNS diseases Parkinson's disease Other extrapyramidal disease Degenerative neurological or muscular diseases Paraplegia or quadraplegia Infantile cerebral palsy Epilepsy Glaucoma Cataract Blindness and low vision Hearing loss **Any disease of the nervous system or sense organs** Cerebrovascular disease or stroke Hypertension Congestive cardiac failure Myocardial infarction Other ischaemic heart disease Chronic pulmonary circulatory disorders Valvular disease Cardiac dysrhythmias or arrest Peripheral vascular disease Haemorrhoids **Any diseases of the circulatory system**	Pneumonia Chronic pulmonary disease Asthma Pleurisy Other respiratory disorders **Any diseases of the respiratory system** Oesophageal diseases Peptic ulcer Gastritis or duodenitis Cholelithiasis and cholecystitis Chronic liver disease Sequelae of chronic liver disease Abdominal hernia Non-infectious digestive disorders Diverticular disease Peritoneal adhesions Other diseases of intestine Gastrointestinal haemorrhage **Any diseases of the digestive system** Chronic renal disease Other kidney, ureter or bladder conditions Urinary tract infection Benign prostatic hypertrophy Breast disorders Pelvic inflammatory disease Cervicitis or vaginitis Genital prolapse Non-inflammatory ovarian conditions Uterine disorders Gynaecological pain Menstrual disorders Menopausal disorders Female infertility **Any diseases of the genitourinary system** All except normal delivery **All pregnancy, childbirth and the puerperium**	Chronic skin ulcer **Any diseases of skin or subcutaneous tissue** Rheumatologic diseases Dorsopathies Other arthropathies or joint disorders Osteoporosis **Any diseases of the musculoskeletal system and connective tissue** **Congenital anomalies** Migraine and headache Symptoms involving alteration-consciousness, syncope, collapse, the cardiovascular system, respiratory system, chest and cyanosis Nausea and vomiting Abdominal pain Symptoms concerning nutrition, metabolism and development **Any symptoms, signs and ill-defined conditions** Late effects of injures **Any injury or poisoning** Potential health hazards related to communicable diseases Housing, household, economic, family or other psychosocial circumstances **Any factor influencing health status and contact with health services**

Optimal Length of Comorbidity Lookback Period

The ***lookback period*** is concerned with the period of time prior to an index admission that a search is conducted to adduce evidence of the presence of comorbid conditions that contribute to a comorbidity score such as Charlson or MACSS. Research by Preen and others using MACSS has examined the impact of different comorbidity lookback periods on modelling post-hospitalisation mortality and readmissions. The results may be summarised as the following guidelines:

- ***Mortality outcomes***: 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. Index admission co-morbidity (ie, essentially no lookback) performs almost as well as a one-year lookback period for mortality outcomes in surgical patients but not in medical patients.
- ***Readmission outcomes***: A five-year lookback period provides the best model-fit, and is therefore the best method of risk adjustment for comorbidity, when evaluating a readmission outcome. This applies equally in surgical and medical patients.

5.7 Methods of Risk Adjustment

Methods Used in Study Design

1. ***Restriction***: The selection of subjects is restricted to a single level of the confounder. For example, a comparison of outcomes after lumpectomy vs radical mastectomy for breast cancer could be restricted to women with clinical stage I disease at the time of surgery (an indicator of disease severity). Restriction may be the only practical approach to addressing health services study questions within patient groups with special needs, such as Aboriginal and Torres Strait Islander people. Another example is to control confounding by comorbidity by restriction of the analysis to patients with no comorbidity.
2. ***Matched comparison group***: Comparison subjects having different interventions are chosen to make the groups identical with respect to the presence of the confounding variable. For example, to every woman who underwent lumpectomy of her breast tumour there could be matched a woman with a tumour of the same clinical size, but who underwent radical mastectomy. Note that here 'matching' refers to a design option for cohort studies. Matching in case-control studies serves entirely different purposes and is not a method to control confounding; rather it improves efficiency (narrower confidence intervals) provided the matching factor is a confounder.
3. ***Case distribution studies***: These were mentioned in Theory Class 4. They represent a special form of matched comparison group, where subjects are compared with themselves. The case-crossover study and other case distribution studies are only of advantage with interventions that are variable over time and have a short latent period and a short risk period. Effects of pharmacotherapeutic agents are generally the most suitable for this type of design.
4. ***Instrumental variables***: These can reduce confounding and other biases in observation studies. In health services research, an instrumental variable is one or more observational factors that influence the type of treatment received, but do not directly affect patient outcomes; eg, differential distances to high and low intervention facilities; or occurrence in high and low intervention time periods (cf quasi-experimental designs). They mimic randomisation of patients to different likelihoods of receiving alternative treatments and thus, like randomisation, produce a sources of variation in treatment that is balanced with respect to confounders. Because the groups being compared differ only in the likelihood of treatment, unlike a division into pure treatment and pure control groups, the method can only estimate a marginal effect of treatment over the range of variation. Another method, akin to instrumental variables, is use of ***propensity scores***, which are multivariate summary scores of the likelihood of receiving different types of treatment.

Methods Used in Study Analysis

Analytic methods of risk adjustment artificially force the groups being compared to have the same status concerning the confounding variable. The adjusted results in the example of lumpectomy for breast cancer would be presented so as to answer the question: If the distribution of tumour size in the group that underwent lumpectomy had really been the same as that in the group that underwent radical mastectomy, what would have been the difference in breast cancer mortality?

1. ***Stratification***: These methods involve an initial calculation of a stratum-specific estimate of effect for each level of the confounding variable, from which a weighted average can be obtained. The most widely used method is precision-based weights (the weight for each stratum is the reciprocal of the variance).
2. ***Multivariate modelling***: These methods involve the use of regression models to adjust simultaneously for a number of confounders. The most common models and their applications to outcome measures are shown below. There is evidence that the more information available for risk adjustment, the better the resulting model – although the gain in predictive abilities of risk adjustment is not directly proportional to the quantity of data points provided for modelling.

Type of Regression Model	Form of Effect Measure	Examples of Applications to Measures in Health Services Research
Multiple linear regression	Mean difference per unit of exposure.	Health expenditure; units of blood; attendances; service costs.
Logistic regression	Odds ratio for count data with binary outcome data.	Coverage by insurance; coverage by procedure; occupancy
Poisson regression	Rate ratio for person-time data with binary outcome.	Multiple procedures; multiple admissions; incidence and mortality rates in dynamic population.
Cox regression	Rate ratio for time to failure data with binary outcome.	First readmission; disease recurrence; revisional procedures; case fatality.

5.8 Advanced Methods of Risk Adjustment Using Multivariate Models

Statistical methods of risk adjustment are thus used in an effort to make valid comparisons, but the success of these methods depends on the completeness and accuracy of available data and the predictive performance of the statistical model. The following are some of the ways in which data accuracy and model performance can be improved in the case of comorbidity adjustment, although the principles apply also to other forms of risk adjustment.

Methods to Improve Completeness and Accuracy of Data

1. Increase the number of cormorbid conditions based on their effects in a large representative clinical cohort.
2. Select comorbid conditions separately in surgical, medical, psychiatric and other patient subgroups.
3. Select comorbid conditions separately for each of a range of outcomes for mortality, morbidity and resource utilisation.
4. Distinguish between prior, co-existent and subsequent comorbidity and complications.
5. Subclassify comorbid conditions by severity, recency and duration.
6. Identify and use the most accurate sources of information based on validation studies of hospital charts, primary medical care notes and administrative health data.

Methods to Improve Performance of the Statistical Model

1. Provide locally accurate empirical weights derived from the data set under analysis.
2. Use non-linear transformations in statistical models, including fractional polynomials.
3. Allow for clustering at zero by inclusion of a binary indicator for no comorbidity.
4. Use modelling techniques that are responsive to interactions, including classification regression, recursive partitioning and neural network analysis.
5. Perform analyses separately in different patient subgroups.

Example: In work using the WA Linked Database, published in Urology International, we studied mortality risks in 19,598 men after surgery for benign prostatic hypertrophy. Closed prostatectomy (TURP) had a higher post-operative mortality rate than open prostatectomy, but this relationship was confounded by higher comorbidity in the TURP group, such that after adjustment with a Charlson index with empirically-derived weights, the RR fell from 1.20 (95%CI 1.08-1.34) to 1.10 (0.99-1.23). However, a further significant improvement in goodness of fit of the Cox proportional hazards model and a further reduction in the RR from 1.10 to 1.07 (0.95-1.19) was achieved simply with the inclusion of fractional polynomials to account for a non-linear relationships of the log(RR) with age and comorbidity. Our suspicion is that, at 1.07, the empirical relationship of TURP with post-operative mortality is still affected by residual confounding, a large part of which could be removed by further improvements in the measurement and control for differences in comorbidity.

References

Australian Bureau of Statistics. **Information Paper. 1991 Census. Socio-Economic Indexes for Areas.** ABS Cat. 2912.0. Canberra: ABS; 1994.

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

Brailer DJ, Kroch E, Pauly MV, *et al*. **Comorbidity-adjusted complication risk. A new outcome quality measure.** *Medical Care* 1996; 34: 490-505.

Charlson M, Szatrowski TP, Peterson J, Gold J. **Validation of a combined comorbidity index.** *J Clin Epidemiol* 1994; 47: 1245-51.

Charlson ME, Pompei P, Ales KL, *et al*. **A new method of classifying prognostic comorbidity in longitudinal studies: development and validation.** *J Chron Dis* 1987; 40: 373-83.

Cleves MA, Sanchez N, Draheim M. **Evaluation of two competing methods for calculating Charlson's comorbidity index when analysing short-term mortality using administrative data.** *J Clin Epidemiol* 1997; 50: 903-8.

D'Hoore W, Sicotte C, Tilquin C. **Risk adjustment in outcome assessment: the Charlson Co-morbidity Index.** *Meth Inform Med* 1993; 32: 382-7.

Deyo RA, Cherkin DC, Ciol MA. **Adapting a clinical co-morbidity index for use with ICD-9-CM administrative databases.** *J Clin Epidemiol* 1992; 45: 613-9.

Einarsdottir K, Preen DB, Holman CDJ, Emery J. **Regular Primary Care Plays a Significant Role in Secondary Prevention of Ischemic Heart Disease in a Western Australian Cohort.** *Journal of General Internal Medicine* 2011; 26(10): 1092-1097.

Einarsdottir K, Preen DB, Holman CDJ, Emery J. Kelman C. **Regular Primary Care Lowers Hospitalisation Risk and Mortality in Seniors with Chronic Respiratory Diseases.** *Journal of General Internal Medicine* 2010; 25(8): 766-773.

Elixhauser A, Steiner C, Harris DR, *et al*. **Comorbidity measures for use with administrative data.** Med Care 1998; 36: 8-27.

Fetter RB, Shin Y, Freeman JL, *et al*. **Case-mix definitions by diagnosis-related groups.** *Med Care* 1980; 18 (Suppl): 1-53.

Ghali WA, Hall RE, Rosen AK, *et al*. **Searching for an improved clinical comorbidity index for use with ICD-9-CM administrative data.** *J Clin Epidemiol* 1996; 49: 273-8.

Gonnella JS, Louis DZ, Gozum M. **Disease Staging Clinical Criteria.** Ann Arbor: Michigan: The MEDSTAT Group, 1997.

Harse JD, Holman CDJ. **Charlson's Index was not predictive of quality of life outcome scores in a study of patients after joint replacement surgery.** *J Clin Epidemiol* 2005; 58: 1142-1149.

Hobbs MST, Fairbairn AS, Acheson ED, *et al*. **The study of disease associations with linked records.** *Br J Prev Soc Med* 1976; 141: 30-44.

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*. **Mortality and prostate cancer risk in 19,598 men after surgery for benign prostatic hyperplasia.** *Br J Urol Int* 1999; 84: 37-42.

Hyndman JCG, Holman CDJ, Hockey RL, *et al*. **Misclassification of social disadvantage based on geographic areas: comparison of postcode and collector's district analyses.** *Int J Epidemiol* 1995; 24: 165-176.

Iezzoni LI, Ash AS, Coffman GA, *et al*. **Admission and mid-stay MedisGroups scores as predictors of hospitalisation charges.** *Med Care* 1991; 29: 210.

Iezzoni LI, Ash AS, Coffman GA, *et al*. **Predicting in-hospital mortality. A comparison of severity measurement approaches.** *Med Care* 1992; 30: 347-359.

Iezzoni LI, Daley JD, Heeren T, *et al*. **Identifying complications of care using administrative data.** *Med Care* 1994; 32: 700-715.

Imbens GW, Angrist JD. **Identification and estimation of local average treatment effects.** *Econometrica* 1994; 62: 467-475.

Jones HJS, de Cossart L. **Risk scoring in surgical patients.** *Br J Surgery* 1999; 86: 149-157.

Kaplan MH, Feinstein AR. **The importance of classifying initial co-morbidity in evaluating the outcome of diabetes mellitus.** *J Chron Dis* 1974; 27: 387-404.

Kieszak SM, Flanders WD, Kosinski AS, *et al*. **A comparison of the Charlson comorbidity index derived from medical records data and administrative billing data.** *J Clin Epidemiol* 1999; 52: 137-142.

Korn EL, Simon R. **Measure of explained variation for survival data**. *Stat Med* 1990; 9: 487-503.

Last JM (ed). **A Dictionary of Epidemiology.** New York: Oxford University Press, 1995.

Leblanc M, Crowley J. **Survival trees by goodness of split.** *J Amer Stat Assoc* 1993; 88: 457-67.

Lucas FL, Wennberg DE, Malenka DJ, *et al*. **The Charlson Comorbidity Index as a predictor of 30-day mortality following carotid endarterectomy.** *Am J Epidemiol* 1997; 145: A257.

Malenka DJ, McLerran D, Roos N, *et al*. **Using administrative data to describe casemix: a comparison with the medical record.** *J Clin Epidemiol* 1994; 47: 1027-1032.

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

McKee M, Petticrew M. **Disease staging: A casemix system for purchasers?** *J Pub Health Med* 1993; 15: 25.

Newschaffer CJ, Bush TL, Penberthy LT. **Comorbidity measurement in elderly female breast cancer patients with administrative and medical records data.** *J Clin Epidemiol* 1997; 50: 725-733.

Newton KM, Wagner EH, Ramsey SD, *et al*. **The use of automated data to identify complications and comorbidities of diabetes: a validation study.** *J Clin. Epidemiol* 1999; 52: 199-207.

Normand ST, Morris CN, Fung KS, *et al*. **Development and validation of a claims based index for adjusting for risk of mortality: the case of acute myocardial infarction.** *J Clin Epidemiol* 1995; 48: 229-43.

Petryshen P, Pallas LLO, Shamian J. **Outcome monitoring: adjusting for risk factors, severity of illness and complexity of care.** *JAMA* 1995; 2: 243-9.

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.

Robertson C, Boyle P, Hsieh C-C, *et al*. **Some statistical considerations in the analysis of case-control studies when exposure variables are continuous measurements.** *Epidemiology* 1994; 5: 164-70.

Romano PS, Mark DH. **Bias in the coding of hospital discharge data and its implications for quality assessment.** *Med Care* 1994; 32: 81-90.

Romano PS, Roos LL, Jollis JG. **Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives.** *J Clin Epidemiol* 1993; 46: 1075-9.

Roos LL, Stranc L, James RC, *et al*. **Complications, comorbidities, and mortality: improving classification and prediction.** *Hlth Services Res* 1997; 32: 229-42.

Rosenbaum PR, Rubin DB. **Reducing bias in observational studies using subclassification on the propensity score.** *J Am Stat Assoc* 1984; 79: 516-524.

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

Royston P, Altman GD. **Regression using fractional polynomials.** *App Statistics* 1994; 43: 429-67.

Royston P, Ambler G, Sauerbrei W. **The use of fractional polynomials to model continuous risk variables in epidemiology.** *Int J Epidemiol* 1999; 28: 964-974.

Satariano WA, Ragland DR. **The effect of comorbidity on 3-year survival of women with primary breast cancer.** *Ann Intern Med* 1994; 120: 104-110.

Shwartz M, Iezzoni LI, Moskowitz MA, *et al*. **The importance of comorbidities in explaining differences in patient costs.** *Med Care* 1996; 34: 767-82.

Stukenborg GJ, Wagner DP, Connors AF. **Comparisons of the performance of two comorbidity measures with and without information from prior hospitalisations.** *Med Care* 2001; 39: 727-739.

Zhang M, Holman CDJ, Price S, Sanfilippo FM, Preen D, Bulsara M. **Comorbidity predicts repeat hospitalisation for adverse drug reactions in the elderly: retrospective cohort study of 28,548 patients.** *British Medical Journal* 2009; 338: 145-158.

TRAINING SESSION 5: RISK ADJUSTMENT WITH LINKED DATA

IMPORTANT NOTICE

The practice data sets provided for training purposes in this course are synthetic. All results based on analysis of these synthetic/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

For the final training session of the introductory course you will emulate a classic analysis in the field of health services research. Among 15 evaluations of surgical outcomes of benign prostatic hypertrophy undertaken in the last two decades of the 20th Century, four large population-based studies found a consistently higher long-term mortality when the 'closed' procedure of transurethral resection of the prostate (TURP) was compared with open prostatectomy (OP). Whether this apparent effect on late mortality was due to patient selection, the procedure itself or other factors remained controversial, with different groups, such as the Oxford Record Linkage Study and the Manitoba Centre for Health Policy and Evaluation publishing different conclusions. The major point at issue was the role of comorbidity – whether men who had co-prevalent illness, which increased their risk of long-term mortality, were more likely to be selected for TURP due to its lower short-term anaesthetic and peri-operative risks.

We also use this final exercise to revise some key syntax skills and to demonstrate the re-use of old syntax from your stored library, which can be modified for application in a new study.

Training Session Resources

Your Course USB pendrive includes the following save files for use in this training session:

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

The file contains hospital admissions with death information loaded on the end of each record. An individual's records, identified by the same rootlpno, are blocked together in chronological order of separation date. The records are those of any person admitted to a WA hospital in 1980-95 with mention of surgery for benign prostatic hypertrophy:

Also in the EOR loading area is a variable, charlson, which is a comorbidity score based on a modified Charlson index with empirically-derived weights for mortality from a larger data set. The presence of each of 21 Charlson-like conditions was assessed on the index hospital separation (ie, the one where the man had his first prostate surgery) and any prior hospital separations overlapping the 12-month period before the index admission date. The value of charlson was calculated as follows:

$$\text{charlson} = \sum_{i=1}^{17} X_i W_i$$

where X_i was a binary indicator for the presence of the ith Charlson condition and W_i was the weight assigned to that condition.

Note that the following procedure codes are relevant to this exercise ('*X*' is a *wild* character, which may represent 0-9):

Procedure	Coding System	Codes
TURP (closed prostatectomy)	ICPM (pre-1988)	5-601
	ICD-9-CM (1988 to June 1999)	60.2*X*
OP (open prostatectomy)	ICPM (pre-1988)	5-602 to 5-605
	ICD-9-CM (1988 to June 1999)	60.3*X* to 60.6*X*

Adjusting for Non-Linearity in the Effects of Continuous Covariates

Traditionally, when continuous covariates, such as age in years, are known to have a non-linear effect within a regression equation, such as in a Cox regression equation, the standard approaches to correcting the problem have been either (i) to categorise age into age groups, (ii) to transform age with a function such as ln(age) that renders its transformed effect to a more linear shape, or (iii) to add an age^2 term to provide a polynomial function consisting of age + age^2.

Fractional polynomials 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 ('m') determines the number and form of the possible terms. For m=2 there are 21 possible combinations of powers ($\beta_1 x^{p1} + \beta_2 x^{p2}$) where p = {-1, -½, 0, ½, 1, 2}, subject to some complex rules:

- $x^p = x^p$ if $p \neq (0)$.
- $x^p = \ln(x)$ if $p=(0)$.
- for (p1, p2) where p1=p2=p, we use $x^p + x^p \ln(x)$.
- (0,0) defines $\ln(x) + \ln(x)^2$.

Some statistical packages (eg, fracpoly in Stata) advise the user on which m=1 and m=2 models offer the best fit. Below are the 21 forms of m=2 fractional polynomials with the terms selected for use in the real WA study of outcomes following surgery for benign prostatic hypertrophy.

	-1	-½	0	½	1	2
-1	$x^{-1} + x^{-1}\ln x$	$x^{-1} + x^{-½}$	$x^{-1} + \ln x$	$x^{-1} + x^{-½}$	$x^{-1} + x$	$x^{-1} + x^2$
-½		$x^{-½} + x^{-½}\ln x$	$x^{-½} + \ln x$	$x^{-½} + x^{-½}$	$x^{-½} + x$	$x^{-½} + x^2$
0			$\ln x + \ln x^2$	$\ln x + x^{-½}$	$\ln x + x$	$\ln x + x^2$
½				$x^{½} + x^{½}\ln x$	$x^{½} + x$	$x^{½} + x^2$
1					$x + x\ln x$	$x + x^2$
2						$x^2 + x^2\ln x$

Charlson comorbidity score

Box-Tidwell transformation

Age

Note that the two fractional polynomial terms are placed in the model as two separate covariates and replace the original continuous covariate. The (1, 1) FP is known as the *Box-Tidwell transformation*, and the xln(x) term in $y = \beta_0 + \beta_1 x + \beta_2 x\ln(x)$ may be used as a general test of non-linearity.

Overview of Exercise 5

Step 1. Open the bphsurgery file.

Step 2. Tag records with TURP or OP.

Step 3. Create sequence variables (file, morb and index).

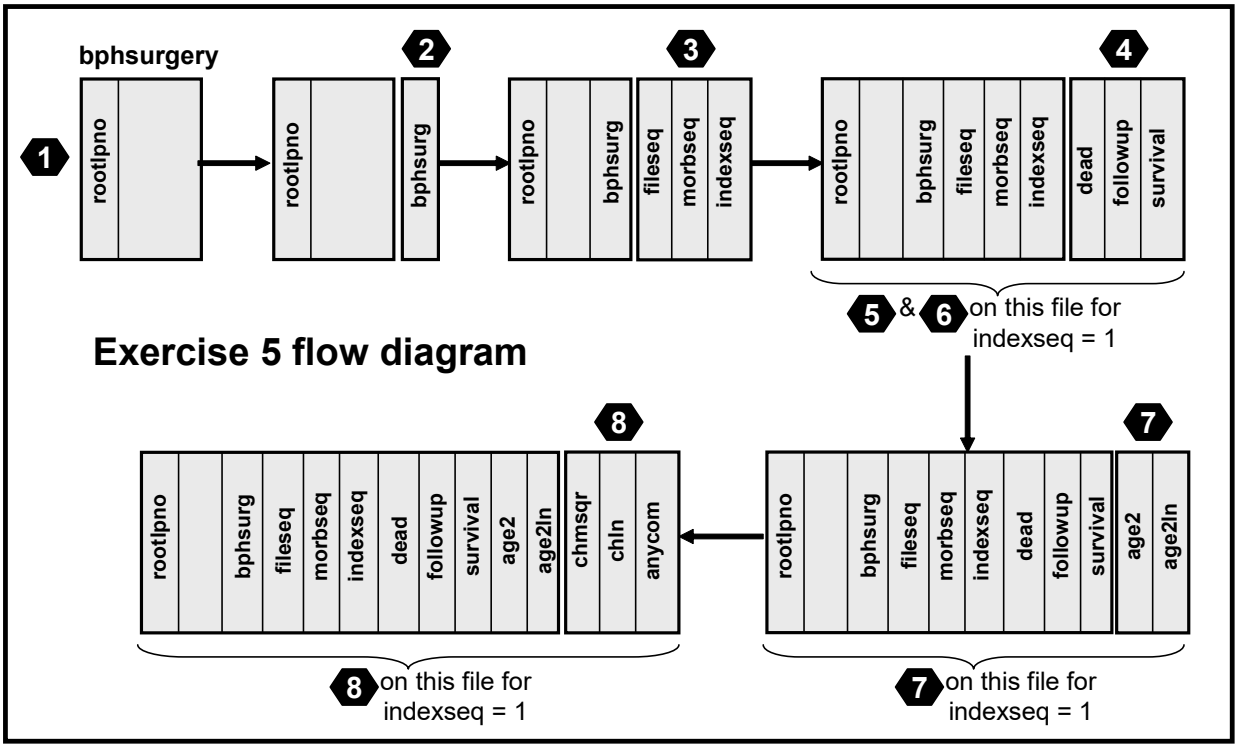
Step 4. Create status and time variables to death and censoring variable.

Step 5. Actuarial survival and unadjusted Cox regression analysis comparing TURP to OP.

Step 6. Examine confounding by age.

Step 7. Risk adjust for age (transformed) and calendar time.

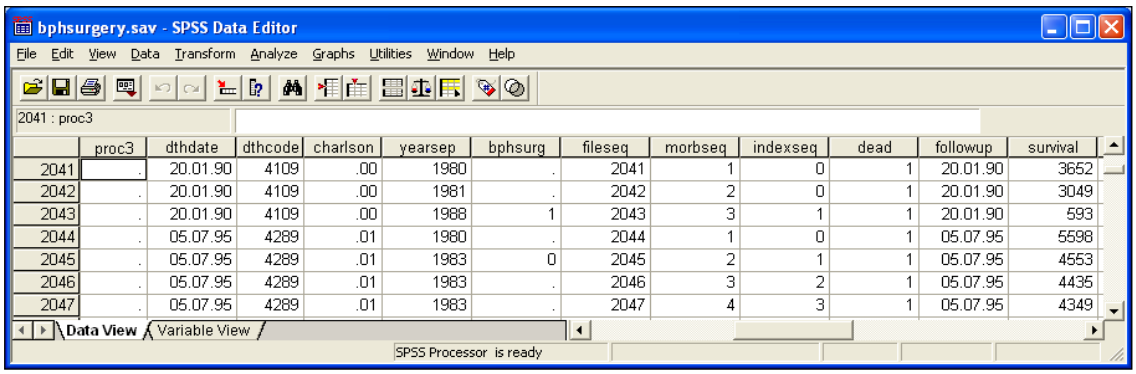
Step 8. Risk adjust for age, calendar time and comorbidity (transformed)



Exercise Instructions

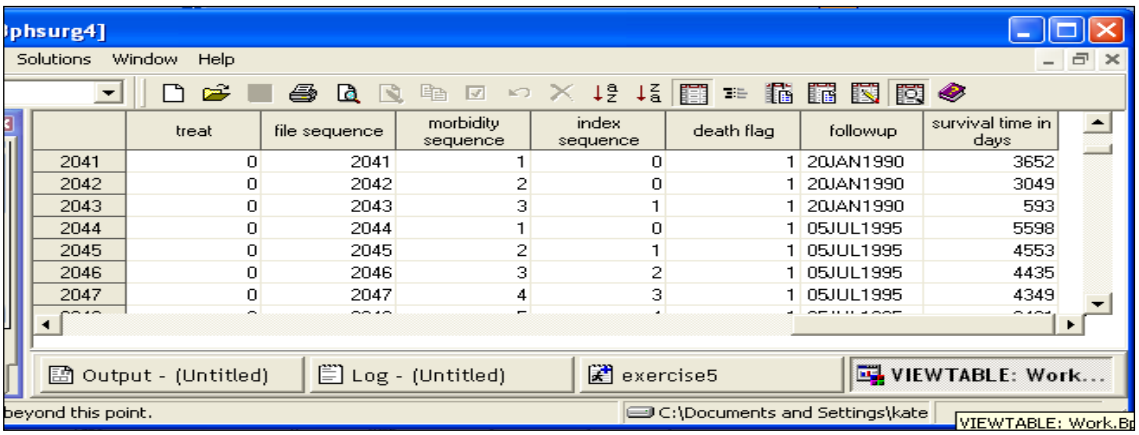
1. Open the bphsurgery file as your working file and familiarise yourself with its structure. Note the death information and the modified charlson comorbidity scores that have already been appended in the EOR loading area.
2. Create a variable, bphsurg, for each record, which is set to 1 for mention of a TURP, set to 0 for mention of an OP, but is otherwise left blank. To do this, you will also need to compose and execute syntax that will create a new variable called yearsep, derived from sepdate using a date function to extract the year.
3. Compose the fileseq, morbseq and indexseq variables. The index record is defined as the first record to mention either a TURP or OP. Note that due to the nature of the research questions addressed in this exercise, inter-hospital transfers are not of consequence and may be ignored. Note that indexseq may be set to a default of 0 or blank as it does not matter in this exercise.
4. For each index record you need to assign a status variable for the outcome event of death (= dead) and a survival time (= survival). The censoring date for followup for deaths was 31 December 1996. Thus any deaths after 31 December 1996 must be ignored and those not yet deceased must be censored on that date. At this point your file should look something like this:

SPSS



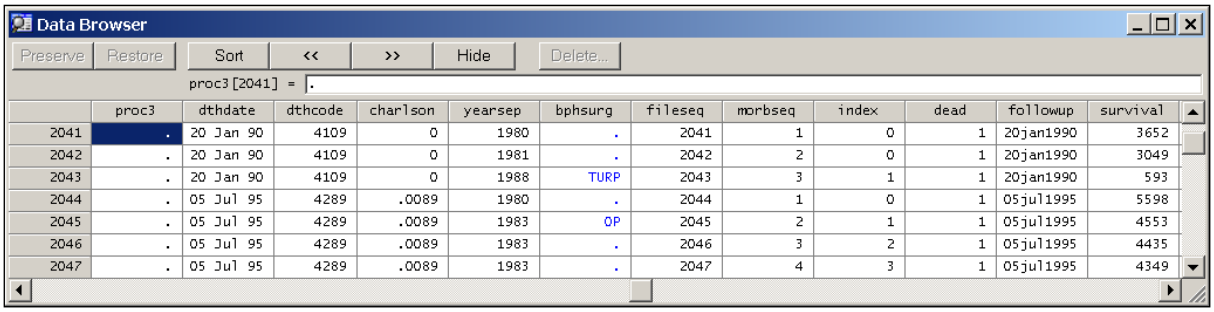
	proc3	dthdate	dthcode	charlson	yearsep	bphsurg	fileseq	morbseq	indexseq	dead	followup	survival
2041	.	20.01.90	4109	.00	1980	.	2041	1	0	1	20.01.90	3652
2042	.	20.01.90	4109	.00	1981	.	2042	2	0	1	20.01.90	3049
2043	.	20.01.90	4109	.00	1988	1	2043	3	1	1	20.01.90	593
2044	.	05.07.95	4289	.01	1980	.	2044	1	0	1	05.07.95	5598
2045	.	05.07.95	4289	.01	1983	0	2045	2	1	1	05.07.95	4553
2046	.	05.07.95	4289	.01	1983	.	2046	3	2	1	05.07.95	4435
2047	.	05.07.95	4289	.01	1983	.	2047	4	3	1	05.07.95	4349

SAS



	treat	file sequence	morbidity sequence	index sequence	death flag	followup	survival time in days
2041	0	2041	1	0	1	20JAN1990	3652
2042	0	2042	2	0	1	20JAN1990	3049
2043	0	2043	3	1	1	20JAN1990	593
2044	0	2044	1	0	1	05JUL1995	5598
2045	0	2045	2	1	1	05JUL1995	4553
2046	0	2046	3	2	1	05JUL1995	4435
2047	0	2047	4	3	1	05JUL1995	4349

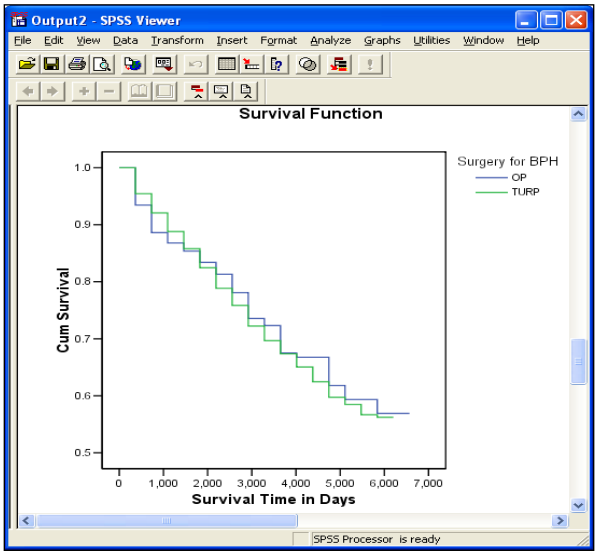
Stata

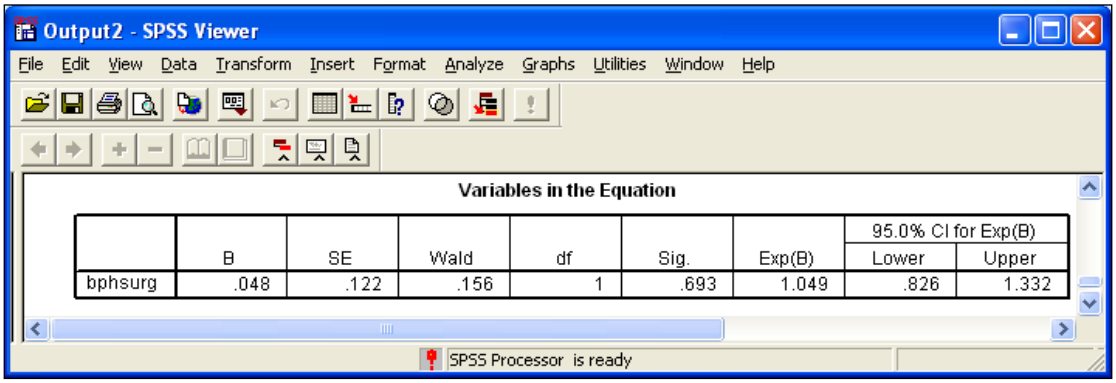


	proc3	dthdate	dthcode	charlson	yearsep	bphsurg	fileseq	morbseq	index	dead	followup	survival
2041	.	20 Jan 90	4109	0	1980	.	2041	1	0	1	20jan1990	3652
2042	.	20 Jan 90	4109	0	1981	.	2042	2	0	1	20jan1990	3049
2043	.	20 Jan 90	4109	0	1988	TURP	2043	3	1	1	20jan1990	593
2044	.	05 Jul 95	4289	.0089	1980	.	2044	1	0	1	05jul1995	5598
2045	.	05 Jul 95	4289	.0089	1983	OP	2045	2	1	1	05jul1995	4553
2046	.	05 Jul 95	4289	.0089	1983	.	2046	3	2	1	05jul1995	4435
2047	.	05 Jul 95	4289	.0089	1983	.	2047	4	3	1	05jul1995	4349

5. Select the index cases and run an actuarial survival analysis of post-operative mortality, using one year intervals, and comparing the survivals of TURP and OP patients. Next run a univariate Cox regression analysis of post-operative mortality with bphsurg as the single 'covariate'. Your results should be similar to those below:

SPSS

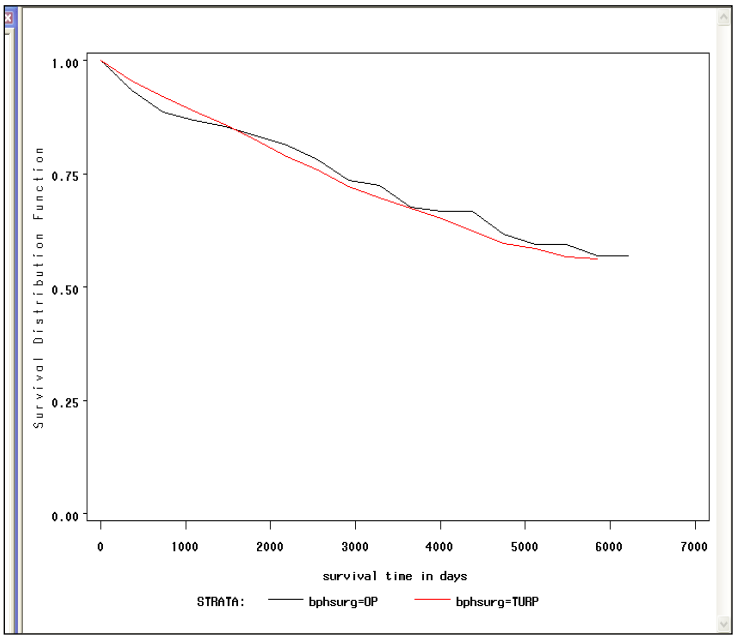




Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
bphsurg	.048	.122	.156	1	.693	1.049	.826	1.332

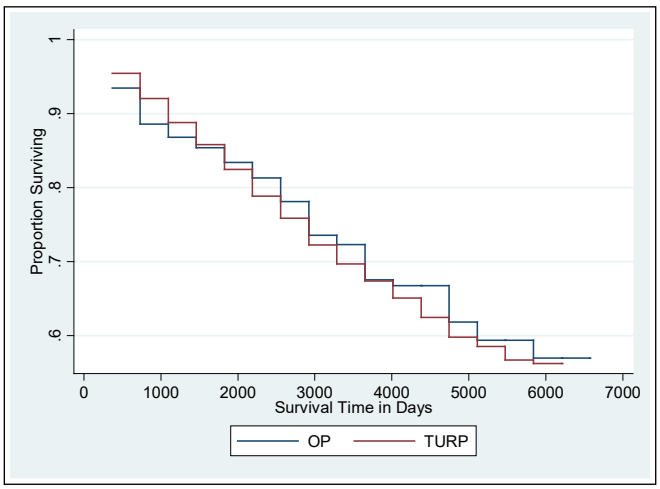
SAS

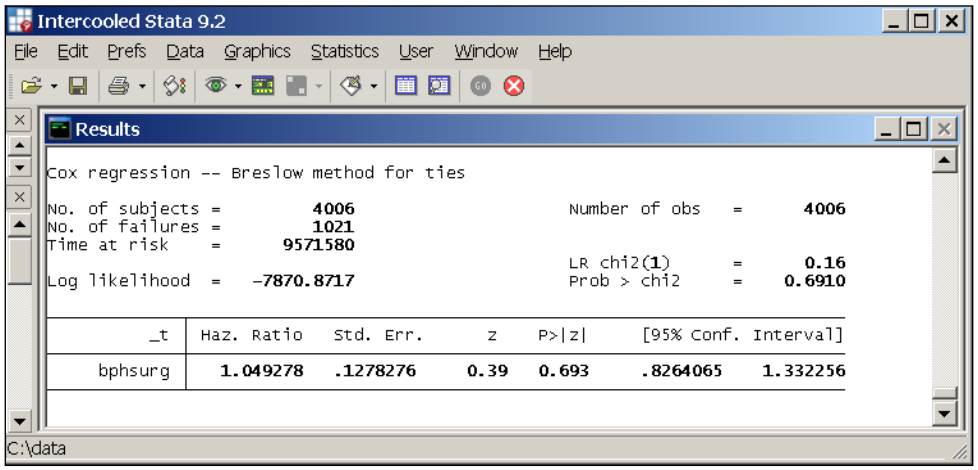


Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Variable Label
bphsurg	1	0.04714	0.12177	0.1499	0.6987	1.048	0.826	1.331	Surgery for BPH

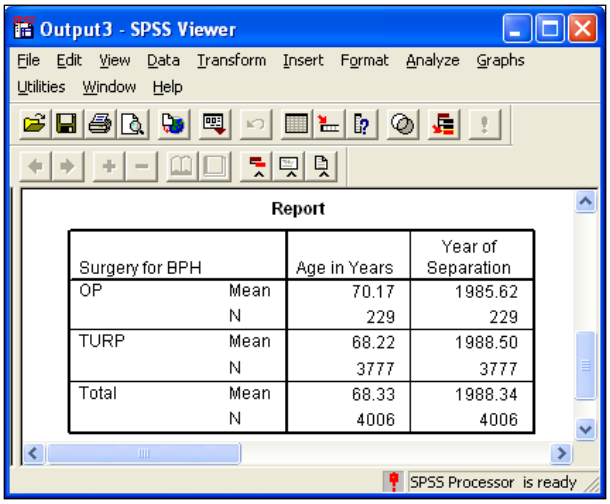
Stata





6. Compare the mean ages and the mean calendar years of separation of index cases undergoing TURP and OP (see below). What does this tell you about the potential for confounding?

SPSS



Surgery for BPH		Age in Years	Year of Separation
OP	Mean	70.17	1985.62
	N	229	229
TURP	Mean	68.22	1988.50
	N	3777	3777
Total	Mean	68.33	1988.34
	N	4006	4006

SAS

```
                    The SAS System                    14:30

        Obs    bphsurg      mean      nobs

         1        0        70.1659     229
         2        1        68.2203    3777

                    The SAS System                    14:30

        Obs    bphsurg      mean      nobs

         1        0        1985.62     229
         2        1        1988.50    3777

                    The SAS System                    14:30

                Obs      mean      nobs

                 1     1988.34     4006

                    The SAS System                    14:30

                Obs      mean      nobs

                 1     68.3315     4006
```

Stata

```
Results
. tab bphsurg if index==1, summarize(age)

Surgery for |      Summary of age in years
        BPH |        Mean   Std. Dev.       Freq.
------------+------------------------------------
         OP |   70.165939   7.7197895        229
       TURP |   68.220281   8.9276566       3777
------------+------------------------------------
      Total |   68.331503   8.8736978       4006
```

```
Results
. tab bphsurg if index == 1, summarize(yearsep)

Surgery for |     Summary of Year of Separation
        BPH |        Mean   Std. Dev.       Freq.
------------+------------------------------------
         OP |   1985.6201   4.1958352        229
       TURP |   1988.4999   4.4900505       3777
------------+------------------------------------
      Total |   1988.3352   4.5229547       4006
```

7. Adjust for age and calendar year as continuous covariates in the Cox regression analysis. Assume that the best fit for calendar year is obtained from a simple linear term and the best fit for age is obtained from a fractional polynomial of valence (2, 2), ie, [$age^2 + age^2 \ln(age)$]. Your results should be equivalent to those below. How does the mortality rate ratio comparing TURP to OP from this adjusted multivariate Cox regression model compare with that obtained from the univariate model in question 5?

SPSS

Output4 - 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
bphsurg	.262	.123	4.559	1	.033	1.300	1.022	1.654
yearsep	-.059	.009	46.556	1	.000	.943	.927	.959
age2	.003	.002	3.578	1	.059	1.003	1.000	1.006
age2ln	-.001	.000	2.399	1	.121	.999	.999	1.000

SPSS Processor is ready

SAS

```
                                   The SAS System          15:02 Sunday, October 22, 2006

                                   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 Variable Label

bphsurg   1         0.26236     0.12288     4.5589     0.0327      1.300     1.022     1.654 Surgery for BPH
yearsep   1        -0.05857     0.00858    46.5560     <.0001      0.943     0.927     0.959
age2      1         0.00300     0.00159     3.5779     0.0586      1.003     1.000     1.006
age2ln    1      -0.0005137   0.0003316     2.3993     0.1214      0.999     0.999     1.000
```

Stata

```
Intercooled Stata 9.2
File  Edit  Prefs  Data  Graphics  Statistics  User  Window  Help

Results
Cox regression -- Breslow method for ties

No. of subjects =        4006                    Number of obs   =      4006
No. of failures =        1021
Time at risk    =     9571580
                                                 LR chi2(4)      =    486.52
Log likelihood  =   -7627.6898                   Prob > chi2     =    0.0000

          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+------------------------------------------------------------
     bphsurg |   1.299998   .1597407     2.14   0.033      1.02176    1.654003
     yearsep |   .9431156   .0080952    -6.82   0.000      .927382     .959116
        age2 |   1.003008   .0015927     1.89   0.059     .9998914    1.006135
      age2ln |   .9994864   .0003315    -1.55   0.121      .998837    1.000136

C:\data
```

8. Now the modified Charlson Index should be added to your Cox regression model as a fractional polynomial with a valence of (-1/2, 0) to provide the best fit. **Useful tip**: Add a small positive value, eg, 0.01, to a covariate with zero values prior to a log or reciprocal transformation to avoid undefined results. Create a covariate for the reciprocal of the square root of charlson and add this plus the natural log of charlson to the regression model (ie, $x^{-½}$ + lnx). Also include a 0-1 binary variable that is set to 1 for the presence of any comorbidity (ie, charlson >0). Explain your results, which should mimic those below.

SPSS

Output8 - 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
bphsurg	.183	.123	2.221	1	.136	1.201	.944	1.528
yearsep	-.060	.009	48.171	1	.000	.942	.926	.958
age2	.002	.002	1.622	1	.203	1.002	.999	1.005
age2ln	.000	.000	.932	1	.334	1.000	.999	1.000
chmsqr	.548	.094	34.001	1	.000	1.730	1.439	2.080
chln	1.093	.137	63.343	1	.000	2.982	2.279	3.903
anycom	.816	.268	9.277	1	.002	2.261	1.338	3.822

SPSS Processor is ready

SAS

The SAS System 6
10:29 Wednesday, December 20, 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	
bphsurg	1	0.18312	0.12289	2.2206	0.1362	1.201	0.944	1.528
yearsep	1	-0.05957	0.00858	48.1708	<.0001	0.942	0.926	0.958
age2	1	0.00202	0.00159	1.6194	0.2032	1.002	0.999	1.005
age2ln	1	-0.0003201	0.0003321	0.9294	0.3350	1.000	0.999	1.000
chmsqr	1	0.54829	0.09403	34.0019	<.0001	1.730	1.439	2.080
chln	1	1.09265	0.13729	63.3448	<.0001	2.982	2.279	3.903
anycom	1	0.81584	0.26785	9.2772	0.0023	2.261	1.338	3.822

Stata

Intercooled Stata 8.0

File Edit Prefs Data Graphics Statistics User Window Help

Stata Results

Cox regression -- Breslow method for ties

No. of subjects = 4006 Number of obs = 4006
No. of failures = 1021
Time at risk = 9571580
LR chi2(7) = 716.46
Log likelihood = -7512.7206 Prob > chi2 = 0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
bphsurg	1.200956	.147579	1.49	0.136	.9439037	1.528012
yearsep	.942165	.0080871	-6.94	0.000	.9264471	.9581496
age2	1.002027	.0015928	1.27	0.203	.9989098	1.005153
age2ln	.9996795	.000332	-0.97	0.334	.999029	1.00033
chmsqr	1.730276	.1626936	5.83	0.000	1.43906	2.080423
chln	2.982134	.4094051	7.96	0.000	2.278605	3.902879
anycom	2.261076	.6056328	3.05	0.002	1.33758	3.822178

C:\DATA

References

Roos NP, Wennberg JE, Malenka DJ, *et al*. **Mortality and reoperation after open and transurethral resection of the prostate for benign prostatic hyperplasia**. *NEJM* 1989; 320**:** 1120-4.

Concato J, Horwitz RI, Feinstein AR, *et al*. **Problems of comorbidity in mortality after prostatectomy**. *JAMA* 1992; 267**:** 1077-82.

Seagroatt V. **Mortality after prostatectomy: selection and surgical approach**. *Lancet* 1995; 346**:** 1521-4.

Holman CDJ, Wisniewski ZS, Semmens JB, *et al*. **Mortality and prostate cancer risk in 19,598 men after surgery for benign prostatic hyperplasia.** *Br J Urol Int* 1999; 84: 37-42.

REVIEW 5:
'RISK ADJUSTMENT USING LINKED DATA'

Please check that you are familiar with the following topics:

- Confounding ☐
- Effect modification ☐
- Objects of risk adjustment (sociodemographics, case severity, comorbidity) ☐
- Methods of risk adjustment used in study design ☐
- Methods of risk adjustment used in study analysis ☐
- Multivariate modelling (multiple linear, logistic, Poisson and Cox regression) ☐
- Methods to improve risk adjustment ☐

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

- Risk profiling and risk adjustment ☐
- Comorbidity ☐
- Complications ☐
- Berkson's bias ☐
- Charlson's index and MACSS ☐
- Lookback period ☐
- Restriction ☐
- Stratification ☐
- Non-linearity ☐
- Fractional polynomial ☐

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

- Empirical assessment of confounding ☐
- Creation and use of fractional polynomial terms in regression models ☐
- Interpretation of variations in results from progression risk adjustment ☐