EPI 5344: SURVIVAL ANALYSIS IN THE HEALTH SCIENCES COURSE OUTLINE

COURSE INSTRUCTOR: N. Birkett,
Department of Epidemiology and Community Medicine,
Room 3230B,
451 Smyth Rd.,
Ottawa, Ontario. K1H 8M5
613-562-5800 ext 8289
613-562-5465 (**fax**)
e-mail: nbirkett@uottawa.ca

COURSE HOURS: Tuesdays, 0900-1200.

COURSE DATES: The course starts on March 2, 2010. The seven classes will be offered weekly with the last class on April 13, 2010. There will be no final examination.

CLASSROOMS: All classes will be held in room 3233 in the Health Sciences Building, 451 Smyth Rd.

OFFICE HOURS: I won't have formal office hours but am available to meet with students anytime that I am in my office and not in a meeting.

WEB SITE/FORUM: I have created a web site for the course. This site contains the full background information for the course. It also has the assignments, including the data for analysis. The URL is:
http://cancer-epidemiology.org/epi_5344_2010/.

PREREQUISITES:

1) A university-level biostatistics course such as EPI 5242 (Biostatistics 1). Students will be expected to be familiar with the fundamentals of hypothesis testing, confidence intervals, and multiple linear regression analysis modeling.
2) Completion of Epidemiology 1 (EPI 5240) or equivalent. Students will be expected to be familiar with confounding and effect modification and the change in coefficient definition of confounding. Prior exposure to the application of the Mantel-Hanzel test for confounding would be useful.
3) Students are strongly encouraged to have completed EPI 5340 (Epidemiological Methods).
4) Students are encouraged to take EPI 5345 (Applied Logistic Regression) in the first half of the winter semester.

5) Familiarity with the use of computers for statistical analysis is strongly recommended. The course assignments will require the student to perform computer-based analysis. Data sets will be provided in SAS format.

NOTE: None of the prerequisite material will be taught during the class. Rather, the class will build on the prerequisite material to demonstrate the application of advanced statistical methods to address core epidemiological issues.

INTRODUCTION

This course provides an advanced (second level) examination of the methods of the analysis of data where the outcome is the time to an event. The course will be oriented towards observational studies in the teaching examples and course assignment. However, these methods have strong applications to RCTs. The focus of the course is on the application of the techniques to solve epidemiological problems rather than on the statistical methods and underlying mathematical principles. The course will examine non-regression methods (e.g. Kaplan-Meier and log-rank test) as well as regression modeling (COX models). Students will be expected to undertake computer analysis of data provided as SAS data sets.

This course is an elective course for all students. All students need permission of the instructor prior to registering. The class will normally be limited to a maximum enrolment of about 20 students.

TEACHING PHILOSOPHY

This course will be taught using interactive lectures (i.e. I will present material to the class but will encourage interaction and discussion during the lecture process). The focus of the course will be on practical applications rather than theory. If enrollment is low, consideration will given to modifying some of the lecture time for seminar discussions.

BACKGROUND EXPECTATIONS

This is a second level course. Students are expected to have a solid foundation in biostatistical methods similar to those covered in EPI 5242. In particular, students will be expected to have covered linear regression methods at a level beyond the simple linear model. This includes topics such as: multiple regression, partial f-tests, residual analysis methods, variable selection strategies.

The course will be a practical course which will require students to use computers for data analysis. Students are expected to have a basic comfort level with the use of computer programmes for simple data management statistical analysis. I recommend the use of SAS for this course; my teaching material will use SAS. However, students can use computer programmes other than SAS for the assignments.

OBJECTIVES

Many important epidemiological designs collect data on the development of new events (e.g. illness, death). These include cohort studies and RCTs. Traditional epidemiological analysis of these studies involves estimation of incidence, based either on count data or person-time. However, these methods are based on assumptions which can be limiting (e.g. that the incidence rate is constant across all follow-up time). More powerful analysis approaches make use of the time-to-event and employ survival methods. The purpose of this course is to introduce students to the underlying approaches of survival analysis and to provide them with the basic analytical methods used for non-regression and regression based analysis methods. These methods will be linked to epidemiological research designs.

By the end of the course, the student will be able to use survival analysis methods to analyze real data and will be able to produce meaningful interpretations of the results in epidemiological terms. They will also be aware of situations which require more advanced statistical help and when the results of an analysis should not be taken on face value. The course **will not attempt** to turn the student into a mathematical statistician. We will not spend time on the detailed underlying theory or derivations of the statistical methods to fit and explore models. Rather, the course will concentrate on the practical aspects of the application of the models and the interpretation of the results. The one theory area we will address is maximum likelihood estimation

since this is essential to understanding the results of the analyses. In addition, students will further develop their ability to analyze data sets using SAS.

More specific objectives include:

1) Understand how to perform Actuarial life table and Kaplan-Meier analyses;
2) Understand the application and interpretation of Cox modeling to survival data;
3) Be able to relate the statistical techniques to the appropriate epidemiological concepts (e.g. confounding and effect modification);
4) Be able to interpret regression coefficients in terms of epidemiological parameters (e.g. relative risk);
5) Understand approaches to variable selection and how these relate to interpretation of the model in biological/epidemiological terms;
6) Be able to use the computer to analyze data sets.

TOPICS NOT COVERED

As in any advanced course, there are many more topics which would be of interest than could be included in the course. I have had to omit a number of topics which are useful in some areas of application and which could easily have been included. My decision was based on an assessment of the techniques which I believe will be most useful to you in a career in epidemiology. Important topics which we will not cover include:

1) Accelerated Failure models;
2) Competing events;
3) Relative survival methods;
4) Multi-level applications of survival analysis;
5) Random effect models;
6) Longitudinal data analysis;
7) Clustering effects, Design effects, etc.;
8) Life events modeling (e.g. when an outcome is a repeating event such as influenza rather than a single event);

We will also not be covering computer techniques for preparing large data files for analysis or techniques for data management in large-scale studies.

EVALUATION METHODS

The course will be evaluated through two assignments. Both assignments will involve simple data management tasks, as well as statistical analysis and interpretation. Students will be expected to apply simple techniques to validate data quality and the accuracy of any data manipulations. The same data set will be used for both assignments. There will be no final examination.

Assignment #1	25%
Assignment #2	75%
Total	100%

The first assignment will relate to non-regression methods of survival analysis. It is the shorter of the two assignments and is intended to provide some interim feedback on progress. It is due to be handed in on March 23. I will attempt to hand-back the marked assignments on March 30.

The second assignment is a larger analysis task and will involve application of Cox regression models. It is due to be handed in on April 23.

Please note that the data set you will using involves real data. It is not made-up data sets designed to give simple or neat answers. Rather, it is a selection of real data from the Framingham study; the data may well be messy or have inconclusive relationships, etc. As such, you should not assume that you should jump in with a complex model without looking at the data first using descriptive and similar simple methods. That is, you are expected to undertake appropriate exploratory data analyses before doing the full regression analysis.

A major aspect of these assignments (especially assignment #2) is the interpretation of the results. I do not want you to simply give me a computer print-out: I may know how to interpret the print-out but you have to show me that **you** know how to interpret it. You should not 'cut and paste' the SAS output results. Rather, re-type the relevant parts of the output into a suitable table. You also should not just type out the regression coefficients and leave it at that. Again, I may know how to interpret the coefficients – your task is to show me that you also know how. At a minimum, the analyses must be converted back into epidemiologically meaningful parameters (e.g. IDR). However, what I really am looking for is that you can interpret what the coefficients and analyses mean in epidemiological terms. This is a key objective of the course. For example, what do the analyses mean in terms of confounding and effect modification issues? How do they impact on the underlying hypotheses?

READING MATERIAL

Core Books

Students are recommended two books for this year's class. The book by Cantor provides a solid introduction to the field, with many excellent SAS examples and macros to improve your analyses. This will be the core basis for the course. The second book (by Machin et al) is a somewhat more advanced presentation of the core material but also includes some topics which are not covered by Cantor. There will also be supplemental reading material used. This material will mainly relate to the non-regression analysis methods but will also cover some of the more advanced Cox model topics.

The primary textbooks recommended for the course are listed here. Copies have been ordered at the Health Sciences bookstore.

> *Cantor AB.* SAS Survival Analysis Techniques for Medical Research. Springer Sciences and SAS Institute, Cary, USA, 2003. ISBN: 978-1-59047-135-7
>
> *Machin D, Cheung YB, Parmar MKB.* Survival Analysis: A Practical Approach. 2nd edition. John Wiley and Sons, Ltd. 2006. ISBN: 0-470-87040-0

To supplement this book and my suggested readings, additional resources and reading material can always be useful. So, here are some other recommended books, followed by a list of supplemental books which might be useful for some students.

First, SAS. You should have been exposed to SAS during Biostatistics 1. If not, or if you want to develop a stronger foundation with SAS, the following two books are very useful. You can explore these books at SAS (http://support.sas.com/publishing/). In fact, large chunks of these books can be read on-line at no charge.

> *Delwiche LD, Slaughter SJ. The Little SAS Book: A Primer, 3rd edition. SAS Institute, Cary, NC, 2003.* **Comments:** This book is not a great reference guide (since it limits discussion to 2 pages on every topic). But, it gives an excellent over-view of most topics in data management, etc. In particular, make sure to look at the ODS material and the discussion on data set manipulations.
>
> *Cody R. Learning SAS by Example: A Programmer's Guide. 2nd edition. SAS Institute, Cary, NC, 2007.* **Comments:** The Cody book is more detailed and provides a slightly more advanced coverage of material.

Next, to the course content. I am not suggesting that you should buy all of these books! But, some of them might be useful, especially if you will be using one of these techniques in your future work.

> *Allison PD. Survival Analysis Using the SAS System: A Practical Guide. SAS Institute Inc., Cary, 1995.* **Comments:** An excellent practical book which you will find extremely helpful if you plan on using SAS to undertake survival analyses. It covers basic concepts rather briefly but is filled with examples of SAS programme to produce a wide range of analyses. I will be using it more some of my teaching in this area. However, it is over 15 years old and uses a version of SAS which is now obsolete. You can buy it directly from 'SAS' at their web site (http://www.sas.com/apps/pubscat/booklist.jsp?attr=category&val=User%27s+Guides). Make sure to select 'Canada' as your country. This web site also lists other books in related areas which may be of interest to you.
>
> *Dupont WD. Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data, Cambridge University Press, Cambridge, 2002.* **Comments:** A good book, covering nearly everything in the course with an appropriate level of presentation. However, all of the examples are worked out using STATA rather than SAS. It is also sometimes rather terse in the explanations.

Supplemental Books

The following section presents a summary of other books addressing the areas covered in this course. You might consider exploring these resources.

1) Rothman KJ, Greenland S. Modern Epidemiology, 2nd edition. Lippincott-Raven, Philadelphia, 1998. **Comments:** The current textbook for Epi II. An excellent and very exhaustive book. While it is primarily an epidemiology text, it takes a very statistical approach and would be a very strong complement to this course. In particular, there are very good sections on model building strategies. We will be using some of this material during the course.
2) Armitage P, Berry G, Matthews JNS. Statistical Methods in Medical Research. Blackwell Science, Inc. 2001. **Comments:** A new up-date (4th edition) of an old stand-by. Most of the book covers basic statistics rather than our material.
3) Holford TR. Multivariate Methods in Epidemiology. Oxford University Press, 2002. **Comments:** An interesting book. It covers both model-based and non-model-based approaches to epidemiological analysis. Lots of good material. Some of it gets a bit mathematically complex at times.
4) Kleinbaum DG. Survival Analysis: A Self-Learning Text. Springer, New York, 1996. **Comments:** This book (a companion to the previous one) covers the basic analytic methods but use a different presentation style from most books. The same comments apply as above.
5) Hosmer DW Jr, Lemeshow S. Applied Survival Analysis: Regression Modeling of Time to Event Data. John Wiley & Sons, Inc, New York, 1999. **Comments**: A companion to their book on logistic regression. This one is a bit more mathematical but still presents a very good overview of survival methods.
6) Collett D. Modeling Survival Data in Medical Research. Chapman & Hall, London, 1994. **Comments**: A bit old but written at a good level. Covers non-regression based methods as well as Cox models, etc. Includes a number of worked examples using SAS, BMDP and SPSS (although these analysis programmes have all been expanded since the book was published).

TIMETABLE (2009/10)

	#	Date	Topic
	1	Mar 2	Introduction to Survival methods. Incidence proportion vs. incidence rate. The hazard function. Censoring. Risk set. Actuarial and Kaplan Maier methods. Confidence intervals.
	2	Mar 9	Estimation of the hazard function. Concept of proportional hazards. Non-regression approaches to comparison of groups.
	3	Mar 16	Non-parametric regression (LOESS), smoothing methods and application to non-linearity analysis. Test for trend. Permutation and randomization tests. Power analysis.
*	4	Mar 23	COX Models I. Introduction to Maximum Likelihood methods. Parameter estimation. Types of predictor variables and modeling approaches. Interpretation of coefficients. 95% confidence interval estimation. Ties.
	5	Mar 30	COX Models II. Application to confounding and effect modification. HR variability by 'predictor' variables vs. 'over time'. 'Trend' analyses with continuous predictors. Non-linear predictor effects. Estimating survival curves from Cox models.
	6	Apr 6	COX Models III. Time dependent covariates. Sample size estimation.
	7	Apr 13	COX Models IV: Numerical problems. Testing of proportional hazards assumption. Schoenfeld residuals. Introduction to non-proportional hazards models and stratified Cox models.

* - Assignment due

April 23 – final assignment due date

TIMETABLE (2008/9)

	#	Date	Topic
	1	Jan. 6	Introduction to Maximum Likelihood methods. Non-parametric regression (LOESS), smoothing methods and application to non-linearity analysis.
	2	Feb 24	Introduction to Survival methods. Incidence proportion vs. incidence rate. The hazard function. Censoring. Actuarial and Kaplan Maier methods.
	3	Mar 3	Estimation of the hazard function. Comparison of groups.
*	4	Mar 10	COX Models I. Basic ideas. Model fitting. Types of predictor variables and modeling approaches. Interpretation of coefficients.
	5	Mar 17	COX Models II. Application to confounding and effect modification. Time dependent covariates.
	6	Mar 24	COX Models III. Numerical problems. Testing of proportional hazards assumption. Interval censoring. Discrete time models
	7	Mar 31	COX Models IV: Non-proportional hazards models and stratified Cox models.

* - Assignment due

April 10 – final assignment due date