Use of Swedish Health Registers in Epidemiology

Paul W. Dickman
Department of Medical Epidemiology and Biostatistics (MEB)
Karolinska Institutet

paul.dickman@meb.ki.se

December 3, 2003

Biostatisticians at MEB

- 3 Professors
  - Yudi Pawitan
  - Juni Palmgren
  - Marie Reilly
- 3 Associate professors (docent/lektor)
  - Rino Bellocco
  - Paul Dickman
  - Keith Humphreys
- 4 Researchers
- 1 Postdoctoral fellow
- 6 Doctoral students

Overview of my presentation

- Examples of the use of health registers in epidemiology
  - Case-control studies
  - Cohort studies
- Multi-generation register
- Web-based data collection at MEB

Legal and ethical issues

- Research involving health registers is approved by Karolinska Institutet's ethics committee (where applicable) and is conducted in accordance with the Personal Data Act (1998), Data Act (1973), guidelines from the Swedish Data Inspection Board, and conditions specified by the data providers (primarily Statistics Sweden and the National Board of Health and Welfare).

Some statistical challenges in epidemiology

- Aim is typically to estimate a relative risk (RR) from observational data

$$RR = \frac{\text{risk(disease} \mid \text{exposed)}}{\text{risk(disease} \mid \text{unexposed)}}$$

- Bias
  - Non-response
  - Incomplete data
  - Misclassification
- Precision
  - Particularly for genetic/molecular exposures and studies of interaction (gene-gene, gene-environment, or environment-environment) – efficient study design is essential
- Confounding

The case-control design

- For example, using the nationwide, population-based Swedish Cancer Register we can identify women diagnosed with ovarian cancer during a 2-year period (the cases).
- Using the total population register (*registret över totalbefolkningen*) select a set of controls — women of a similar age resident in the same region as the cases but without a diagnosis of ovarian cancer.
- Obtain information on exposures of interest (e.g., reproductive history, use of hormone replacement therapy, oral contraceptive use, smoking, diet) and compare the distribution of exposure between cases and controls.
- The availability of population-based registers from which to sample cases and controls strengthens validity compared to other countries where, for example, cases may be hospital-based and controls selected from other patients in the same hospital or by random digit dialling.

- Details of exposures of interest are typically obtained by questionnaire or personal interview but can be obtained by matching with population-based registers such as the
  - Hospital discharge register (e.g., for information on gynecological surgery)
  - Medical birth register (for information on reproductive history)
  - Census (*folk- och bostadsräkningar*) for information on socioeconomic factors
- Major issues are
  - Misclassification of exposure or outcome (particularly differential misclassification of exposure)
  - Incomplete data (item non-response and unit non-response)
- Use of registers provides high quality information on exposure and outcome and also provides a first stage set of covariates for adjustment for incomplete data using a missing at random assumption.

Cohort studies

- The case-control design is hindered by potential problems due to misclassification of exposure.
- An alternative is the cohort design where we sample from a source population and follow-up disease-free individuals for the event of interest.
  - Information on exposure is collected prospectively (no problems with recall bias)
  - Information on exposure is obtained before the outcome is known (no problems with differential misclassification of exposure)
- Disadvantages of the cohort design are that a large number of individuals are required (particularly for rare disease) and they must be followed for a long time (particularly for exposures with a long latency).
- An ideal design is the historical cohort study in which a past cohort is identified and their experience up to the present is obtained.

- Douglas Altman (in *Practical Statistics for Medical Research* [1]) writes that 'few studies like this are carried out since the necessary data are rarely available'.
- Such studies are possible in Sweden and are one of the greatest strengths of Swedish Epidemiology.
- For example, the association between birthweight and adult blood pressure can be studied using the medical birth register to ascertain birthweight and the conscription register to ascertain adult blood pressure.
- With a similar design we can study the association between birthweight (or other pregnacy/birth characteristics) and performance on an intelligence test.
- Since prenatal ultrasound was routinely used in one county before other parts of Sweden we can also study the association between prenatal ultrasound and performance on an intelligence test.

A register-based cohort study of ovarian cancer

- Evidence from epidemiologic and experimental studies indicates that ovarian carcinogenesis in large part is due to factors associated with reproduction and ovulation: the risk is reduced with increasing parity, increasing maternal age at first birth, and oral contraceptive use, while use of certain hormone replacement therapies increases the risk of ovarian cancer.
- The exact biological mechanisms are, however, unknown – there are presently at least five hypotheses for the pathogenesis of ovarian cancer, all of which have shortcomings.
- For example, the 'incessant ovulation hypothesis' suggests that the risk of ovarian cancer increases with the number of ovulations. It is supported by the protective effects of parity and oral contraceptives but contradicted by the lack of protective effect of an early menopause, spontaneous abortions, and the uncertain association with lactation.

- We hypothesize that markers of hormonal exposures during pregnancy such as gestational age, offspring birthweight, placental weight, and preeclampsia influence risk of ovarian cancer and may explain the protective effect of parity and possibly also influence the protective effect of a high maternal age at first birth on ovarian cancer risk.
- We will establish a cohort of more than 1.4 million women who delivered their first infant between 1973 and 2000, and were included in the Swedish Medical Birth Register.
- We aim to retrieve information about maternal characteristics, pregnancy complications, placental weight and birth characteristics for all births to these women from 1973 through 2000.
- The women will be followed up in the Hospital Discharge Register to retrieve information about gynecological surgery (oophorectomy, hysterectomy, and other ovarian or tubal operations), in the Cause of Death and Emigration Registers to ascertain vital status, and in the Cancer Register to identify diagnoses of ovarian cancer up to the year 2002.

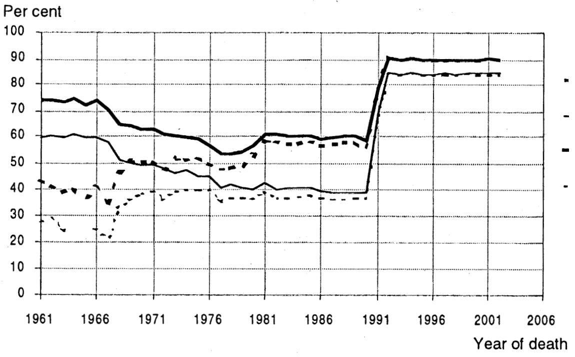
Risk set sampling

- If we need to collect additional information not available in registers then we can sample from the cohort using a nested case-control design or a case-cohort design.
- In our register-based cohort study of ovarian cancer we expect to observe 1000 cases of ovarian cancer among the 1.4 million women in the study.
- Since most of the information is in the cases (individuals who experience the event of interest) we may wish to collect additional information on all cases but a sample of the controls.
- Since the study population is sampled from an enumerable cohort we know all relevant sampling fractions and can estimate, for example, both absolute and relative risks.
- The nested case-control and case-cohort designs ignore the available information on the controls not selected for the study – we can also utilise such information using a two-stage design.

The multi-generation register

- The multi-generation register (maintained by Statistics Sweden) contains information on family relationships for individuals who were born 1932 or later and registered in Sweden at some time since 1961 [2].
- It is a valuable resource for research in genetic epidemiology, such as studies of clustering of disease within families.
- We have, for example, used this register (together with other registers) to study whether colorectal cancer and inflammatory bowel disease share a common genetic etiology by studying the risk of colorectal cancer among relatives of individuals with inflammatory bowel disease (and vice versa) [3, 4].
- An issue with the register is that information on family relationships is not known for all index individuals and information is less complete for individuals who have died.

Figure 3: Deceased 1961-2002, born in Sweden
Number with information on parents, before and after additions



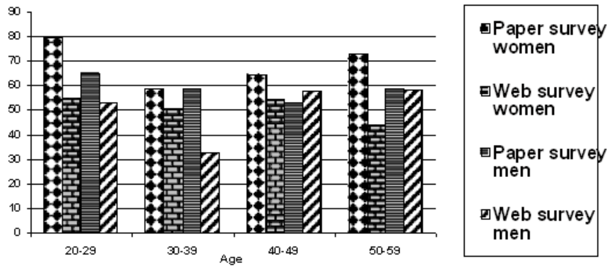
- If we can model missingness than we can apply methods for analysing incomplete data [5].
- There are also issues with truncation and censoring due to the limited coverage periods of the registers.
- Using a binary indicator 'positive family history' may not be the most efficient means of modelling family history.

Data collection using the web

- We are studying the feasibility of collecting data using web-based questionnaires as an alternative to mail questionnaires or interviews.
- Our aim is to use this alternative method of data collection in standard epidemiological designs.
- For example, we recently randomised 900 individuals to one of three methods of data collection
  - Standard paper questionnaire
  - 'Paper in a box' web questionnaire
  - 'Enhanced' web questionnaire

  (Contact: Olle Bälter, `olle.balter@meb.ki.se`)

Response rates by age, sex, and type of questionnaire



A large web-based study

- Over 10 years ago 50,000 women completed a questionnaire for a cohort study aimed at examining risk factors for various diseases, including cancer, diabetes, and circulatory and psychiatric diseases. In the spring of 2003, all women were asked to complete a similar web-based questionnaire.
- After one reminder, 45% have responded to the questionnaire – women who do not complete the web questionnaire will be given the opportunity to complete a paper questionnaire.
- We are studying the effect on response rates of alternative approaches for ordering the questions and methods for follow-up.
- We will match with the LOUISE database (*Longitudiell databas för utbildning, inkomst och sysselsättning*) maintained by Statistics Sweden to obtain information on characteristics of responders and non-responders which will be used to adjust for non-response.

  (Contact: Alexandra Ekman, `alexandra.ekman@meb.ki.se`)

References

[1] Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall, 1991.

[2] Statistiska centralbyrån (Statistics Sweden). *Multi-generation register 2002: A description of contents and quality*. Statistiska centralbyrån, 2003. `http://www.scb.se/Statistik/BE/OV9999/2003M00/BE96ST0305.1.pdf`.

[3] Askling J, Dickman PW, Karlén P, Broström O, Lapidus A, Löfberg R, Ekbom A. Colorectal cancer rates among first-degree relatives of patients with inflammatory bowel disease: a population-based cohort study. *The Lancet* 2001;**357**:262–266.

[4] Askling J, Dickman PW, Karlén P, Broström O, Lapidus A, Löfberg R, Ekbom A. Family history as risk factor for colorectal cancer in inflammatory bowel disease. *Gastroenterology* 2001;**120**:1356–1362.

[5] Andersen EW, Andersen PK. Adjustment for misclassification in studies of familial aggregation of disease using routine register data. *Statistics in Medicine* 2002;**21**:3595–3607.