REVIEW ARTICLE

Data Analysis of Epidemiological Studies

Part 11 of a Series on Evaluation of Scientific Publications

Meike Ressing, Maria Blettner, Stefanie J. Klug

SUMMARY

Introduction: An important objective of epidemiological research is to identify risk factors for disease. Depending on the particular question being asked, cohort studies, case-control studies, or cross-sectional studies are conducted.

Methods: Methods of data analysis in different types of epidemiological studies are illustrated through examples with fictive data. Important measures of frequency and effect will be introduced. Different regression models will be presented as examples of complex analytical methods.

Results: Important frequency measures in cohort studies are incidence and mortality. Important effect measures such as the relative risk (RR), hazard ratio (HR), standardized incidence ratio (SIR), standardized mortality ratio (SMR), and odds ratio (OR) can also be calculated. In case-control or cross-sectional studies, the OR can be calculated as an effect measure. In cross-sectional studies, prevalence is the most important frequency measure. The interpretation of different frequency measures and effect measures will be discussed.

Conclusion: The measures to be calculated and the analyses to be performed in an epidemiological study depend on the research questions being asked, the study type, and the available data.

Cite this as: Dtsch Arztebl Int 2010; 107(11): 187–92
DOI: 10.3238/arztebl.2010.0187

Institut für Medizinische Biometrie, Epidemiologie und Informatik, Universitätsmedizin der Johannes Gutenberg-Universität Mainz: Dr. med. Ressing, Prof. Dr. rer. nat. Blettner, PD Dr. rer. nat. et med. habil. Klug

Epidemiology is used to describe the distribution of diseases in the population and to analyze the causes of these diseases. One important objective is to identify risk factors and to quantify their significance. A risk factor can influence the probability that a specific disease develops. Risk factors include:

- Environmental influences (for example, exposure to radon)
- Predisposition (for example, genes), or
- Behavioral characteristics (for example, hormone intake).

Epidemiological research employs various different types of study (1–3), depending on the question asked. The most important are

- Cohort studies
- Case-control studies, and
- Cross-sectional studies

In cohort studies, persons exposed to specific risk factors are compared with persons not exposed to these factors. The occurrence of diseases or deaths in these two groups is observed prospectively. Data from cohort studies allow the estimation of incidence rate and mortality rate as descriptive measures of frequency, as well as relative risk (RR) or hazard ratio (HR) as comparative effect measures. Standardized incidence ratios (SIR) or standardized mortality ratios (SMR) are used for comparison with the general population.

In case-control studies, persons suffering from the studied disease are compared with controls who do not have the disease. Exposure is recorded retrospectively. The odds ratio (OR) is calculated as a comparative effect measure.

In cross-sectional studies, the exposure and disease status are examined for a sample from a defined population at the same time point. The prevalence of various diseases and the risk factors, as well as the OR can be determined.

Effect estimates, such as RR, are normally calculated with regression models, taking influencing factors into consideration. These lead to statements about the extent of changes in the frequency of a disease due to a specific risk factor. To assess whether the observed effect is statistically significant, the confidence interval (CI) should, for example, be considered for all effect estimates (4). If a statement is to be made about the number of cases of the disease caused by the risk factor, then the risk difference (RD) is considered.

Material and Methods

Studies on the link between hormone replacement therapy (HRT) and breast cancer will be used to illustrate the difference in analysis of the different study types. Various articles and textbooks can be recommended for more advanced reading (3, 5–10).

Whatever the study design, the study population should first be described (description) (11). For example, age can be given as the mean value and standard deviation (for normal distribution), or as the median and range, or in a histogram. Studies on breast cancer and HRT normally also examine influence factors such as menopausal status, family history, marital status and education. These variables should be included in the analysis, as they may be risk factors for breast cancer and are potential confounders (12). Risk factors may also be effect modifiers. Effect modification means that the influence of one factor (for example, HRT) on a disease (for example, breast cancer) is modified by the presence of another factor (for example, smoking). In other words, there is an interaction between the two factors. The effects should be examined in different subgroups (stratification), each with the same analysis.

An analysis plan must be prepared when the study is being planned and this must include a detailed description of the study design and the planned data analysis.

Example: cohort study

Between 1996 and 2001, the Million Women Study in Great Britain included 828 923 postmenopausal women aged between 50 to 64 years and without breast cancer (13). The occurrence of breast cancer in this group was then monitored with the help of the Cancer Registry (follow-up).

Incidence and Mortality—The incidence describes the number of persons in a defined population who develop a disease for the first time during a defined period in time. A distinction is made between the cumulative incidence and the incidence rate (incidence density).

It is decisive for the cumulative incidence estimate *(Figure)* that all study participants were at risk of developing breast cancer at the start of the observation. The cumulative incidence is also often interpreted as the risk that an individual develops a specified disease within a period in time. Women who had already fallen ill are excluded from the calculation of the cumulative incidence. They can be used to calculate the prevalence of breast cancer in the study population (*Figure)*.

The incidence rate considers the period in which each individual was in fact at risk of developing breast cancer and could be monitored. This period is designated as person-years (the years which the study participants contribute to the cohort study) and added up for the whole group examined. Not all women were at risk of developing breast cancer throughout the period of the study. One reason might be that they had already died from other causes before the end of the study. If each participant can be followed throughout the period of the study, the incidence rate is the same as the cumulative incidence.

The mortality rate is calculated from the number of deaths, rather than the number of new diseases (*Figure)*. For the disease-specific mortality (in this case, from breast cancer), solely the deaths from a defined disease (in this case, from breast cancer) are counted. Case-fatality is a measure of the mortality from a specific disease (*Figure)*.

Relative risk (RR) and risk difference (RD)—The calculations of RR and RD are shown in the *Figure*. The RR is calculated by dividing the risk of disease for an exposed person by the risk of disease for a non-exposed person. To calculate the RD, these two risks are subtracted (14).

Standardized incidence ratio (SIR) and standardized mortality ratio (SMR)—The aim of the calculation of the SIR or the SMR is to compare the incidence or mortality in the cohort with the general population. It is investigated whether the incidence or mortality in the cohort differ from the values for the general population. It is calculated how many cases or deaths would be expected in the cohort if the incidence or mortality were the same as in the general population *(Table 1a/b)*. The SIR or SMR are calculated by dividing the observed number of cases (or number of deaths) in the cohort by the expected number of cases (or number of deaths) *(Figure)*.

Regression models—Comparison between the users and non-users of HRT is only permissible if there is no difference between these groups, except with respect to exposure. This means that these two groups should be equivalent with respect to other factors relevant to breast cancer, such as age.

These influence factors are considered in the data analysis by employing subgroup analyses or by adjustment in regression models (5, 6, 15, 16). The principle of regression analysis is to investigate the common influence of several potential influence factors on the target parameter. For example, Cox regression or Poisson regression can be used for the data analysis of cohort studies, depending on the target parameter (5, 15) *(Table 2)*.

In a Cox regression, the target parameter is the time until the occurrence of an event (for example, disease or death). The data are censored, meaning that not all participants could be observed throughout the entire study duration. Cox regression uses a proportional hazard model to calculate the hazard ratio (HR). The underlying assumption is that the risk in the two groups differs by a specific factor. The interpretation of HR and RR is similar. The Million Women Study considered not only HRT intake, but also other factors, such as age.

If an effect modification (interaction) is examined, this interaction is considered in the regression model (interaction term). This approach can be used to reveal interactions between different factors (10).

Poisson regression is used if the target parameter is the number of observations of a rare event, for

FIGURE

Measure		Calculation
Incidence rate*1	=	$\frac{\text{Number of new cases in the time period} \times 100\,000}{\text{Number of person-years}}$
Cumulative incidence per period*1	=	$\frac{\text{Number of new cases in the time period} \times 100\,000}{\text{Number of persons in the cohort}}$
Mortality rate*1	=	$\frac{\text{Number of deaths in the time period} \times 100\,000}{\text{Number of person-years}}$
Cumulative mortality per period*1	=	$\frac{\text{Number of deaths in the time period} \times 100\,000}{\text{Number of persons in the cohort}}$
Case-fatality*2	=	$\frac{\text{Number of deaths from the defined disease in the period} \times 100}{\text{Number of new cases of the defined disease in the population}}$
Odds ratio	=	$\frac{\left(\frac{\text{Number of exposed diseased patients}}{\text{Number of non-exposed diseased patients}}\right)}{\left(\frac{\text{Number of exposed controls}}{\text{Number of non-exposed controls}}\right)}$
Prevalence*2	=	$\frac{\text{Number of diseased patients in the study population} \times 100}{\text{Number of persons in the study population}}$
Relative risk	=	$\frac{\left(\frac{\text{Number of exposed diseased patients}}{\text{Number of exposed persons}}\right)}{\left(\frac{\text{Number of non-exposed diseased patients}}{\text{Number of non-exposed persons}}\right)}$
Risk difference	=	$\frac{\text{Number of exposed diseased patients}}{\text{Number of exposed persons}} - \frac{\text{Number of non-exposed diseased patients}}{\text{Number of non-exposed persons}}$
Standardized incidence ratio (SIR)	=	$\frac{\text{Observed number of cases}}{\text{Expected number of cases}}$
Standardized mortality ratio (SMR)	=	$\frac{\text{Observed number of deaths}}{\text{Expected number of deaths}}$

Important epidemiological frequency measures and comparative measures;
*1 per 100 000,
*2 per cent

example, the number of breast cancer cases within a defined period.

Logistic regression models can also be used in cohort studies (see below).

Example: case-control study

As part of a large study by the WHO, a case-control study was performed in Switzerland on the influence of HRT intake on the development of breast cancer (17). Between 1990 and 1995, 230 breast cancer patients and 507 controls (patients with other diagnoses) aged between 24 and 75 years were enrolled in this study at Lausanne University Hospital and asked about their intake of HRT.

For the binary target variable of a case-control study (disease yes/no), logistic regression is the best suited statistical model to estimate OR *(Figure, Table 2)*. In this example, a multivariate model was used to consider additional potential risk factors for breast cancer (17).

It is not possible to calculate RR in a case-control study, as no incidence can be calculated (14). OR can be interpreted as RR, if the disease is rare.

Example: cross-sectional study

In a cross-sectional study in the USA, 800 women aged between 50 and 70 years were randomly selected from the administrative records of a primary care practice (18). They were then sent a questionnaire on the intake of HRT. The main outcome here was not the diagnosis of breast cancer (yes/no), but the intake of HRT (yes/no).

The prevalence can be calculated in cross-sectional studies as a measure of frequency *(Figure)*. The prevalence describes how frequently a specific disease or a

TABLE 1

Calculation of the standardized incidence and mortality ratios based on a cohort study on the association of hormone replacement therapy (HRT) and breast cancer (13)

a) Expected incidence in exposed study population				
Age group	Number of exposed persons in each age group in the cohort (fictive) A	Breast cancer incidence in each age group per 100 000 (general German population* [19]) B	Expected number of exposed cases of breast cancer in exposed members of each age group per year in the cohort	Expected number of exposed cases of breast cancer in exposed members of each age group in 2.6 years in the cohort
50–54	181 736	221.1	401.8	1 044.7
55–59	138 119	286.9	396.3	1 030.4
60–64	116 311	299.1	347.9	904.5
Total	436 166	–	1146.0	2 979.6
b) Expected mortality in exposed study population				
Age group	Number of exposed persons in each age group in the cohort (fictive) A	Breast cancer mortality in each age group per 100 000 (general German population* [19]) B	Expected number of deaths from breast cancer in exposed members of each age group per year in the cohort	Expected number of deaths from breast cancer in exposed members of each age group in 2.6 years in the cohort
50–54	181 736	39.7	72.1	187.6
55–59	138 119	58.3	80.5	209.4
60–64	116 311	75.8	88.2	229.2
Total	436 166	–	240.8	626.2

*taken as approximation for breast cancer incidence and mortality in Great Britain

specific risk factor occurs in a population at a defined point in time. The prevalence OR can be calculated in a cross-sectional study as a measure of effect. It needs to be emphasized that the prevalence OR can only be interpreted as RR when the prevalence is low.

Interpretation of estimates and confidence intervals

The various effect estimates described above state the extent of change in the frequency of a disease due to a specific risk factor. A value of 1 means that exposed persons have the same risk of falling ill as non-exposed persons. If the value is above 1, this means that this risk factor increases the frequency of the disease. If the value is less than 1, the factor is considered to be protective. In other words, this factor reduces the risk of disease. The confidence intervals (CI) and p-values need to be considered for all effect estimates, to help assess whether the observed effects are statistically significant (4).

The confidence interval includes the true value with a specific probability, usually 95%. If the confidence interval does not include 1, the effect estimate is considered statistically significant (4).

If a statement is to be made about the number of cases of the disease caused by a risk factor, the risk difference (RD) is calculated. If the RD is 0, this means that there is no difference between exposed and non-exposed persons.

Results

Cohort study

Incidence—In the cohort study, 7140 of 828 923 postmenopausal women developed breast cancer within the observation period of six years (13). This corresponds to a cumulative incidence for this period of 861 per 100 000, or an average of 144 per 100 000 per year *(Figure)*.

The 828 923 women could be observed for a mean period of 2.6 years, or 2 155 200 person-years in all. Therefore, the incidence rate is 331 per 100 000 person-years.

Relative risk (RR) and risk difference (RD)—The calculation of the crude (unadjusted) RR is shown in the *Figure*. As the crude RR does not permit final conclusions, data analysis normally presents adjusted estimates from multiple regression models (see below).

Standardized incidence ratio (SIR) and standardized mortality ratio (SMR)—SIR and SMR were not calculated in the Million Women Study. For this reason, the following parameters are used to explain these estimates *(Table 1a/b)*:

- A fictive age distribution in the cohort of the Million Women Study
- The incidence and mortality of breast cancer in the general population (19)
- The resulting expected number of cases of breast cancer and deaths from breast cancer.

4246 cases of breast cancer were observed in the exposed women within 2.6 years, giving an SIR of 1.43, as calculated from the formula in the *Figure*. Thus, with the above assumptions, 1.43 times as many cases of breast cancer would occur in the exposed women in the cohort than would have been expected in the general population.

For the calculation of the SMR, it was assumed that 780 exposed women died of breast cancer within the period of 2.6 years, giving an SMR of 1.25. This is interpreted analogously to SIR.

Regression—After adjusting for other factors in this example, Cox regression gave a statistically significant HR of 1.66 with a 95% confidence interval of 1.60–1.72 for women currently taking HRT, in comparison to women who had never taken HRT (13). This HR means that women currently taking HRT have a 1.66-fold increased risk of developing breast cancer.

Case-control study

Calculation of an unadjusted OR is described in the *Figure*. As, however, an unadjusted OR does not allow final conclusions, the publication on this study only presents the adjusted OR (17). The OR was adjusted for further potential risk factors and protective factors for breast cancer, using a logistic regression model, yielding an OR of 1.2 (95% CI: 0.8–1.8). Although this is increased, it is not statistically significant.

Cross-sectional study

In the cross-sectional study, the prevalence of HRT was 48% in women aged 50 to 59 years. If the influence of age on the intake of HRT is examined, the age-adjusted OR for the 55- to 59-year olds compared to the 50- to 54-year olds was 1.3 (95% CI: 0.8–2.2). Here, the "cases" were the women who use HRT and the "controls" the non-users. The group of 55- to 59-year old women took HRT 1.3-fold more often than women aged 50 to 54 years. Nevertheless, the result is not statistically significant.

Discussion

The selection of the study design has a decisive influence on the analysis of the study. Important effect measures have been presented in epidemiological studies. The emphasis was on the descriptive frequency measures of incidence, mortality, prevalence and the comparative effect measures RR, OR, SIR, and SMR. These comparative effect measures are mostly determined by regression analysis.

In cohort studies, the RR makes a statement about the extent of the change in the probability of developing the disease due to a defined risk factor. Taken together with the confidence interval, the RR shows the relevance of the risk factor for the disease. The OR is only an approximation of RR for rare diseases. The RD depends on the frequency of the disease. In this way, statements can be made about the number of cases of a disease caused by a defined risk factor. RR and RD must be evaluated very differently when communicating risks.

TABLE 2

Overview of the calculation of the effect estimates, depending on the target parameter and the comparison group

Type of target parameter	Effect estimate	Examples	Model or type of calculation
Comparison within the study population (internal comparison)			
Dichotomous	Odds ratio, Relative risk	Breast cancer (yes/no)	Two by two table
Dichotomous	Odds ratio	Breast cancer (yes/no)	Logistic regression
Time to first event	Hazard ratio	Time to death, Time to recurrence, Time to disease	Cox regression
Rare events	Relative risk	Number of breast cancer cases	Poisson regression
Comparison of the study population with the general population (external comparison)			
Dichotomous	Standardized incidence ratio, Standardized mortality ratio	Breast cancer (yes/no)	Age standardization

These methods can also be used in clinical epidemiology. These may investigate specific interventions, i.e., therapies or diagnostic procedures, as influence factors. In cohort, case-control and cross-sectional studies, the influencing factors are only observed, without any intervention taking place. Nevertheless, the statistical analysis is similar.

Whatever the study type, study planning and procedure must always avoid the various forms of bias, such as systematic errors (for example, selection of study population) and confounding factors (12). If this is not successful, these problems must be considered during data analysis, if possible. Moreover, possible interactions (effect modifications) are examined.

Detailed planning must be performed ahead of time and an analytical protocol laid down in writing in advance. These are important requirements if the study is to provide adequate answers to the questions being asked.

Conflict of interest statement
The authors declare that no conflict of interest exists according to the guidelines of the International Committee of Medical Journal Editors.

Manuscript received on 31 July 2009, revised version accepted on 2 November 2009.

Translated from the original German by Rodney A. Yeates, M.A., Ph.D.

REFERENCES

1. Blettner M, Heuer C, Razum O: Critical reading of epidemiological papers. A guide. Eur J Public Health 2001; 11: 97–101.
2. Röhrig B, du Prel J-P, Wachtlin D, Blettner M: Study design in medical research—Part 3 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009;106(11): 184–9.
3. Klug SJ, Bender R, Blettner M, Lange S: Wichtige epidemiologische Studientypen – Artikel Nr. 18 der Statistik-Serie in der DMW. Dtsch Med Wochenschr 2007; 132: e45–7.

4. du Prel J-B, Hommel G, Röhrig B, Blettner M: Confidence interval or p-value? Part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(19): 335–9.
5. Ziegler A, Lange S, Bender R: Überlebenszeitanalyse: Die Cox-Regression – Artikel Nr. 17 der Statistik-Serie in der DMW. Dtsch Med Wochenschr 2007; 132 (Suppl 1): e42–4.
6. Bender R, Ziegler A, Lange S: Logistische Regression – Artikel Nr. 14 der Statistik-Serie in der DMW. Dtsch Med Wochenschr 2007; 132: e33–5.
7. Bender R, Ziegler A, Lange S: Multiple Regression – Artikel Nr. 13 der Statistik-Serie in der DMW. Dtsch Med Wochenschr 2007; 132: e30–2.
8. Bender R, Lange S: Die Vierfeldertafel – Artikel Nr. 6 der Statistik-Serie in der DMW. Dtsch Med Wochenschr 2007; 132: e12–4.
9. Gordis L: Epidemiology. 3rd edition. Philadelphia: Elsevier Saunders 2004.
10. Rothman KJ, Greenland S, Lash TL (eds.): Modern Epidemiology. 3rd edition. Philadelphia: Lippincott Williams & Wilkins 2008.
11. Spriestersbach A, Röhrig B, du Prel J-P, Gerhold-Ay A, Blettner M: Descriptive statistics: the specification of statistical measures and their presentation in tables and graphs—part 7 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(36): 578–83.
12. Hammer GP, du Prel J-P, Blettner M: Avoiding bias in observational studies—part 8 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(41): 664–8.
13. Beral V, Million Women Study Collaborators: Breast cancer and hormone-replacement therapy in the Million Women Study. Lancet 2003; 362: 419–27.
14. Sauerbrei W, Blettner M: Interpreting results in 2×2 tables: extensions and problems—part 9 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(48): 795–800.
15. Breslow NE, Day NE: Statistical methods in cancer research. Vol. II: The analysis of cohort studies, IARC Sci Publ No. 82. Lyon: International Agency for Research on Cancer 1987.
16. Breslow NE, Day NE: Statistical methods in cancer research. Vol. I: The analysis of case-control studies, IARC Sci Publ No. 32 ed. Lyon: International Agency for Research on Cancer, 1980.
17. Levi F, Lucchini F, Pasche C, La Vecchia C: Oral contraceptives, menopausal hormone replacement treatment and breast cancer risk. Eur J Cancer Prev 1996; 5: 259–66.
18. Finley C, Gregg EW, Solomon LJ, Gay E: Disparities in hormone replacement therapy use by socioeconomic status in a primary care population. J Community Health 2001; 26: 39–50.
19. Gesellschaft der Epidemiologischen Krebsregister in Deutschland e.V., Robert Koch-Institut: Krebs in Deutschland 2003–2004. Häufigkeiten und Trends, 6th revised edition. Berlin: 2008.

Corresponding author
PD Dr. rer. nat. et med. habil. Stefanie Klug, MPH
Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI)
Universitätsmedizin der Johannes Gutenberg-Universität Mainz
55101 Mainz, Germany
klug@imbei.uni-mainz.de