# Rate and rate ratio

**Pamela Warner**

## Background

Analysis of findings from a cohort study by Vessey and Yeates[1] are reported in this issue of the Journal. These notes are intended to provide some background to and explanation of the statistical methods used. [See Box 1 for a glossary of terms used in this article.]

## What are they?

In human health it is generally the case that the more people studied, and the longer the surveillance of them, the greater the number of new occurrences of any specified condition there are likely to be. This is because of the increase in total time 'at risk' of having an event, that is, the increased persons-and-time of observation. A rate is an arithmetic way of quantifying the number of events occurring in relation to both persons and time exposed to risk of the event. In order to avoid very small and hence difficult decimal fractions, the rate will often be expressed in terms of some multiple of the person-time unit. For example, Vessey and Yeates report a rate of first hospitalisations for cervicitis, among women aged 40–44 years, of 33 per 10 000 woman-years.[1] This is a simpler way of expressing the underlying but exactly equivalent rate of 0.0033 cases per (single) woman-year of observation.

A rate *ratio* (RR) quantifies a relative comparison between the rates (of occurrence of an event) in two groups, most usually of a group exposed to some risk factor, and a group not exposed.[2] The RR is simply obtained by dividing the rate in the one group by the rate in the other. So, for example, the RR for first hospitalisations for cervicitis, in women aged 40–44 years, relative to the youngest age group (i.e. 25–34 years), is 33/11 = 3, which indicates that the older group have an event rate three times that of the younger women.[1]

## When/why are they useful?

A rate is the natural summary to calculate in cohort studies when the feature of interest is occurrence of events in relation to people exposed, and how long they have been exposed.[2] In order to calculate a rate it is necessary to know the true follow-up period for each individual in the study, and whether this culminated with the event occurring.

An RR is useful when there is an implicit wish to compare event rates between groups. RRs lend themselves to multivariable modelling that can examine the effect of an exposure of interest [say oral contraceptive (OC) use] at the same time as adjusting for potential confounding factors (e.g. age).[2] The multivariable analysis of the event rate for a specified condition involves, implicitly, both a giant table of counts of events that have occurred, cross-classified in terms of the risk factor(s) under study and all other potential explanatory factors, and also a corresponding table accumulating total person-time of follow-up for every cross-classification cell. Therefore, for any specified cell the rate of events can be calculated as the number of events divided by accumulated follow-up. Examination of the effect of the risk factor (say, exposure versus not) involves comparison of rates in relevant cells (that is, calculation of RRs between the cells), adjusted for other factors.

The exposure and other variables might be binary (two groups), categorical (more than two groups), ordered categorical (such as the risk factors presented in Vessey and Yeates' Table 2)[1], or continuous (for example, body mass index, if it had not been recoded into a categorical variable). Some risk factors will be constant for an individual throughout the follow-up period (such as 'social class at entry to study'). Other risk factors are time varying, in that they can change over the course of follow-up. While it is difficult to imagine the dataset – let alone the mathematics involved – methods for multivariable analyses of rates can accommodate all these types of explanatory factors, including time-varying risk factors. Finally, it is common for cohort studies to investigate more than one type of event, for example, uterine polyp, cervicitis, cervical erosion and vaginitis/vulvitis.[1] This makes scientific sense, and is cost-effective, given that the major effort of such a study is recruitment and follow-up procedure, and relatively little extra work is required to enquire about four event types rather than one.

## What precautions are needed?

The analysis of event rates for a specific condition is much more complex if more than one occurrence of that condition is possible per person. While a condition such as death can happen only once for an individual, many health conditions can and do recur. For such conditions, useful clinical information can be obtained without the need for excessively complex analysis, if ascertainment of an event is restricted to the *first occurrence* within the follow-up period. This approach has been adopted by Vessey and Yeates[1] in that they have specified each of their events as 'first hospitalisation for' uterine polyp, etc.

The most crucial aspect of calculation of rates and RRs is accurate recording of person-time of follow-up, in particular, any changes in time-varying covariates (e.g. parity, OC use), and the point when loss-to-follow-up occurs (e.g. due to emigration or accidental death) or when follow-up accumulation should be terminated. 'Termination' of follow-up should apply if only 'first' events are being studied *and* there has been an occurrence of that condition. No further occurrences *of that same event* will be counted for that patient and so, in respect of that event, neither should any further follow-up she has. This is to avoid the estimate of the event rate being biased downwards by the addition of (spurious) follow-up time to the denominator in a situation when, even if an event occurred, it would *not* be added to the numerator. The clock must also be 'stopped' (for all events under study) if a study participant is lost to follow-up for any reason, since in such cases any subsequent event(s) that might have occurred would have been unknown, and so could not have been added to the numerator. Similar bias-avoidance steps must be taken in respect of situations where even if a person continues being followed-up, they have no chance of experiencing the event in question, by excluding such follow-up ('at risk') time from calculations of the relevant rates. For example, in respect of event rates for conditions

*J Fam Plann Reprod Health Care* 2009; **35**(2): 111–113

**Public Health Sciences, University of Edinburgh Medical School, Edinburgh, UK**
Pamela Warner, BSc, PhD, *Reader in Medical Statistics and Associate Editor, Journal of Family Planning and Reproductive Health Care*

**Correspondence to:** Dr Pamela Warner, Public Health Sciences, University of Edinburgh Medical School, Teviot Place, Edinburgh EH8 9AG, UK. E-mail: p.warner@ed.ac.uk

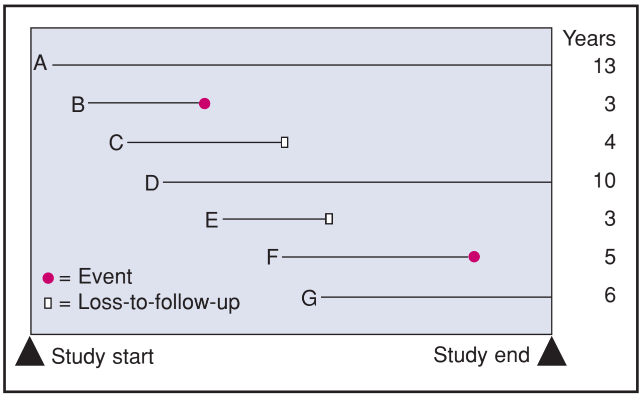

**Figure 1** Hypothetical follow-up of seven participants in a cohort study

other than vaginitis/vulvitis, Vessey and Yeates had to ignore follow-up time post-hysterectomy, since having had this surgery meant that the only study event that could still occur was vaginitis/vulvitis.[1]

## Example

Figure 1 illustrates the follow-up of seven hypothetical members of a cohort, recruited over the first 7 years of the study, which lasted just over 13 years in total. Subjects B and F experienced events, after 3 and 5 years of follow-up, respectively. Subjects A, D and G illustrate a common situation, namely that follow-up continues until the end of the study observation period without any event occurring, nor loss-to-follow-up. Two individuals (C and E) were lost to follow-up after 4 and 3 years, respectively, before experiencing an event, but if this had not been recorded properly then their follow-up would have been presumed to be 11 years and 8 years, respectively. The total follow-up was 44 person-years, so the rate of occurrence of the event is 2/44 = 0.045 per person-year. It is notable that if loss to follow-up had not been recorded properly then the rate would have been calculated as 2/56 = 0.036, an underestimate by 20%!

It is necessary also to keep track of time-varying risk factors. Some factors change steadily throughout the follow-up period (age, calendar year), while some can switch (OC user or not), and others change in a step-wise way (in an OC user, duration of use increases steadily year by year, until OC use ceases, from which point cumulative use is unchanging, until the next resumption of OC use, if this transpires).

Figure 2 illustrates the 12-year follow-up course for a hypothetical participant, and corresponding values for her time-varying covariates. For simplicity, time is segmented/counted in whole years but in practice if exact dates were available then finer measurement would be used (e.g. months). At study entry in 1969 this woman was 26 years of age, had used OC for the past 3 years and had given birth to one child. If the event 'uterine polyp' (E) occurred in the 7th year of her follow-up, then the factor values corresponding to that event would be: current non-user of OC, age 32 years, 5 years of OC use, 3 years since stopping OC, parity 3 and calendar year 1975.

While it is relatively easy to locate the timing of occurrence of the event in terms of the risk factor under study (e.g. this event would be placed in the third OC-use row of Vessey and Yeates' Table 3[1] corresponding to 5 years duration), the important part of calculating accurate rates and RRs is the correct apportioning, to the appropriate denominators ('cross-classification' cells – see above), of all that person's follow-up until that event. For uterine polyp, the illustrated patient had 2 person-years of follow-up while she was classified 'up to 4 years of OC use', and 5 years while 'from 4 to up to 6 years of OC use'. A similar exercise is undertaken for the adjustment factors: for parity the woman had 5 years of follow-up at parity 1 to 2 and 2 years of follow-up at parity 3, but for age all 7 years of follow-up was during 'age 25–34 years', and so on. This needs to be accumulated across all cohort participants, and fortunately this can be accomplished relatively easily these days using available computer technology, provided data are recorded suitably. Once all uterine polyp events and follow-up periods are attributed to the appropriate cells of the cross-classification tables implicit in the multivariable model, then rates of the event per person-time of follow-up can be analysed in relation to risk factors of interest (e.g. OC use), adjusting for nuisance effects of other factors (such as age, parity, calendar period).

In order to understand the importance of calendar year, one needs to consider another participant almost identical to the one illustrated (also developing a polyp), except for the fact that this woman entered the study 5 years later, in 1974. If over the 5-year time lag in their life-courses there had been either some increase in clinical enthusiasm for investigation for polyps in post-pill women, or improvement in the technology to detect polyps, then there would be a tendency for the later recruit to have her polyp detected and dealt with earlier than the illustrated patient, perhaps at 2 years since stopping OC, rather than at 4 years. Such a drift, across the time-course of a cohort study, would tend to blur real patterns in timing of events relative to pill use that might otherwise have become apparent on analysis. Calendar year is therefore relevant to analyses of cohort rate rates, to enable adjustment for potentially confounding trends in medical diagnosis or management (in that these could play a part in the ostensible 'timing' of events). Inclusion of calendar period thus ensures a better estimate of the true association of event timing with the exposure factor of interest, here OC use.

If more than one condition is being studied then follow-up counted and apportioned can differ across conditions,

| Follow-up years | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Subject X | 🟥 | 🟥 | 🟥 | 🟥 | ⬜ | ⬜ | ⬜ E | ⬜ | 🟥 | 🟥 | 🟥 | ⬜ |
| Time-varying factors | | | | | | | | | | | | |
| Calendar year | 69 | **70** | **71** | **72** | **73** | **74** | **75** | **76** | **78** | **79** | **80** | **81** |
| Age (years) | 26 | **27** | **28** | **29** | **30** | **31** | **32** | **33** | **34** | **35** | **36** | **37** |
| Years of OC use | 3 | **4** | **5** | **6** | 6 | 6 | 6 | 6 | **7** | **8** | **9** | 9 |
| Years since OC use | NA | NA | NA | NA | **1** | **2** | **3** | **4** | **NA** | NA | NA | **1** |
| Current OC user | Y | Y | Y | Y | **N** | N | N | N | **Y** | Y | Y | **N** |
| Parity (n) | 2 | 2 | 2 | 2 | 2 | **3** | 3 | 3 | 3 | 3 | 3 | 3 |

🟥, OC use. ⬜, OC non-use. Bold values indicate a change from the previous year.
E, event; N, no; NA, not applicable; OC, oral contraceptive; Y, yes.

**Figure 2** Time-varying risk factor values for a hypothetical participant

for example, because follow-up for a specific condition ceases at first occurrence of it. For the participant in Figure 2, in respect of polyp, only 7 years follow-up is apportioned (until the first polyp event), but for the other conditions under study, her full 12 years counts. The apportioning and accumulation of follow-up therefore needs to be undertaken separately for each condition.

## Overview

Rates and RRs provide a powerful method for analysis of event rates in relation to potential risk factors, allowing for simultaneous adjustment for nuisance factors. Of key importance is accurate measurement of follow-up, and the apportioning of this and events to the correct values (or grouped ranges) of time-varying covariates.

### Statements on funding and competing interests

**Funding** None identified.

**Competing interests** None identified.

### References

1 Vessey M, Yeates D. Some minor female reproductive system disorders: findings in the Oxford-Family Planning Association contraceptive study. *J Fam Plann Reprod Health Care* 2009; **35**: 105–110.

2 Kirkwood BR, Sterne JAC. *Essential Medical Statistics*. Oxford, UK: Blackwell Science, 2003.

3 Porta M. *A Dictionary of Epidemiology*. Oxford, UK: Oxford University Press, 2008.

4 Hennekens CH, Buring J. *Epidemiology in Medicine*. Boston, MA: Little Brown & Company, 1987; 56.

5 Warner P. Testing and quantifying association in binary data. *J Fam Plann Reprod Health Care* 2009; **35**: 26–27.

### Box 1: Glossary of statistical terms used in this article

**'Adjusted' association** See *Association* and *Logistic regression*.

**Association** Relationship between two variables, here an event rate and a risk factor. Association means that the occurrence of a particular value of the risk factor variable, in a segment of the person-time follow-up [say those with longer duration of oral contraceptive (OC) use], is associated with a higher (or lower) rate of the event.

**Binary variable** Has only two possible values (e.g. OC user or not, female or not).

**Categorical variable** Has a set of distinct values, which can be nominal or simply descriptive (such as blood group) or ordered (such as grouping by duration of OC use).

**Cohort study** A prospective study design where individuals are recruited and followed up over time, to observe the incidence of cases of some condition of interest. Since exposure to factors under study is ascertained for each individual at the outset, and/or monitored over time, the association of the occurrence of the condition with exposure factors can be established.

**Confounding variable** A variable that influences the occurrence of the event under study (positively or negatively) and is also associated with the risk factor being studied, but is not on the causal pathway between that risk factor and the event.[3] Such a 'nuisance' variable can *confound* efforts to ascertain the true association of the risk factor with the event, leading to spurious inflation or damping of the rate ratios estimated. However, if the potential confounding variable (e.g. age) is measured then its effect can be 'adjusted' away at the analysis stage.

**Denominator** for a ratio, proportion or rate The divisor; the number being divided *into* the other number; the number on 'the bottom'.

**Explanatory variable** A feature potentially associated with outcome (rate of event).

**Incidence** The number of *new* occurrences of a condition (onset) in a specified study group. The 'condition' might be an illness (e.g. a sexually transmitted infection) or might be contraceptive choice (e.g. undergoing sterilisation).

**Incidence rate** See *Rate.*

**Multivariable regression modelling** A method of analysis that allows assessment of the association between some risk factor of interest and an outcome (or here, event rate), while taking account also of the effects of other variables with potential influence. The analysis is comparing rates between risk factor subgroups, and these comparisons are expressed as rate ratios. If the risk factor is binary, there will be only one rate ratio (RR) (exposed to unexposed). However, if it is categorical, then a number of RRs will arise, each one compared to the reference category (e.g. OC non-user, or non-smoker). The number of RRs estimated will be one less than the number of levels of the risk factor, whereas for the comparison of the reference category to itself, the RR is known to be 1 (since numerator and denominator must be equal!), so no calculation/estimation is needed. Where the risk factor has an ordinal effect, such as smoking or duration of OC use, then an even more powerful test is possible, for trend in rates across the levels of the risk factor.

**Numerator** for a ratio, proportion or rate The number being divided *by* the other number; the number on 'the top'.

**Proportion** This is a special kind of ratio where the divisor (denominator) is the 'whole' and the quantity being divided (numerator) is part *of that whole*.[4] Therefore, by definition the two quantities are in the same units. For example, number of those surveyed currently using OC divided by the total number surveyed (including those using OC).

**Rate** 1. General – *a rate* summarises a change in some quantity in relation to another, usually ***but not always*** in relation to time.[3] For example, population growth (of so many thousands per year, say) is a rate per time unit, whereas perinatal mortality rate expresses neonatal deaths *per live births.*

2. Epidemiological – in health research a *rate* quantifies the occurrence of health events/conditions in relation to persons as well as ***time***. Specifically, 'incidence rate' expresses the occurrence of *new* cases by persons and time. For example, if 1200 persons are followed up for a total of 6600 years (an average follow up of 5.5 years per person), then if there are 132 occurrences of the condition the incidence rate is 132/6600 person-years of follow-up = 0.02 cases per person-year of follow-up. To avoid small numbers/excess zeroes, rates are often expressed in terms of multiples of persons or of years, say per 1000 persons or per 5 years. So the rate of 0.02 occurrences per person-year could instead be expressed as 0.1 per person per 5 years, or 2 per 100 person-years (or 2 per 100 persons per year).

**Rate ratio (RR)** The RR is the ratio of two health event rates, typically the rate in the group exposed to some risk factor under study, relative to the corresponding rate in the unexposed group. Since both rates are in the same units, the ratio will be dimensionless. If the exposure has no effect then it would be expected that the two rates calculated should be approximately equal and hence the ratio approximately 1, which is therefore the 'null' value for an RR. This is analogous to odds ratios.[5] Also similarly, the more extreme the RR (away from 1, i.e. 0.5 vs 0.7 or 2.4 vs 1.3), the greater the degree of association. If the RR is greater than 1 then exposure is associated with more occurrences across time; if it is less than 1, the exposure is protective against occurrences.

**Ratio** This is an arithmetic summary obtained by dividing one quantity by another, with no implication that they are related, such as ratio of males to females opting for sterilisation, or in the same units (pregnancies to storks observed flying overhead!). Rate and proportion are special types of ratio.