Paper AS11

Incidence Rate - Incidence Proportion, Prevalence Rate and Other Related Statistics

Nadiia Pukhliar, Intego Group LLC, Kyiv, Ukraine

ABSTRACT

There is a tangle with incidence related statistics in clinical trial reporting. In the analysis of Adverse Events one often uses “incidence rate“, but actually means “incidence proportion”. Whereas incidence density rate is the same as incidence rate with only one qualification: denominator is based on subject-time at risk. The algorithm for time at risk calculation is completely different for Adverse Events analysis compared with PD or death analysis. For Adverse Events analysis, FDA uses terminology “incidence rate” alongside with “reporting rate”. How to sort out which statistics name implies the correct object is an open question. The choice of inappropriate algorithm leads to undercounting number of events.

Statistics for incidence can be queried for both safety and efficacy analysis. While working on TLFs for reporting, we found out that there are various approaches and different meanings for the same statistics depending on the context and analysis needs. To produce high quality results we had to browse through plenty of articles and documentation. Now we are able to provide the guide with calculating algorithm and interpretations of results.

First, we will describe statistical considerations for each statistical quantity and method. This will be the first step to understand why and what should be calculated. Next, we will present programming implementation of each algorithm, and results with interpretation for statistics.

INTRODUCTION

Let us distinguish two classes of statistics depending on their purpose. The first class includes measures of disease status. It means that we assess population at one specific moment in time. Such measures are important in epidemiology and could be applied to assess disease status and take necessary actions in case of an epidemic. Marker of statistics in the first class is the word “prevalence” that means the fact of existing or being very common at a particular time or in a particular place.

The second class includes measures that assess the frequency of disease onsets. The numerator of any measure is the frequency of events that are defined as the occurrence of disease. We have to observe a population during certain time to obtain data needed for the calculation of such statistics. For instance, collect data during a clinical trial to analyze the treatment effect. Marker of the second class statistics is the word “incidence”, which means the extent to which something happens or has an effect.

Next, we will investigate statistics from both classes more precisely.

FIRST CLASS: DISEASE STATUS

Prevalence (or prevalence proportion) in epidemiology is the proportion of a particular population found to be affected by a medical condition.

There are three statistics describing prevalence.

POINT PREVALENCE

Point prevalence is the proportion of a population that has the condition at a specific point in time. The key concept is the fact that condition is relevant at the specific moment.

As was described above, prevalence statistics measure disease status. This implies that only the number of subjects is presented in the formulas for calculation, and such factors as period of time population was observed, number of events happened to each subject, or disease duration do not affect the proportion.

Let us say that an event happened to a subject if the subject was affected by an investigated medical condition.

Point prevalence proportion is calculated as number of subjects with event at specified point in time divided by total number of subjects in the population:

$$P = \frac{number\ of\ subjects\ with\ event\ at\ time\ point}{total\ number\ of\ subjects\ in\ population}$$

For example, we are investigating children in a school and examining the proportion of people affected by seasonal flu at the end of a month. There are 500 students in the school, and 75 of them currently have flu. Then point prevalence equals 75/500 = 0.15 or 15%. It means that 15% of the population currently has flu.

PERIOD PREVALENCE

Period prevalence is the proportion of a population that has the condition at some time during a given period (e.g. 12-month prevalence), and includes people who already have the condition at the start of the study period as well as those who acquire it during that period. In other words, we disregard the time of condition occurrence and consider only the fact that the condition exists at any moment of the observation period.

The numerator for period prevalence calculation is equal to the number of subjects having the medical condition at any moment during a specified period. In other words, for these subjects the event began during that period or was ongoing at the start of the period. The denominator is the same as for point prevalence and equal to the number of subjects in the population.

$$P = \frac{number\ of\ subjects\ with\ event\ at\ any\ moment\ of\ period}{total\ number\ of\ subjects\ in\ population}$$

Let us go back to the example with 500 students in a school. We know that 75 students have flu at the end of a month, but other students could have had it and recovered sometime before the end of the month. To calculate one-month prevalence we have to take into account students who were sick at any moment during that month in addition to those 75 students. It does not matter whether the disease began before the start of the considered period or after it. If 100 students had flu in the past month then one-month prevalence equals 100/500 = 0.2 or 20%. Thus, 20% of the population were sick at least one day of the month.

LIFETIME PREVALENCE

Lifetime prevalence (LTP) is the proportion of a population that at some point in their life (up to the time of the assessment) has experienced the condition.

This is the last type of prevalence proportion, which describes if an event happened to subjects sometime in their life. The numerator of lifetime prevalence is the number of subjects with the event that has ever occurred. As in the last case, the denominator equals the total number of investigated subjects.

$$P = \frac{number\ of\ subjects\ with\ event\ at\ any\ moment\ of\ their\ life}{total\ number\ of\ subjects\ in\ population}$$

Students in a school are from 7 to 17 years old. If a subject has ever experienced the flu in his life, he will be added to the numerator. If in the population 400 out of 500 people had flu in their life, then lifetime prevalence proportion is equal to 400/500 = 0.8 or 80%.

SECOND CLASS: DISEASE ONSET

Incidence is one of the key concepts in epidemiology. It is a measure of the probability of occurrence of a given medical condition in a population within a specified period of time. We take into account the time of the event occurrence, in contradistinction to prevalence for which only the fact that subject was having a condition is investigated.

INCIDENCE PROPORTION

Incidence proportion (also known as cumulative incidence) is the number of new cases within a specified time period divided by the size of the population initially at risk. The first thing to notice in the definition is that the population is observed during a strictly limited period of time.

Next, let us explain in more details meaning of the phrase “to be at risk”. A subject is considered at risk if it is possible for him to develop a given disease. The population initially at risk includes all subjects who were considered at risk at the beginning of a specified time period. Most often, all subjects in an investigated group are at risk. Let us describe cases when they are not.

For example, chickenpox is a once-in-a-lifetime disease. In other words, if a child was ill with chickenpox in childhood, then he will not get sick with it anymore in the life. Let us suppose that one child out of 25 in a class gets chickenpox. Since it is a viral disease, most likely several more students will get it, and we want to calculate incidence proportion for a month. The population initially at risk includes all children in the class who have not had chickenpox before.

The last important thing in the definition of incidence proportion is new cases. For period prevalence proportion we count subjects who were having a condition for at least one day of a specified period. However, for incidence proportion we count events, and the event should occur during a specified period to be considered, i.e. start date of

the event falls into the investigated time range. Furthermore, for subjects who experienced the target event twice per period, both events are added to the numerator.

Thus, formula for incidence proportion is as following:

$$I = \frac{number\ of\ new\ cases\ during\ period}{number\ of\ subjects\ in\ population\ initially\ at\ risk}$$

For example, we want to examine incidence proportion for seasonal flu during a month in a school with 500 students. We know that 100 students had flu, but there were 120 cases, i.e. 20 students had flu, recovered and then got sick again. As a result we get incidence proportion 120/200 = 0.6 or 60%. It means that there is 60% of event per subject per month in a school.

INCIDENCE RATE

Incidence rate is a measure of frequency with which a disease or other incident occurs over a specified time period. The numerator is the same as for incidence proportion, i.e. number of events that occur in population. The denominator is different and equals to summation of time each subject in population was observed.

When the denominator is the sum of the person-time of the at risk population, it is also known as the incidence density rate or person-time incidence rate. It allows to consider cases when time at risk differs among subjects.

Incidence rate can be calculated as follows:

$$I = \frac{number\ of\ new\ cases\ during\ period}{total\ time\ subjects\ in\ population\ at\ risk}$$

There are several different ways to calculate time in the denominator. As the first option, we can observe a population during one strictly defined period. For example, 100 subjects have been followed for two years, then the denominator equals 100*2 = 200. Next, there are two options depending on the type of investigated event. There are events that can occur several times during the period, and the other type of events that can only happen once or that require time to recover for the subject. This difference has influence on time at risk defined for each subject.

Let us assume that we are investigating food supplement and have to assess incidence rate of adverse events of special interest during supplement administration, i.e. digestive system problems. Such adverse events can occur several times for the subject, thence the subject remains in at risk group even after the event has happened. In this example, the denominator is calculated as a sum across subjects of treatment periods duration.

As the second example, let us consider an oncology study and incidence rate of such event as death. It is obvious that time for each subject is equal to the difference between death date and initial date of observation of the subject, because death is an event that can only occur once. The same logic of calculation is applied in case when progression of a disease is the target event. Progression of a disease is such kind of events that requires time for recovering, i.e. subject is excluded from the at risk population as soon as the event occurs. For subjects who did not experience the target event, the time is calculated as the overall time for which the subject was followed. It should be noted that for events that can only happen once or that require recovering time, the number of cases in the numerator of incidence rate is equal to the number of subjects who have experienced the event.

PREVALENCE OR INCIDENCE

One can assume that incidence and prevalence behave in the same way. However, high incidence does not always mean high prevalence. Another important factor that influences prevalence is disease duration. The shorter disease duration, the lower prevalence proportion.

Even for a big number of new cases, i.e. incidence proportion is high, prevalence can be low and steady in case when disease is rapid. On the other hand, long time needed for recovering leads to an increase in prevalence, because one case of disease is taken into account in several consequent prevalence assessments.

Thus, prevalence accumulates both incidence and disease duration.

Since prevalence reflects both disease duration and incidence rate, it is not as good as incidence rate alone for studying of disease causes. However, prevalence is useful to assess population needs, especially when subjects who have a disease require specific medical procedures. For example, high prevalence of end-stage renal disease in a population predicts the need for dialysis facilities in the population[2].

REPORTING RATE

Reporting rate of adverse events is widely used in pharmacoepidemiologic and post marketing studies in context of spontaneous reports. Reporting rate consists of the numerator derived from spontaneous adverse event report, and the denominator representing the population at risk for the adverse events. The denominator is estimated from available drug use data with adjustments to estimate the total number of patients exposure to the drug[3].

Reporting rate has similar calculation formula with incidence rate, but due to the limitation in getting exact values for reporting rate, there is no sense to compare their values.

The limitation in assessing risk in pharmacoepidemiologic studies are:

Accurate national estimates of the number of patients exposed to a medical product and their duration of exposure may not be available. It may be difficult to exclude patients who are not at risk for an event, for example, because their exposure is too brief or their dose is too low. A product may be used in different populations for different indications, but use estimates are not available for the specific population of interest[4].

Therefore, it was suggested to calculate crude reporting rates using available data, i.e. total number of spontaneous reported cases as the numerator and estimates of patient exposure to a medical product, i.e. number of subjects or person-time exposed as the denominator:

$$R = \frac{number\ of\ spontaneous\ reported\ events}{estimates\ of\ subjects\ exposure}$$

PROGRAMMING IMPLEMENTATION

After identifying what should be calculated, the next question is how to implement mathematical formulas in SAS® and generate tables. Let us use real word data for illustration, namely Vaccine Adverse Event Reporting System[1] (VAERS) which collects information about adverse events caused by U.S.-licensed vaccines. Just for educational purposes we considered two seasonal influenza vaccines with safety data collected during 2019.

PREPARATION OF DATA

We consider two data sets: subject-level data set with one record per subject – ASL; and analysis data set with one record per subject per adverse event – AAE. Next, we list variables needed for calculations in each of the data sets.

ASL should contain USUBJID – unique subject identifier; TRT – treatment name which is name of vaccine in the VAERS data; TRTSDT – start date of subject to be observed or treated, that is vaccination date; TRTEDT – end date of subject to be observed or treated, that is today's date; TRTDUR – duration of treatment or following the subject, can be calculated as TRTEDT – TRTSDT + 1.

AAE should contain USUBJID – unique subject identifier; AETERM – name of adverse event which is symptom name in VAERS data; AESDT – start date of the event; AEEDT – end date of the event which is omitted in the example.

Described set of data and variables can be adjusted to fit any structure of database.

ASSUMPTIONS

In the investigated example we assume that events are still ongoing till observation end date. In the other case, end date has influence only on point prevalence values, i.e. subject falls out of the numerator if he is not having the condition at the moment of point prevalence assessment.

Another assumption is that all subjects are considered at risk throughout observation period. In studies where the investigated event can only happen once during observation period, e.g. death or progression of a disease, treatment duration should be replaced with time from the treatment start to the event occurrence. All other derivations stay the same.

The last assumption required for accurate prevalence estimation is that all subjects did not have any symptoms at the moment of study beginning, i.e. vaccination date.

IMPLEMENTATION

All statistics are calculated as proportions with two types of the denominators. Let us define two variables for denominators first:

```
proc sql;
    /* define denominators */
    create table denoms as
        select distinct TRT,
               /* total number of subjects in each treatment group */
               count(USUBJID) as TRTTOT,
               /* total treatment duration for subjects in each group in years */
               sum(TRTDUR)/365.25 as DURTOT
        from asl
        group by TRT;
```

```
    /* create data set for analysis which combines AAE with denominators */
    create table anl as
        select aae.*, TRT, TRTTOT, DURTOT
        from aae
        left join denoms
            on aae.TRT = denoms.TRT;
quit;
```

After the above manipulation we get data set ANL which contains all variables needed for derivations. The easiest way to implement simple mathematical formulas in SAS is to use PROC SQL.

PREVALENCE PROPORTION

To calculate period prevalence proportion we should divide the number of subjects who were having a condition sometime during a period by the number of subjects in a population observed. Since VAERS provides data collected during one year, we can calculate one-year-prevalence.

```
proc sql;
    /* period prevalence proportion */
    create table prev_prop as
        select distinct TRT, AETERM,
            /* number of subjects with event */
            count(distinct USUBJID) as S_COUNT,
            /* number of subjects with event/number of subjects in population */
            100*count(distinct USUBJID)/TRTTOT as P_PROP
        from anl
        group by TRT, AETERM
        order by PERCENT desc;
quit;
```

The main thing to highlight in this block of code is DISTINCT option just before USUBJID variable. It allows to calculate the number of unique values of USUBJID cross data set, i.e. subject who has several occurrences of the same symptom is calculated only once for this symptom.

Another DISTINCT option at the first place in SELECT statement specifies that result data set PREV_PROP does not contain full duplicates, i.e. result data set PREV_PROP contains one record per set of values TRT, AETERM, S_COUNT, P_PROP.

INCIDENCE PROPORTION

As one may remember, the main difference in the formula for incidence proportion is the number of events in the numerator rather than the number of subjects with the event. Whereas, the denominator is the same as for prevalence proportion, i.e. the number of subjects in the population at risk. In the proposed example, the population at risk includes all subjects during an observation period.

```
proc sql;
    /* incidence proportion */
    create table inc_prop as
        select distinct TRT, AETERM,
            /* number of events */
            count(USUBJID) as E_COUNT,
            /* number of events/number of subjects in population */
            100*count(USUBJID)/TRTTOT as I_PROP
        from anl
        group by TRT, AETERM
        order by PERCENT desc;
quit;
```

For incidence proportion option DISTINCT is not specified near USUBJID variable. It means that a subject is calculated as many times as the event occurs with him, i.e. we get the number of events as result.

INCIDENCE RATE

To calculate incidence rate we should divide the number of new events by total time subjects in a population at risk. We have no data before the start of observation of the population, therefore all events are considered as new events.

One of the assumptions is that all subjects are at risk throughout period of observation, thus for each subject time at risk is equal to time being observed, or treatment duration.

```
proc sql;
    /* incidence rate per 100 patient-years */
    create table inc_rate as
        select distinct TRT, AETERM,
            /* number of events */
            count(USUBJID) as E_COUNT,
            /* number of events/total treatment duration */
            100*count(USUBJID)/DURTOT as I_RATE
        from anl
        group by TRT, AETERM
        order by PERCENT desc;
quit;
```

The only difference compared with incidence proportion is in the denominator. For incidence rate we use variable DURTOT, which contains summation of observation time across subjects in each treatment group.

ANY SYMPTOM

The very common situation is when we need a summary across symptoms or names of the events. General structure of tables provides first row with statistics related to “Any Adverse Event”, including number of subjects with at least one event, number of event with any name, and corresponding incidence statistics. To calculate incidence or prevalence statistics for any event we should exclude AETERM from the query. As an example for prevalence proportion:

```
proc sql;
    /* period prevalence proportion for any event */
    create table prev_prop_any as
        select distinct TRT,
            /* number of subjects with any event */
            count(distinct USUBJID) as S_COUNT,
            /* number of subjects with any event/number of subjects in population */
            100*count(distinct USUBJID)/TRTTOT as P_PROP
        from anl
        group by TRT
        order by PERCENT desc;
quit;
```

Variable AETERM is removed from both SELECT clause and GROUP BY clause.

OUTPUT TABLE

A wide variety of table templates for adverse events is found in clinical study reports. The most common is the table with number of subject with events and prevalence period proportion of each adverse event by group of variables. For instance, summary of adverse events by preferred term, system organ class and treatment group. Such table can be a subset with the most significant events based on statistics value.

Let us consider for example a table including the number of events and incidence proportion of each adverse event and overall by treatment group, but only with incidence proportion greater than or equal to 10%:

Summary of Adverse Events with Incidence Proportion at least 10% by Term and Treatment

Event Name	Alfuria (N= 917)	Fluarix (N= 997)
Any Adverse Event	3916 (427.04)	4168 (418.05)
Injection site pain	139 (15.16)	174 (17.45)
Pain	153 (16.68)	127 (12.74)
Pain in extremity	109 (11.89)	130 (13.04)
Pyrexia	113 (12.32)	108 (10.83)
Injection site erythema	0	123 (12.34)
Headache	96 (10.47)	0

Subsetting of tables and sorting by statistics in descending order are great tools to reveal events with the most frequent occurrence or the highest degree of influence.

CONCLUSION

We described the structure to help in choice of appropriate statistics for analysis needs. The first thing we have to determine is what the goal is, i.e. safety endpoint can be either disease status or disease onset.

Next, target statistics should be selected from the list. There are two possible ways depending on the input information. Thus, knowing statistics name we can find out the right formula to implement, or having information about calculation algorithm we can determine the correct statistics name and avoid confusion among the programming team, statistician and authorities organs.

Finally, we provided programming implementation for beginners or for ones who would like to learn one more application of SQL procedure in SAS.

In conclusion, we would like to add that the list of statistics described above is far from full in safety analysis. It is only a starting point in structuring and collaboration between statistician and programming team.

REFERENCES

1) The Vaccine Adverse Event Reporting System. URL: https://vaers.hhs.gov/eSubDownload/index.jsp?fn=2019VAERSData.zip
2) Kenneth J. Rothman. 2012. Epidemiology: An Introduction.
3) Rodriguez EM, Staffa JA, Graham DJ. 2001. The role of databases in drug postmarketing surveillance, Pharmacoepidemiology and Drug Safety.
4) U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). 2005. Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Nadiia Pukhliar
Intego Group, LLC
23 Bahhoutivska st.
Kyiv 04107, Ukraine
Work Phone: +1 407 512 1006 (ext 2434)
Email: nadiia.pukhliar@gmail.com