HHS Public Access

Author manuscript

Am J Nurs. Author manuscript; available in PMC 2022 October 06.

Published in final edited form as:

Am J Nurs. 2021 December 01; 121(12): 45–48. doi:10.1097/01.NAJ.0000803196.49507.08.

Overview: Cohort Study Designs

Bernadette Capili, PhD, NP-C [Director],
Heilbrun Family Center for Research Nursing, The Rockefeller University, 1230 York Avenue, Hospital, Room 106, New York, NY 10065

Joyce K. Anastasi, PhD, DrNP, FAAN [Independence Foundation Professor of Nursing]
New York University Rory Meyers College of Nursing, 380 Second Avenue, Suite 305, New York, NY 10010

This paper continues the series on the observational study designs, focusing on the cohort design. The word ‘cohort’ was adopted from the Roman term of 300 to 600 fighting soldiers who march together (Hood, 2009; Hulley, 2013). The epidemiology community-initiated using ‘cohort’ during the 1930s to mean a “designated group which are followed or traced over a period of time “(Hood, 2009, p. E2). The term is currently defined as a group of people with pre-defined common characteristic(s) (i.e., smokers, exposure to lead in drinking water, ICU nurses) followed longitudinally with periodic measurements to determine the incidence of specific health outcomes or events (Alexander, 2015; Hulley, 2013; Song & Chung, 2010). Since cohort studies are observational, study participants are monitored, and study interventions are not provided. This paper describes the prospective and retrospective cohort designs, examines the strengths and weaknesses, and discusses methods to report the results.

Cohort Design

The cohort study design is an excellent method to understand an outcome or the natural history of a disease or condition in an identified study population (Mann, 2012; Song & Chung, 2010). Since participants do not have the outcome or disease at study entry, the temporal causality between exposure and outcome(s) can be assessed using this design (Hulley, 2013; Song & Chung, 2010). A vital feature of a cohort study is selecting the study participants based on mutual characteristics such as geographic location, birth year, or occupation (Song & Chung, 2010). Cohorts are also selected based on exposure and non-exposure status (Setia, 2016). Ideally, both groups are similar except for the exposure status. Additionally, the cohort can be divided based on exposure categories at study entry.

For example, an investigator could recruit people living with HIV (PLWH) who smoke and do not smoke (never smoked) from the same community and follow them over five years to determine the relationship between smoking status and HIV and the incidence of heart disease and stroke in this population. Alternatively, at study entry, the smokers could be categorized based on the smoking pack-years (less than five pack-years or greater than five

(212) 327-8405, bcapili@rockefeller.edu.

pack-years) to determine whether heart disease and stroke are associated with the amount and duration of smoking.

Prospective Cohort Design

The prospective cohort studies are also referred to as longitudinal studies. It is used to answer a specific question(s) in a selected area. Investigators recruit a sample of participants and follow them over time, from the present to the future. At pre-determined time-points, characteristics are measured (using interviews, questionnaires, biological assays, physiologic measures) to understand the relationship between the cohort and study outcome. See figure 1.

During the recruitment phase, the investigator must identify potential participants who plan to move and difficult to reach during the study's follow-up phase. The eligibility criteria should reflect this consideration. The investigator should collect contact information from the enrolled participants, telephone, email address, mailing address, and at least two friends or family members the investigator can contact if they move or die during the follow-up phase (Hulley, 2013). Additionally, the study protocol should schedule periodic contact with the participants, such as telephone calls to provide assessment results, study newsletter, or study incentives (gift cards) to keep the participants engaged.

In continuing with the HIV study example, study participants are recruited from local New York City HIV primary care clinics. The study plans to evaluate participants annually for ten years to determine heart disease and stroke incidence. PLWH are eligible to join if they smoke cigarettes with well-controlled HIV (undetectable viral load). At study entry, individual exposures for smoking are determined (smoking pack-years), medical history and cardiovascular health are evaluated. Participants identified at baseline to have heart disease or a history of stroke are excluded from the study. Participants are categorized into two groups based on smoking exposure, less than five pack-years or greater than five pack-years for this study. The independent variables ((predictor variables) (smoking pack-years, blood pressure, weight, waist circumference, lipid levels), and the dependent variable ((outcome), history of heart disease, and stroke) are assessed annually. The longitudinal design allows investigators to compare changes over time (Fitzmaurice, 2008) and determine if the level of exposure (smoking pack-years) and other variables are associated with the outcome (incidence of heart disease and stroke).

Prospective Cohort Design: Strengths and Weaknesses

A primary strength of the prospective cohort design is that it allows investigators to determine the number of new cases (incidence) occurring over time. From our example, the incidence of new-onset heart disease and stroke among the study participants. Additionally, measuring the predictor variables before the onset of the outcome (heart disease and stroke) strengthens the ability to assess the sequence of events and infer the causal basis of an association between the predictor variables and the outcome (Hulley, 2013).

A limitation of using this design is that it requires a large sample size. Alexander and colleagues (2015) recommend at least 100 participants. Additionally, the cost of conducting the study may be costly in terms of participant recruitment, the number of staff to conduct

the research, and the collection, storage, and analysis of the outcome measurements. Moreover, some conditions (i.e., breast cancer, chronic obstructive disease), despite being relatively common, could occur at low rates in any given evaluation period and not provide meaningful results. Therefore, participants need to be followed for a longer duration, thus increasing cost and the possibility of participants withdrawing from the study or losing them during follow-ups (Hulley, 2013).

Retrospective Cohort Design

Retrospective cohort studies are also called historical cohort studies. The term historical is fitting since data analysis occurs in the present time, but the participants' baseline measurements and follow-ups happened in the past (Hulley, 2013). This type of study is feasible if an investigator has access to a dataset that fits the research question. The dataset must also have adequate measurements about the predictor variables. See figure 1.

Generally, the participants for a retrospective cohort design are generated for other purposes, such as electronic medical records or an administrative database like medicare (Hulley, 2013). This design's primary goal is to review past data (predictor variables) to examine events or outcomes. Institutional review board approval is required for this design even though actual patient interactions do not occur. For example, to ascertain the incidence of heart disease and stroke among PLWH who smoke, electronic medical records of 500 HIV patients from a local HIV primary clinic are examined over ten years, 2010–2020. For this illustration, HIV patients are categorized by their smoking exposure status: smoking less than five pack-years or greater than five pack-years. The outcome of interest is the incidence of heart disease and stroke.

Retrospective Cohort Design: Strengths and Weaknesses

A strength of the retrospective cohort design is the immediate ability to analyze the outcome since it is already assembled with collected measurements and the participants' follow-ups. This type of design is also inexpensive to conduct. A primary limitation of this study is that the available dataset may be incomplete, inaccurate, or measurements undertaken that do not match the research question (Hulley, 2013). In other words, the investigator(s) do not have control over the data collection methods and procedures.

Method to Report Results

During the scheduled evaluation periods, investigators count the *incidence* or the number of participants who develop the outcome of interest (i.e., heart disease and stroke). The methods to measure *incidence* are *risks* and *rates* (Alexander, 2015). Both terms can provide additional information about the exposure of interest (smoking, nonsmoking) by calculating the *risk ratio* and *rate ratio* (Alexander, 2015).

Risk and Risk Ratio

The term *risk* is also known as *cumulative incidence*. It is defined as the number of participants who develop the outcome of interest divided by the total population (participants from the cohort) at risk (Alexander, 2015). For instance, investigators conduct

a study to evaluate the association between smoking and heart disease and stroke among PLWH who attend an HIV primary clinic in lower Manhattan. The investigators follow a total of 1000 PLWH for ten years. Among the 1000 PLWH, 500 were smokers, and 500 were nonsmokers. Participants were evaluated annually. A total of 125 heart disease cases and stroke were diagnosed in the smoking group, while 25 heart disease cases and stroke were diagnosed in the non-smoking group. All the cases of heart disease and stroke were diagnosed at the fifth year follow-up. (See Table 1 for calculations).

$$Risk = \frac{\text{number of participants who develop the outcome}}{\text{total number of participants at risk}}$$

From the above example, 150 cases of heart disease and stroke were identified from the cohort sample size of 1000. Based on the calculations, the risk for developing heart disease and stroke was 15% among the study participants. Additional analyses using the *risk ratio* compared the risk between participants exposed (smoker) and unexposed (nonsmoker) to provide further information about the data. The *risk ratio* illustrates the relative increase or decrease in the incidence between the exposed and unexposed groups (Alexander, 2015). (See Table 1 for calculations).

$$Risk\ Ratio = \text{Risk}_{\text{exposure}}/\text{Risk}_{\text{unexposed}}$$

Using the formula from table 1, the *risk ratio* was 5. The results demonstrate that PLWH who smoke (exposed) were five times more likely to be diagnosed with heart disease and stroke than PLWH who were nonsmokers. To further understand the meaning of the *risk ratio* results, if the result was equal to 1, then the exposure (smoker) did not affect the outcome. In other words, the risk was the same for the exposed and unexposed groups. Similarly, if the risk ratio was less than 1, it indicates that the exposed (smoker) group was protective for heart disease and stroke. When the results are further away (see figure 2)

Rate and Rate Ratio

The term *rate* is also known as an *incidence rate* (IR). It is defined as the number of participants who develop the outcome of interest (heart disease and stroke) divided by the person-time (days, months, years) at risk during follow-up (Alexander, 2015). Person-time is the sum of each participant's total time free (no heart disease and no stroke) from the outcome of interest. This measure provides the accumulated events (cases of heart disease and stroke) and the speed at which new health outcomes transpire in a study cohort. Another analysis used to compare and understand the rate of speed (increase or decrease) of a health outcome between the exposed and unexposed groups is the *rate ratio*.

$$Rate = \frac{\text{number of new cases}}{\text{total person}-\text{time risk}}$$

In continuing with the example from above, the calculated *rate* was 0.016 (see Table 1). The result indicates that 0.016 cases of heart disease and stroke per person-year occurred in the sample, with a *rate ratio* of 5.2. This result indicates that heart disease and stroke rates were 5.2 times greater in the exposed group than in the unexposed group. Similar to the *risk ratio*, if the result was equal to 1, then the smoking exposure did not affect the outcome. If the *rate ratio* was less than 1, smoking exposure was protective for heart disease and stroke. The greater the rate ratio is from 1 (null association, the exposure is not preventive or harmful), the exposure had more impact on the study cohort. (see figure 2).

$$\textit{Rate Ratio}(\text{Incidence Rate Ratio (IRR)}) = \text{IRRe}_{\text{exposed}}/\text{IRR}_{\text{unexposed}}$$

Reporting Recommendations

In continuing the *Step by Step Research* column with the observational studies, the cohort design also has a reporting guideline to explain how a study was conducted and how the results were obtained. Like the cross-sectional study, the cohort study uses the same guideline, *Strengthening the Reporting* of *Observational Studies in Epidemiology* (STROBE) (von Elm et al., 2014). The report provides specific recommendations for cohort studies in the 22-item checklist to guide investigators in what to include in their manuscript. For consumers of the research, the checklist helps the reader understand the paper better regarding study planning, conduct, findings, and conclusions (von Elm et al., 2014). Additionally, the checklist contains information to allow a study to be replicated, useful to make clinical decisions, and sufficient information to be included in a systematic review (https://www.equator-network.org/reporting-guidelines/strobe/).

Conclusion

The cohort design is an appropriate method to determine the incidence of a health outcome or an event. This design is especially helpful in understanding the natural history of disease and conditions in an identified study population. Additionally, this design allows an investigator to examine the timing between an exposure and outcome(s).

Acknowledgments

This manuscript is supported in part by grant # UL1TR001866 from the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH) Clinical and Translational Science Award (CTSA) program, and by the National Institutes of Health/National Institutes for Nursing Research #R01NR017917

References

Alexander L, Lopes B, Richetti-Masterson K, Yeatts KR. (2015) Risk and Rate Measures in Cohort Studies. In: Vol. 2nd. ERIC Notebook. Durham, NC: Department of Epidemiology at the UNC Gillings School of Global Public Health.

Hood MN (2009). A review of cohort study design for cardiovascular nursing research. J Cardiovasc Nurs, 24(6), E1–9. doi:10.1097/JCN.0b013e3181ada743 [PubMed: 19858946]

Hulley S, Cummings SR, Browner WS, Grady DG, Newman TB (Ed.) (2013). Designing Clinical Research (4th ed.). Philadelphia, PA: Wolters Kluwer/Lippincott Williams & Wilkins.

Mann CJ (2012). Observational research methods—Cohort studies, cross sectional studies, and case–control studies. African Journal of Emergency Medicine, 2(1), 38–46. doi:10.1016/j.afjem.2011.12.004

Setia MS (2016). Methodology Series Module 1: Cohort Studies. Indian J Dermatol, 61(1), 21–25. doi:10.4103/0019-5154.174011 [PubMed: 26955090]

Song JW, & Chung KC (2010). Observational studies: cohort and case-control studies. Plastic and reconstructive surgery, 126(6), 2234–2242. doi:10.1097/PRS.0b013e3181f44abc [PubMed: 20697313]

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, & Vandenbroucke JP (2014). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg, 12(12), 1495–1499. doi:10.1016/j.ijsu.2014.07.013 [PubMed: 25046131]

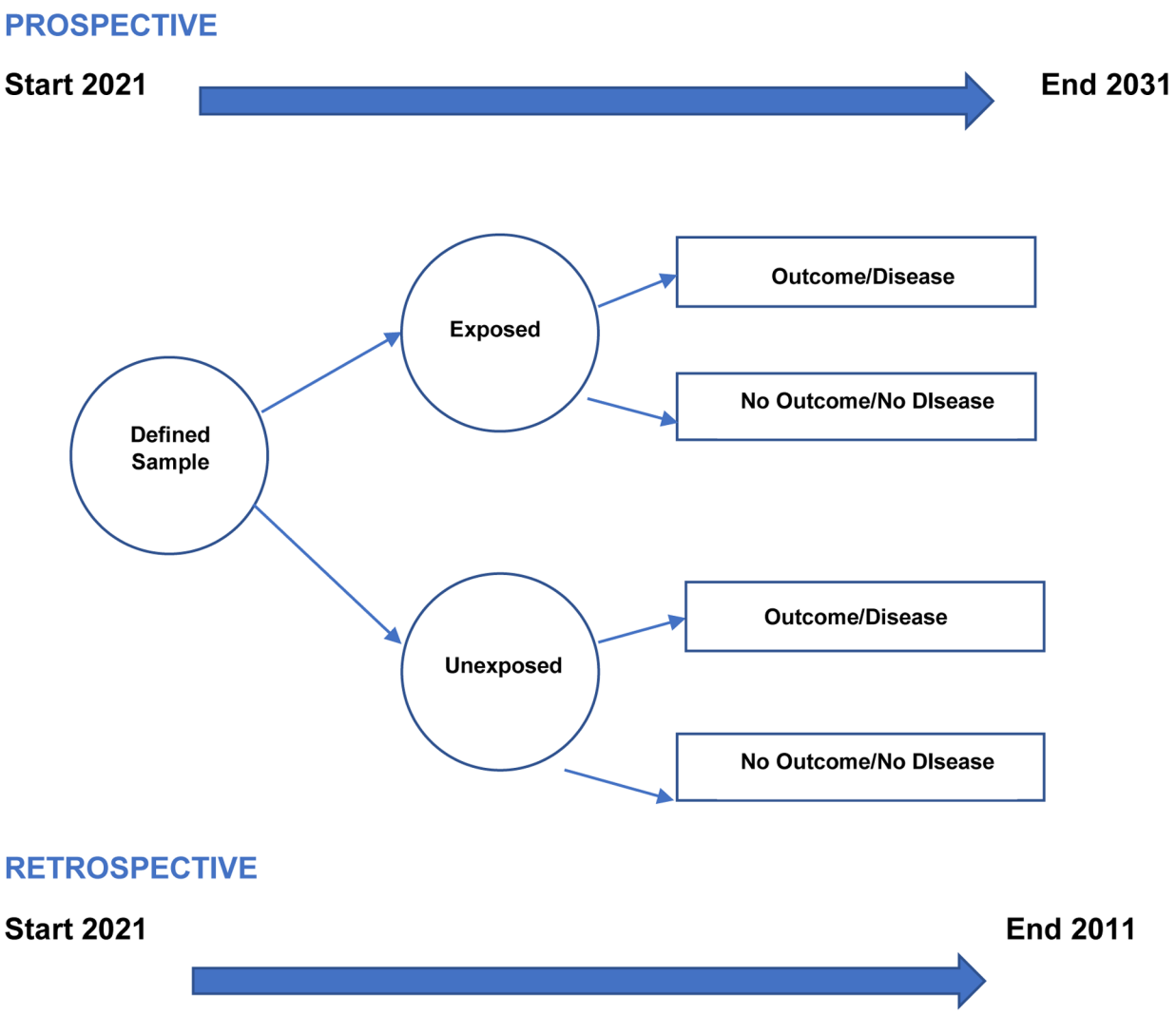

Figure 1.
Prospective and Retrospective Cohort Designs

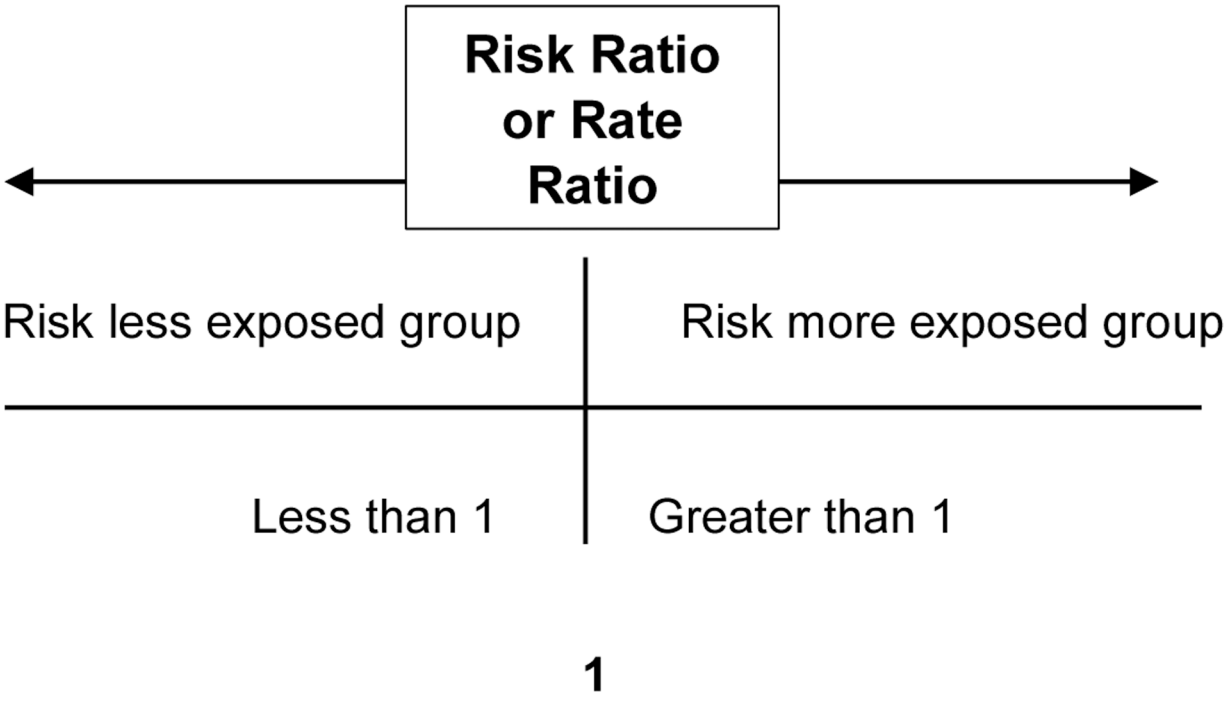

Figure 2.
Risk Ratio or Rate Ratio Interpretation

Table 1:

Calculation Example

	Disease **Heart Disease/Stroke**	**No Disease** **No Heart Disease/Stroke**	**Total**	**Total Person-Time** **(years)**
Exposed Smoker	125 a	375 *b*	500 (a + b)	(125×5) + (375×10) = 4375 (a × 5[+]) + (b × 10[$])
Unexposed Nonsmoker	25 *c*	475 *d*	500 (c + d)	(25×5) + (475×10) = 4875 (c × 5[+]) + (d × 10[$])
Total	150 (a + c)	859 (b + d)	1000 (a + b + c + d)	9250 [(a × 5[+]) + (b × 10[$])] + [(c × 5[+]) + (d × 10[$])]

a = exposed participant and acquires the outcome of interest

b = exposed participant and does not acquires the outcome of interest

c = unexposed participant and acquires the outcome of interest

d = unexposed participant and does not acquire the outcome of interest

Risk (Cumulative Incidence) of PLWH diagnosed with heart disease/stroke: (a+c)/(a+b+c+d) = 150/1000 = .15 × 100 = 15%

Risk Ratio among PLWH who smoke for heart disease and stroke: [a/(a+b)] / [c/(c+d)] = (125/500)/(25/500) = .25/.05 = 5

Interpretation Risk Ratio or Rate Ratio

Risk Ratio or Rate Ratio = 1 Exposure is not preventive or harmful

Risk Ratio or Rate Ratio > 1 Exposure is harmful

Risk Ratio or Rate Ratio < 1 Exposure is protective

Rate (Incidence Rate) of heart disease/stroke among PLWH over a ten year period: a + c/ [(a × 5^{+}) + (b × 10$^{\$}$)] + [(c × 5^{+}) + (d × 10$^{\$}$)] =150/9250 = 0.016 cases/Person-year

Rate Ratio (Incidence Rate Ratio (IRR)): a/[(a × 5^{+}) + (b × 10$^{\$}$)] c[(c × 5^{+}) + (d × 10$^{\$}$) = 0.026/0.005= 5.2

[+]Participants all diagnosed with heart disease/stroke at end of fifth year follow-up

[$]Duration of the study follow-up ten years