CHEST

Cohort Studies
Design, Analysis, and Reporting

Check for updates

Xiaofeng Wang, PhD; and Michael W. Kattan, PhD

Cohort studies are types of observational studies in which a cohort, or a group of individuals sharing some characteristic, are followed up over time, and outcomes are measured at one or more time points. Cohort studies can be classified as prospective or retrospective studies, and they have several advantages and disadvantages. This article reviews the essential characteristics of cohort studies and includes recommendations on the design, statistical analysis, and reporting of cohort studies in respiratory and critical care medicine. Tools are provided for researchers and reviewers.

CHEST 2020; 158(1S):S72-S78

KEY WORDS: bias; cohort studies; confounding; prospective; retrospective

General Overview of Cohort Study Design

The term "cohort" in modern epidemiology refers to "a group of people with defined characteristics who are followed up to determine the incidence of, or mortality from, some specific disease, all causes of death, or some other outcome."[1] A cohort study observes people as two or more groups, from exposure to outcome.[2] A key feature of the cohort study design is that subjects are followed up over time. It begins with subjects who are exposed and not exposed to a factor and then evaluates the subsequent occurrence of an outcome. Unlike cross-sectional studies, which are often used to determine prevalence, cohort studies are used to study incidence, causes, and prognosis.

In clinical research, cohort studies are appropriate when there is evidence to suggest an association between an exposure and an outcome, and the time interval between exposure and the development of outcome is reasonable. Cohort studies are the design of choice for determining the incidence and natural history of a condition. Due to their longitudinal design feature, one can look at disease progression and natural history.[3] Cohort studies allow us to calculate the incidence rate, cumulative incidence, relative risk, and hazard ratio. Causality cannot be established definitively through a cohort study.[2] Nevertheless, cohort studies are useful to provide evidence that suggests causality and information regarding the strength of the association between the risk factors and the outcome.

Description of Subtypes of Cohort Studies

Cohort studies can be either prospective or retrospective. The type of cohort study is determined by the outcome status. If the outcome has not occurred at the start of the study, then it is a prospective study; if the outcome has already occurred, then it is a retrospective study.[4] Figure 1 presents a

ABBREVIATION: CAP = community-acquired pneumonia

AFFILIATIONS: From the Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH.

CORRESPONDENCE TO: Xiaofeng Wang, PhD, Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, 9500 Euclid Ave/JJN3-01, Cleveland, OH 44195; e-mail: wangx6@ccf.org

Copyright © 2020 American College of Chest Physicians. Published by Elsevier Inc. All rights reserved.

DOI: https://doi.org/10.1016/j.chest.2020.03.014

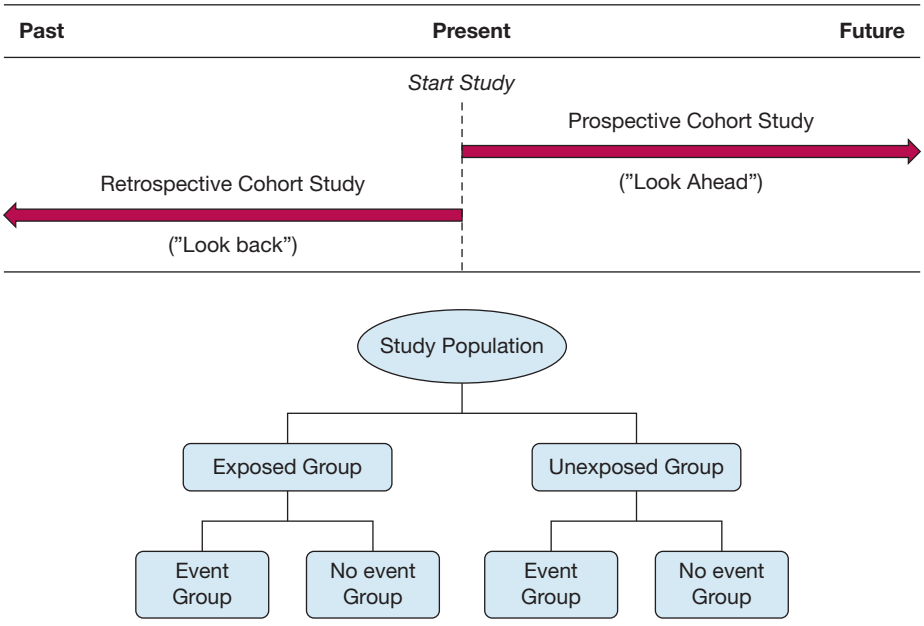

Figure 1 – *Graphical representation of the timeline in a prospective vs a retrospective cohort study design.*

graphical representation of the designs of prospective and retrospective cohort studies. The distinguishing feature of a prospective cohort study is that at the time that the investigators begin enrolling subjects, none of the subjects has developed the outcome of interest. In contrast, a retrospective study is conceived after subjects have already developed the outcome. The investigators jump back in time to identify a cohort of subjects at a point in time when they did not have the outcome. A prospective cohort study design is ranked higher in the hierarchy of evidence than a retrospective design because the outcome, predictor, and confounding variables can be better measured and controlled.[5] Information gained from a retrospective study can be helpful in planning a future prospective study.[6]

A study combining two study designs, the case-cohort design, is a combination of a case-control and cohort design that can be either prospective or retrospective. The case-cohort design can be viewed as a variant of the nested case-control design.[7] In a nested case-control study, one starts with identifying cases that have already occurred (retrospective) or as they occur (prospective) in a defined cohort. A specific number of control subjects are then selected from among those in the cohort. Limitations in this type of design include: (1) inefficiency due to the need to align each selected case subject to its matched control subject; and (2) when there is more than one outcome considered, strict implementation of the design requires the selection of a new set of control subjects for each distinct disease outcome. The case-cohort design was proposed by Prentice[8] as a cost-effective alternative to the nested case-control design. In a case-cohort design, a subcohort is randomly drawn from the full cohort, and the case-cohort sample consists of the subcohort plus those subjects from the entire cohort whose outcome occurred during the study period. Figure 2 illustrates the subject selection process of a case-cohort sample. The case-cohort study design is efficient when only a very small fraction of the full cohort develops the outcome in the given study time frame and the exposure measurement of interest is expensive to obtain.[9]

Use Cases of Cohort Studies

Example 1

Nijkeuter et al[10] conducted a prospective cohort study to understand the natural course of hemodynamically stable pulmonary embolism (PE). The study aimed to evaluate the incidence of recurrent VTE, hemorrhagic

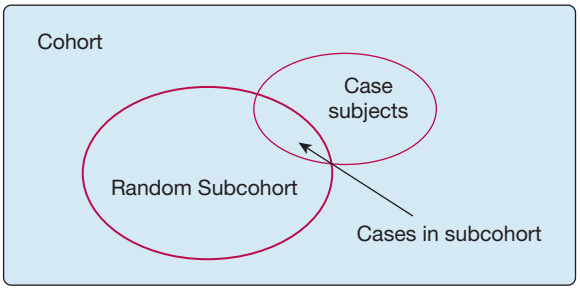

Figure 2 – *An illustration for subject selection in a case-cohort study. The case-cohort sample consists of the subcohort members as well as all the case subjects who are outside of the subcohort.*

complications, and mortality in patients with PE, and to identify risk factors and the time course of these events. Between November 2002 and September 2004, a total of 3,503 patients with clinically suspected PE were screened, and PE was diagnosed in 674 patients. Three-month follow-up was completed in 673 of the 674 patients with PE. The authors found that recurrent VTE occurred in a small percentage of patients treated for an acute PE, and the majority of recurrent VTEs were fatal. Immobilization, hospitalization, age, COPD, and malignancies were risk factors for recurrent VTE, bleeding, and mortality.

Example 2

Short et al[11] performed a retrospective cohort study to examine the effect of β-blockers in the management of COPD. They searched a disease-specific database of patients with COPD and linked to the Scottish morbidity records of acute hospital admissions, the Tayside community pharmacy prescription records, and the General Register Office for Scotland death registry. A total of 5,977 patients aged $>$ 50 years with a diagnosis of COPD were identified and divided into two groups according to β-blocker use. The study found that β-blockers might reduce mortality and COPD exacerbations when added to established inhaled stepwise therapy for COPD, independently of overt cardiovascular disease and cardiac drugs, and with no adverse effects on pulmonary function.

Example 3

Skull et al[12] described the epidemiology of community-acquired pneumonia (CAP) in elderly Australian subjects. Using a case-cohort design, cases with CAP were identified as in-patients aged $\geq$ 65 years with International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification codes J10 to J18 admitted over 2 years to two tertiary hospitals. The cohort sample was randomly selected from all hospital discharges and frequency-matched to case subjects according to month. A total of 4,772 inpatients were studied. The strongest predictors of CAP were previous pneumonia, history of other respiratory disease, and aspiration. ICU admission, renal disease, and increasing age were the strongest predictors of mortality, whereas influenza vaccination conferred protection.

Benefits and Downside of Cohort Studies

A major advantage of the cohort study design is the ability to study multiple outcomes that can be associated with a single exposure or multiple exposures in a single study. Even the combined effect of multiple exposures on the outcome can be determined. Cohort study designs also allow for the study of rare exposures. Investigators can specifically select subjects exposed to a certain factor. Furthermore, cohort studies often have broader inclusion and fewer exclusion criteria compared with randomized controlled trials. The investigators may obtain large samples and reach greater power in statistical analysis relative to a randomized controlled trial. For these reasons, results from cohort studies may be more generalizable in clinical practice. Finally, the longitudinal nature of cohort studies means that changes in levels of exposure over time, and changes in outcome, can be measured to provide insight into the dynamic relation between exposure and outcome.

Prospective and retrospective studies have different strengths and weaknesses. Prospective cohort studies are conducted from the present time to the future, and thus they have an advantage of being accurate regarding the information collected about exposures, end points, and confounders. The disadvantage could be the long period of follow-up while waiting for events to occur, leading to vulnerability to a high rate of loss to follow-up.

Retrospective studies rely on data collected in the past to identify both exposures and outcomes. These studies use data that have already been collected, such as would be obtained using a database extracted from electronic medical records. Thus, cohort studies are often time-efficient and cost-effective. However, many retrospective cohort studies use data that were collected in the past for another objective. Hence, the investigators lack control over the collection of data. The measurement of variables might be inaccurate or inconsistent, which results in a source of information bias. Table 1 summarizes the advantages and disadvantages of cohort studies.

Study Subject Considerations

There are several considerations related to the subjects of a cohort study. These include selection of an appropriate sample of the population of interest, the sampling method that will be used, access to longitudinal data for the subjects chosen, and the sample size required to properly power the study. The criteria for inclusion and exclusion should be determined at the study design stage. The study subjects selected should be appropriate for the study question and should be generalizable to the population of interest. Avoiding bias

TABLE 1] Advantages and Disadvantages of Cohort Studies

Advantages	
Can investigate multiple outcomes that may be associated with multiple exposures	
Able to study the change in exposure and outcome over time	
Good for examining rare exposures	
Can measure incidence of outcome	
May be able to infer causality	
Prospective Study	Retrospective Study
Able to control design, sampling, data collection, and follow-up methods	Time-efficient and inexpensive
Can measure all variables of interest	Easy to obtain large sample
Disadvantages	
Susceptible to loss to follow-up compared with cross-sectional studies	
Confounding variables are the major problem in analyzing the data compared with RCTs	
Prospective Study	Retrospective Study
May be expensive to conduct	Less control over variables
Time-consuming	Susceptible to information bias and recall bias

RCTs = randomized controlled trials.

in subject selection, ensuring generalizability of the results, and determining the feasibility of performing an adequately powered study are crucial elements of the study design.

Sample size determination for cohort studies has been widely discussed in the literature.[13-15] A comparison of incidence rates is usually the major aim of a cohort study. Assume that and $p_1$are the incidence rates of the end point of interest in the exposed and unexposed samples. The sample size is typically calculated based on the following statistical hypothesis:

$$H_0 : p_1 = p_2 \; vs \; H_1 : p_1 \neq p_2$$

The sample size formula can be found in Fleiss et al.[16] For paired cohort studies or case-cohort designs, the formulae can be found in Kasiulevičius et al[17] and Cai and Zeng.[9] When the outcome of interest in a cohort study is continuous (although it is less common), we would like to compare the means of two cohorts. The formula based on the minimum detectable difference can be found in Woodward.[18]

It is also important to consider subject loss to follow-up in designing a cohort study. Any sample size calculated should be inflated to account for the expected dropouts. For instance, if the dropout rate is expected to be 10%, the estimated sample size would be *N* multiplied by 1/(1-0.1). A general discussion about sample size determination is presented in the article by Wang and Ji[19] included in this supplemental issue of *CHEST*. As part of that article, an online calculator has been developed to help readers to perform the sample size estimation for cohort studies. It can be found at http://riskcalc.org:3838/samplesize/.

Statistical Considerations

Investigators often use cohorts to assess the association between multiple exposures and multiple outcomes over time and to build prognostic/prediction models. The modeling and analysis strategy could be sophisticated in cohort studies. Here we emphasize a few important aspects of statistical analysis.

Bias

Bias may be defined as any systematic error in a clinical study that results in an incorrect estimate of the true effect of an exposure on the outcome. A major source of potential bias in cohort studies is due to loss to follow-up. This occurs due to dropouts or death, which often occurs in studies with long follow-up durations. A general rule of thumb requires that the loss to follow-up rate does not exceed 20% of the sample.[20] It is recommended that investigators examine any systematic differences related to the outcome and/or exposures between those who completed the study and those who were lost to follow-up. Methods of minimizing loss to follow-up in a prospective cohort study have been comprehensively discussed by Hulley et al.[21] We suggest that the investigators report median follow-up for patients without the event or the number followed up without an event at a given follow-up time.
For example, consider the case of a cohort of 1,000

patients with COPD treated in 1970 and followed up until 2010. The median follow-up for all patients might be far less than the median follow-up for patients who survived. The latter statistic may provide a more accurate impression of how long the cohort had been followed up. Now assume that in 2009, a second cohort of 2,000 patients were added to the study. The median follow-up for survivors will now be around 1 year, which is again misleading. An alternative would be to report a statistic such as "312 patients have been followed up without a death event for at least 35 years."

There are many other types of bias in clinical studies. Examples include allocation bias, prevalence-incidence bias, recall bias, and detection bias. In the accompanying cross-sectional study article included in this supplemental issue of *CHEST*, Wang and Cheng[22] provide a detailed discussion regarding common types of biases and their definitions in clinical studies.

Confounding

Confounding often occurs in cohort studies. For a variable to be a confounder, it should meet three conditions: (1) be associated with the exposure being investigated; (2) be associated with the outcome being investigated; and (3) not be in the causal pathway between exposure and outcome. Confounding could result in a distortion of the effects; it may lead to overestimation or underestimation of an effect, or even reverse the direction of an effect. For example, a study found that alcohol consumption was associated with lung cancer. A person who drinks alcohol is more likely to smoke, and smoking is a risk factor for lung cancer. Controlling for the potential confounding effect of smoking may show that there is no association between alcohol consumption and lung cancer. Figure 3 shows the relation among the exposure, confounder, and outcome in this example.

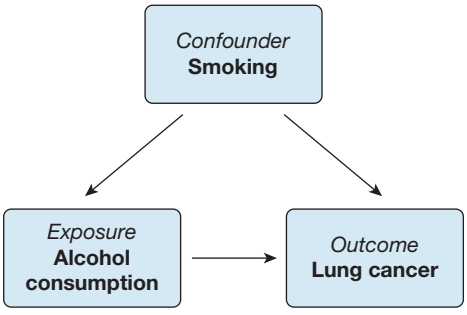

Figure 3 – *An example of illustrating the relation among exposure, confounder, and outcome.*

Many statistical methods can be applied to control for confounding factors, both at the design stage and in the data analysis. The aim of controlling for confounding is to make the groups as similar as possible with respect to the confounders. At the design stage, restriction is a common method for controlling confounders. The investigators first identify potential confounding factors based on previous studies or the knowledge that confounding is biologically plausible. The investigators then limit participation in the study to individuals who are similar with respect to those confounders. For example, a lung cancer study restricted to smokers will eliminate any confounding effect of smoking. A drawback of this method is that it may be difficult to generalize the findings to the rest of the population.

At the analysis stage, stratification is one of the popular controlling methods. Stratification allows the association between exposure and outcome to be examined within different strata of the confounding variables. For example, a study is conducted to examine the association between lung cancer and exposure to asbestos. To control for smoking, the study population could be stratified according to smoking status. The association between exposure to asbestos and cancer can then be assessed separately within each stratum. An issue with stratifying is that strata with more individuals will tend to have a more precise estimate of the association (with a smaller SE) than strata with fewer individuals. For this reason, the Cochran-Mantel-Haenszel method is often used in stratification analysis. It allows calculating an overall and adjusted effect estimate of a given exposure for a specific outcome by combining (pooling with weight) stratum-specific relative risks or OR.[23]

Multivariable regression analysis is a model-based method to control for confounding. One builds a multivariable regression model for the outcome and exposure as well as other confounding variables. Based on the regression equation, the effect of the variable of interest can be examined with confounding variables held constant statistically. Multivariable regression has the advantage in that it can control simultaneously for more confounding variables than can stratification. It has the disadvantage in that this model may not fit the data well. The investigators have to be careful to use accepted variable selection procedures.

The propensity score method is also popular for controlling confounding.[24] The propensity score is the probability of treatment/exposure assignment conditioned on observed baseline characteristics. It allows investigators to mimic some of the characteristics

of a randomized controlled trial in a cohort study. In practice, propensity score analysis involves a two-step procedure. The first stage of the analysis is to estimate the propensity score. We consider the exposure variable as the response variable and build a regression model with the variables that influence exposure group membership, such as sex and age. This model is used to give each subject a propensity score that measures the propensity (probability) to be exposed given the subject's characteristics. In the second stage, the outcomes of interest are compared between exposed and unexposed following adjustment for propensity scores. There are different approaches on using propensity scoring but they all yield similar results: matching on the propensity score, stratification on the propensity score, inverse probability of treatment weighting using the propensity score, and covariate adjustment using the propensity score.[25] Although propensity score methods are powerful, they involve sophisticated statistical techniques. A deep understanding of the methodology is necessary when implementing the specific analysis. There are other methods of controlling for confounding such as instrumental variable analysis and regression discontinuity design; details are provided in Merrill[1] and Rothman et al.[2] The application of directed acyclic graphs in observational studies assessing associations is described in a separate article included in this supplemental issue of *CHEST* (Etminan et al[26]).

Model Building

Model building is often crucial in cohort studies. Investigators may need to build explanatory models or predictive models. In explanatory modeling, one is interested in identifying variables that have a scientifically meaningful and statistically significant relation with an outcome. In predictive modeling, the goal is to predict the probability of or the risk for the presence (diagnosis) or future occurrence (prognosis) of an outcome for an individual. When building a model (explanatory or predictive), the variables selected for inclusion should be based on the critical consideration of relevant literature or knowledge of medical experts. Use of stepwise selection should be restricted to a limited number of circumstances, such as during the initial stages of developing a model, or if there is poor knowledge of what variables might be predictive.[27] Modern shrinkage or penalization procedures such as LASSO/least absolute shrinkage and selection operator, elastic net, and their variants are recommended for the study of rare events or when there are a large number of predictors. If predictive models are built, one should include some form of internal validation, such as cross-validation or bootstrapping, particularly in the situation that has no additional external validation performed. Controlling for confounding when building a prediction model is less common than that when modeling to assess for associations/causality. The article by Kattan and Gerds[28] included in this supplemental issue of *CHEST* offers guidance on building prediction models.

Reporting Considerations

We suggest that investigators report their cohort studies following the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, which contains a checklist of 22 items that are considered essential for reporting of observational studies.[29] If multivariable prognostic prediction models are developed in a cohort study to be used in predicting future outcomes in individuals at risk, we recommend that investigators consult the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement.[30,31]

Short List of Questions to Guide the Reviewer

When reviewing a cohort study, consider commenting on the following:

1. **The study cohort.** Was the study cohort well described? Was the method for selection of cohort members and the inclusion/exclusion criteria appropriate? Were there potential biases introduced by the methods chosen? Was the sample size adequate for the primary study question? How was subject dropout, death, and missing data handled?
2. **The exposures and outcomes.** Were they clearly defined? Are there concerns about the accuracy of their measurement? Were there potential biases introduced by the definitions and measurements?
3. **Potential confounders.** Were potential confounders identified based on prior knowledge? Were they properly controlled for in the study design and/or analysis? Are causal directed acyclic graphs included or required?
4. **The interpretation of the strength of the association(s) identified.** Were the measures used to describe the association between the exposure and the outcome clearly described and appropriate? Was the interpretation of the association(s) identified appropriate?

Acknowledgments

Financial/nonfinancial disclosures: None declared.

References

1. Merrill RM. *Introduction to Epidemiology*. 7th ed. Burlington, MA: Jones & Bartlett Publishers; 2015.
2. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2012.
3. Carlson MDA, Morrison RS. Study design, precision, and validity in observational studies. *J Palliative Med*. 2009;12(1):77-82.
4. Euser AM, Zoccali C, Jager KJ, Dekker FW. Cohort studies: prospective versus retrospective. *Nephron Clin Prac*. 2009;113(3): c214-c217.
5. Vandenbroucke JP. Observational research, randomised trials, and two views of medical science. *PLoS Med*. 2008;5(3):e67.
6. Hess DR. Retrospective studies and chart reviews. *Respir Care*. 2004;49(10):1171-1174.
7. Ernster VL. Nested case-control studies. *Prev Med*. 1994;23(5):587-590.
8. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986;73(1):1-11.
9. Cai J, Zeng D. Sample size/power calculation for case–cohort studies. *Biometrics*. 2004;60(4):1015-1024.
10. Nijkeuter M, Söhne M, Tick LW, et al. The natural course of hemodynamically stable pulmonary embolism: clinical outcome and risk factors in a large prospective cohort study. *Chest*. 2007;131(2): 517-523.
11. Short PM, Lipworth SIW, Elder DHJ, Schembri S, Lipworth BJ. Effect of β blockers in treatment of chronic obstructive pulmonary disease: a retrospective cohort study. *BMJ*. 2011;342:d2549.
12. Skull SA, Andrews RM, Byrnes GB, et al. Hospitalized community-acquired pneumonia in the elderly: an Australian case-cohort study. *Epidemiol Infect*. 2009;137(2):194-202.
13. Altman DG. *Practical Statistics for Medical Research*. Boca Raton, FL: CRC Press; 1990.
14. Everitt BS. *Statistical Methods for Medical Investigations*. London, England: Edward Arnold; 1994.
15. Kahn HA, Sempos CT. *Statistical Methods in Epidemiology*. New York, NY: Oxford University Press; 1989.
16. Fleiss JL, Tytun A, Ury HK. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*. 1980;36(2):343-346.
17. Kasiulevičius V, Šapoka V, Filipavičiūtė R. Sample size calculation in epidemiological studies. *Gerontologija*. 2006;7(4): 225-231.
18. Woodward M. Formulae for sample size, power and minimum detectable relative risk in medical studies. *J R Stat Soc Ser D Stat*. 1992;41(2):185-196.
19. Wang X, Ji X. Sample size estimation in clinical research: from randomized controlled trials to observational studies. *Chest*. 2020;158(suppl 1):S12-S20.
20. Conato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research design. *N Engl J Med*. 2000;342:1887-1892.
21. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. *Designing Clinical Research*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2007.
22. Wang X, Cheng Z. Cross-sectional studies: strengths, weaknesses, and recommendations. *Chest*. 2020;158(suppl 1):S65-S71.
23. Sullivan LM. *Essentials of Biostatistics in Public Health*. Burlington, MA: Jones & Bartlett Learning; 2017.
24. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1): 41-55.
25. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavio Res*. 2011;46(3):399-424.
26. Etminan M, Collins GS, Mansournia MA. Using causal diagrams to improve the design and interpretation of medical research. *Chest*. 2020;158(suppl 1):S21-S28.
27. Kattan MW, Vickers AJ. Statistical analysis and reporting guidelines for *Chest*. *Chest*. 2020;158(suppl 1):S3-S11.
28. Kattan MW, Gerds TA. A framework for the evaluation of statistical prediction models. *Chest*. 2020;158(suppl 1):S29-S38.
29. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Int J Surg*. 2014;12(12):1495-1499.
30. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.
31. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63.