Publication status: Not informed by the submitting author

Case-control and case-cohort study designs: methodological considerations

Artur Iuri Alves de Sousa, Elisabeth Carmen Duarte

https://doi.org/10.1590/SciELOPreprints.7327

Submitted on: 2023-11-07
Posted on: 2023-11-08 (version 1)
(YYYY-MM-DD)

Case-control and case-cohort study designs: methodological considerations

Artur Iuri Alves de Sousa
aia.desousa@gmail.com | Universidade de Brasília
https://orcid.org/0000-0001-6688-568X

Elisabeth Carmen Duarte
eduarte@unb.br | Universidade de Brasília
https://orcid.org/0000-0001-9148-5063

Abstract

The "case-control" study is not a homogeneous entity in epidemiology. This article presents a narrative review of characteristics that differentiate "traditional case-control", "population-based case-control", "nested case-control", and "case-cohort" studies. The use of a secondary basis for the choice of controls makes case-control studies particularly vulnerable to selection biases and limits the representativeness of the control group relative to the study population base. The choice of prevalent cases, although useful for maximizing statistical power, undermines the interpretation of associations since determinants of incidence are mixed with determinants of disease duration. "Nested case-control" and "case-cohort" studies partly overcome these limitations because they are based on a true cohort and often use incident cases. "Case-cohort" studies give emphasis to the contrast between case exposure experience and population-based exposure experience, providing additional benefits when there is temporal matching between cases and the reference group. This article provides insights into the methodological arsenal of case-control studies and highlights some of the methodologies discussed - particularly those in which controls are included - favors the use of secondary databases without the need for a key variable that correlates them.

Keywords[1]: Analytical Epidemiology, Case-Control Studies, Cohort Studies, Longitudinal Studies

Delineamentos caso-controle e caso-coorte: considerações metodológicas

Resumo

O delineamento tipo "caso-controle" não é uma entidade homogênea em Epidemiologia. Este artigo apresenta uma revisão narrativa de características que diferenciam os estudos caso-controle, tradicional e de base populacional, caso-controle aninhado e caso-coorte. O uso de uma base secundária para a escolha dos controles deixa os estudos caso-controle particularmente vulneráveis ao viés de seleção e limita a representatividade do grupo controle em relação à base populacional do estudo. A opção por casos prevalentes, útil para maximizar o poder estatístico, prejudica a interpretação das associações já que determinantes da incidência são mesclados com os determinantes da duração da doença. Os "casos e controles aninhados" e "caso-coorte", superam em parte essas limitações, pois são baseados em uma coorte verdadeira e usam frequentemente casos incidentes. Os "caso-coorte" enfatizam o contraste entre a experiência de exposição dos casos e de exposição da base populacional, fornecendo benefícios adicionais quando existe pareamento temporal entre casos e controles. Este artigo oferece elementos para a discussão sobre o arsenal metodológico dos estudos “caso-controle” e destaca as metodologias discutidas – particularmente aqueles em que os controles são inclusivos - favorece o uso de bases de dados secundários sem que haja a necessidade de uma variável chave que as correlacione.

Palavras-chave[1]: Epidemiologia Analítica, Estudos de Casos e Controles, Estudos de Coortes, Estudos longitudinais

Introduction

When randomized experimental epidemiological studies, such as clinical and community-based trials, are well conducted they provide results with a high degree of evidence about causality thanks to their strong internal validity[2-6]. However, in general their cost is high, are time consuming, have low external validity and considerable ethical limitations. Cohort studies have become a more acceptable alternative from the ethical point, in general, have greater external validity than experimental studies. However, it is generally costly and time consuming as it often requires a large number of study subjects and a long follow-up time. This is especially true when the outcomes of interest are rare and/or have long induction and/or latent periods[2-6].

Case-control studies arose as a more efficient alternative for overcoming the cost and time-consuming limitations. Studies retrospective have been documented since the 1920s[7]. However, it was after the 1950s that the used more evident, driven by the need to investigate the increasing chronic degenerative diseases in developing countries[7-9]. In the 1960s the term “case-control” was proposed by Sartwell[10] to unify the different terms used at that time[7]. Nevertheless the need still persists to enhance conceptual and methodological systematization involving case control-studies. The objective of this study was to contribute to the systematization of different designs of case-control studies regarding selected methodological issues.

We carried out a narrative review[11] based on selected articles having an empirical, theoretical and/or conceptual approach that discussed the different types of case-control studies, namely: “traditional", population-based, nested or case-cohort. We searched scientific articles available through the Capes portal, PubMed and Google Scholar. Our search also included books and academic dissertations and theses. The keywords used for the search were: case-control, case-cohort and nested case-control. We did not restrict the language or year of publication.

General consideration and methodological challenges

Case-control studies are relatively quicker than cohort and experimental as well as usually being less costly because they require a smaller numbers subjects to achieve a given level of statistical power[6]. Another advantage is that they do not require prospective follow-up time. However, require periods of retrospective-historical "monitoring" between exposure and outcome by means of subject memory recall, document reviews or some pertinent type of retrospective observation. It makes them very attractive because of their efficiency[12], especially for rare diseases and/or diseases with long latent and induction periods.

When case-control studies are adequately designed and conducted they can produce evidence of causality just as robust or even more than cohort studies can. On the other hand, owing to their retrospective design, which is reverse to the natural history of a disease, some methodological challenges exist when carrying out a case-control study in order to increase its internal validity[2-6].

Adequate retrospective evaluation of determinants of interest

Exposures of interest – as well as other determinants, confounders and effect modifiers – are measured with potential information bias[2,6], because it depends on the memory of the subjects and/or on the quality of document reviews. Moreover, exposures of interest will be measured after the outcome has already occurred and can, therefore, be affected by it[6].

Adequate identification of the "population-base" or "reference population" or "study-base"

For the purposes of standardizing this article, the term "study-base", as proposed by Miettinen[13], will be referred to as "population-base", which is defined as a particular population that has had exposure-experience over time, but which does not need to be representative of the "general

population" nor even of the "general patient population", need to be comparable and representative with regard to the exposure-experience[13]. According to Miettinen[13], the population-base can be primary or secondary of the choice of case series. Primary base refers to a census of the population-base with case of interest identified and selected later, previously identified. Following this, a sample of the population-base is taken to serve as a control group.

Secondary base refers to the fact that the population-base is identified and sampled only after the identification of cases, having the challenge arises with regard to the adequate definition of the population-base: ensuring that the population-base identified correspond to the population from which the case series arose, in order to preserve comparability and representativeness of the subjects[13].

Case series selection

A case-control study should be conceived as a census of the population-base, at, later, selecting a sample, in order for these subjects to form the control group and also measure the same variables of interest among them[13]. In this manuscript this will be called the "case-base", this is cases and controls must come from the same population-base under study. One point in particular that needs to be considered is the option to use prevalent cases or incident cases.

Stopped here

Case-control studies are, in general, considered to be useful for studying rare events, as such, choosing prevalent cases can appear attractive in terms of increasing the size of the case series to be studied. However, has two such important limitations are: i) Prevalence of disease is directly proportional to incidence and to average duration of the disease, whereby duration is influenced by the likelihood of cure and survival time associated with the natural and/or clinical history of the disease[14]. We therefore need to recognize that "prevalent cases" are cases that

have "survived" in the study population with the disease and that they systematically (or are likely to) exclude fatal acute cases and cases that were cured quickly (in the case of curable diseases), being able to introduce bias defined as "survival bias"[15,16]. ii) Prevalent cases can change their habits/behaviors as a result of being diagnosed as having the disease, thus changing the pattern of their past exposure[2,4-6]. If current exposure is measured, a bias called "reverse causality bias" may occur - the “chicken or egg” dilemma - common to all cross-sectional data[3]. Taking this into consideration, investigation of exposure in a case-control study should seek the "time window" of interest to the analysis.

Because of these limitations, choosing incident cases should be prioritized whenever possible. Incident cases are new (or recently diagnosed) cases selected in a sequential manner respecting the definition of primary or secondary base[13]. On the other hand, this methodological option tends to minimize the “survival bias" and the “reverse causality bias”. A guide this decision, it is important to assess existing knowledge about the natural history of the disease and the expected average incubation/induction periods following exposure. For similar reasons, "reverse causality bias" is unlikely given that it is possible to ascertain more clearly that exposure has not been affected by the disease knowledge because exposure in fact precedes diagnosis.

Ratio between the number of cases and controls

In a case-control study, the ratio of the number of controls per case is a relevant issue for overcoming limited case availability (in the context of a rare disease) and protecting the study’s statistical power and efficiency[17-19]. Increasing the number of controls per case is not a useful solution for all types of case-control studies, and it may even have negative consequences for the study. We have some specific situations:

When few cases are available – rare disease context

When case availability is low, the option may be to choose multiple controls for each case, to benefit the study with a certain gain in statistical power. However, potential gain is often jeopardized, as information from a few cases will continue to strongly drive the amount of variance in the measures of associations, limiting the precision of the estimates. If, for example, we consider one of the formulae for the standard error of the natural logarithm of the odds ratio (SE ln(OR)), it is observed this statistics is strongly inflated in the presence of at least one small cell in the 2x2 table.

$$SE_{\ln(OR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Formula 1 – Calculation of Standard Error (SE) of the natural logarithm of the Odds Ratio (ln(OR))[20]

Examples of estimated 95% Confidence Intervals of odds ratios (OR), table 1 shows, taking two hypothetical situations: 1) taking a fixed number of 10 cases and changing the case:control ratio to 1:1(1a), 1:3(1b), 1:6(1c) and 1:9(1d); 2) taking a fixed number of the total sample size (n=100) and using a ratio of 1:9(2a) or an equal ratio of 1:1(2b) for cases and controls (Table 1).

In Context 1 (Table 1) precision is improved (narrowing of the confidence intervals) between situation "1a" and "1b", but not so much between situation “1b” and “1c” and “1d”, there is not much point in increasing the controls indefinitely if the small number of cases will, to a large extent, strongly influence the study’s power, keeping small numbers in cells a (exposed cases) and c (unexposed cases). There are indications that there is no increase in the study estimates that justifies the inclusion of a further four controls per case[17-19,21]. In Context 2, for a fixed total sample size (e.g. n=100), the gain in precision occurs at situation (2b) – with equal

distribution of cases and controls – when compared to situation (2a) – control oversampling. It can be seen that it is more efficient, in terms of precision vs. effort (expense), to distribute the subjects equally between cases and controls. Obviously this would only be possible if the number of cases available – albeit limited – is sufficient to ensure enough power to access the study hypothesis.

Table 1 – Simulation of OR estimates considering the results of studies with cases and controls in different contexts.

Context		Case and control ratio	n cases	n controls	n total	OR (95% Confidence Interval)(1)	
1a		1:1	10	10	20	3.9	(0.3 – 45.6)
1b		1:3	10	30	40	3.9	(0.6 – 23.4)
1c		1:6	10	60	70	3.9	(0.8 – 19.0)
1d	2a	1:9	10	90	100	3.9	(0.8 – 17.6)
	2b	1:1	50	50	100	3.9	(1.3 – 11.6)

Note (1) OR = The odds ratio of 3.9 is constant taking, for all examples, exposure prevalence rates set at 30% for cases and 10% for controls. 95% Confidence Intervals are based on Conover, 1999.

When exposure increases risk of the outcome – exposure is a risk factor

It is advisable to optimize the ratio in favor of the controls. Given that few exposed controls are expected, having a ratio in favor of the controls (e.g. 1:2 or 1:3) would protect the study from very small cells and thus protect the study power.

When the exposure factor reduces the risk of the outcome – exposure is a protective factor

It is advisable to optimize the ratio in favor of the cases (given that few exposed cases are expected). This is important in order to increase the number of exposed cases, avoiding few subjects with this profile and thus protecting the study power.

When it is not known whether the factor is a risk or a protective factor

The best scenario would be to select 1:1 (one control per each case), as discussed (1.4.i, scenario 2b – Table 1). For a given fixed sample size, dividing these subjects equally between cases and controls will potentially avoid, to a certain extent, very small cells in the 2x2 table. This is good in terms of the study's efficiency and statistical power. Obviously, this balance will depend on how exposure is in fact associated in the end with the status of being case or control, as also discussed (1.4.ii and 1.4.iii).

Exclusive or inclusive controls

A general aspect when defining the choice of controls in case-control studies is that, controls are subjects who, if they had the disease that is being studied, they would be captured by the study to comprise the case series13. A point to be discussed is whether controls are exclusive or inclusive, i.e., can controls also be or not be cases.

Exclusive controls (controls are necessarily non-cases):

Represented by a group of subjects who do not have the disease that is being studied (i.e. they are necessarily non-cases), who therefore represent the non-ill population of the "population-base". Two situations can be considered:

a) controls represent subjects arbitrarily selected from a non-case group – secondary base;

b) controls represent necessarily non-cases directly selected from the population-base (or from a "cohort study") from which the incident cases emerged - primary base.

Inclusive controls (controls can be cases):

Represented by a sample of the population-base of the study from which the cases emerged - incident cases - and can become cases. These controls can be time matched, or not, with cases. Matching according to potentially confounding variables is a traditional procedure in case-control studies22-24. Matching as discussed in this article refers to “time” matching controls and (incident) cases. This situation can be clearly exemplified based on a cohort in progress

a) time matched controls: the control is identified the time incident case is identified. It should be noted that this currently non-ill control may become a case during the follow-up of the subjects of the cohort. This cohort may actually be a formal cohort study in the context of a nested case-control study, or it is a "by definition" cohort considering the natural history of the disease.

b) non-time matched controls: the controls will be selected at baseline of the cohort (time zero of the follow-up) and cases will be identified later as cohort follow-up occurs. As case selection takes place after control selection, the controls selected at baseline may become cases. In this case, time matching is absent, even though some level of stratification by other variables may be done.

Case-control study typologies

"Traditional" Case-control study (also called “Cumulative” Case-control or “Classic” Case-control).

The “traditional” case-control studies can be classified as population-based or not population-based studies.

a) “Traditional” case-control not population-based study: cases will be based on a random sample or on the totality (census) of eligible cases that have hypothetically emerged from a

given population base. The choice of controls will be based on a random sample of all controls also hypothetically originating - in theory - from the same population-base (from which the cases emerged)(Table 2-a). Given that this is not a population-based study, an effort must be made to represent the same base population as that of the cases, and whereby the exposure experienced by the controls is a good estimate of the exposure experience of the hypothetical population-base. Owing to the difficulty in guaranteeing such representativeness, sometimes controls from different populations are chosen in the sense of validating (to a certain extent) estimated association between exposure and outcome: e.g. hospital and neighborhood controls.

b) "Traditional" population-based case-control study:

The population-based case-control study is a case-control in which the totality or a representative random sample of the population-base is identified (primary base). Later, individuals are classified as cases (usually prevalent cases) or controls (non-cases). Finally, experience of exposure will be measured and compared between the case group and the control group[23](Table 2-b). In this design is possible that the ratio between cases and controls tends to reproduce what happens in the study population. This usually is no possible in other case-control study designs, where the ratio case:control is judged by the researcher as discussed earlier. For example, if disease prevalence is 25% in the population-base, the case and control ratio in the population-based study will tend to be one case per three controls.

Cohort-based case-control studies:

Case-controls studies within a cohort or nested case-control studies are being denominated by different terminologies[2-6,23]. They are classified regarding the use or not of inclusive controls and the time-matching option. Briefly, they select incident cases and an appropriate subsample of controls from a true cohort study. The choice of this type of design is especially advantageous

when baseline biologic specimens are stored for future analysis and/or when predictors variables can be measured at the end of the study follow-up[23,25]. In this case, cohort-based case-control studies are cost saving because controls come from a subgroup of potential controls sampled from the entire cohort at baseline or at risk when cases are identified, and explanatory variables will only be "processed" for this subgroup and not for the entire cohort population. Study efficiency is optimized particularly when expensive tests to assess exposures are involved.

c) Nested case-control study with exclusive controls:

One way of ensuring that incident cases and controls are comparable and originate from the same population-base is to select them from a true cohort. This type of design is called a nested case-control study, which means a case-control nested in a cohort[22,25]. In this design, cases are necessarily incident cases, identified and selected from a population-base which is monitored prospectively (true cohort). Moreover, the controls are individuals who have not developed the disease (exclusive controls) during follow-up on the same cohort. The controls therefore represent a sample of total non-cases up to a given time t during the cohort[14,23] (Table 2-c). At the researcher's criteria, controls can also be selected at the time each case is identified, i.e. case and control time matching; making sure that they did not become a case latter on during follow-up (e.g. exclusive controls)[25](Table 2-c).

d) Case-cohort study (nested case-control study with inclusive controls).

These types of design is also called: case-base study, case-referent study, "inclusive" case-control study, hybrid case-control study or nested case-cohort study. For simplicity we will call them as "case-cohort study" in this article.

The case-cohort study is an alternative to the case-control described above and an alternative to the cohort study. Was introduced by Prentice[22] and its particularity is the fact of the controls

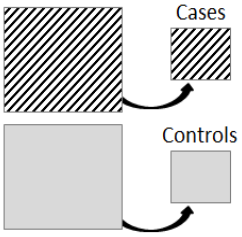
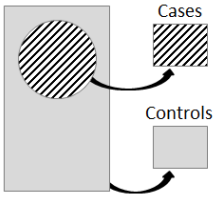
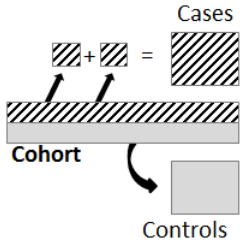
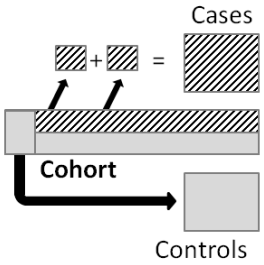
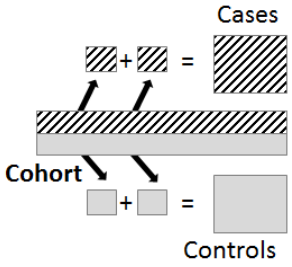
being selected randomly from the study population, regardless of being or not being a case of the study in question[26], i.e. controls are inclusive, meaning that the same individual can be included in the case group and in the control group. Similarly, to nested case-control, cases and controls also originate from a true cohort, whereby cases are necessarily incident and identified and selected from a given cohort. The controls are inclusive controls and identified either at baseline or during monitoring of the cohort from which the cases emerge. Controls therefore no longer represent the group of non-cases in the cohort, but rather the total of the subjects (the sum of cases and non-cases) of the entire cohort. Consequently, in this type of design, the terms "reference group" or "population-base" or "cohort" are, at times, preferable to the use of the term "control". The advantage of method is that the controls provide a reference base for estimating incidence in the population-base from which the cases have emerged[23].

Design can be unmatched or time matched. i.e. selection of the "controls" can be at the cohort baseline or at the time cases are identified[24].

d.1.) Non-time matched case-cohort study: the "controls" represent a random sample of the entire cohort at the baseline (beginning of follow-up), regardless of whether they will become cases or not in the future during follow-up (Table 2-d.1).

*d.2) Time matched Case-cohort study***:** the "controls" will be selected at the time each case is identified, i.e., time matched[24]. Selection will take place from among all the cohort subjects, regardless of whether the subject was identified or not as a case at the time of selection, or whether they will or will not become a case in the future[24] (Table 2-d.2).

Table 2 – Case-control design typologies and selected characteristics.

Name	Basic architecture	Population-base	Types of control and cases	Time matching
a) ”Traditional” case-control		Non-population-based - Secondary base	Exclusive controls Cases can be incident or prevalent	Does not apply
b) Population-based case-control		Population-based - Primary base	Exclusive controls Cases can be incident or prevalent	No
c) Nested case-control with exclusive controls		Population-based - Primary base	Exclusive controls Incident cases	Yes or No
d.1) Non-time matched case-cohort		Population-based – Primary base	Inclusive controls Incident cases	No
d.2) Time matched case-cohort		Population-based – Primary base	Inclusive controls Incident cases	Yes

Below we discuss some of the advantages and disadvantages of the different designs.

Table 3 – Summary of advantages and disadvantages of the different designs.

Name	Advantages	Disadvantages
a) Traditional" case-control	Ease of finding an adequate number of cases and controls, due above all to the use of prevalent cases and institutionalized cases	Primary base is not identified: potential for selection bias. Included "prevalent cases": survivor bias and possibility of temporality error
b) Population-based case-control	Can be ease in finding cases and controls when cases are prevalent. Base is primary: minimize selection bias	Cases are prevalent: survivor bias and possibility of temporality error
c) Nested case-control with exclusive controls	Use primary base (defined by the previously identified cohort): minimize selection bias. The use of incident cases: improves measurement of exposure variables and minimizes reverse temporality bias	Controls can only be identified after cases have been identified Exclusive controls Controls represent "non cases" and not the "population-base"
d.1) Non-time matched case-cohort	Use primary base: minimize selection bias The use of incident cases: improves measurement of exposure variables and minimizes reverse temporality bias. Controls (population-base) represent all the exposure experience of the reference population at the baseline of the cohort	The outcomes of interest of individuals in the cohort need to be known. Controls are identified from the base (inclusive controls), they represent all the exposure experience of the reference population at the cohort baseline and not at the time when the case occurred Non-time matched: can increase the potential for confounding and not ensure representativeness of the base at the time the case occurs
d.2) Time matched case-cohort	Use primary base: minimize selection bias The use of incident cases: improves measurement of exposure variables and minimizes reverse temporality bias. Controls represent the exposure experience of the whole reference population (and not just that of the "non-cases") at the time cases are identified. This time matching minimizes confounding due to factor (temporal) and ensures population-base representativeness for each case	Outcome frequencies in the complete group of cohort individuals need to be known. Design structure may be more complex than the previous ones

Discussion

The demand for more efficient analytical observational studies to address outcomes that are rare or have long induction and latent periods has been perceived for decades and is gradually being met through the development of retrospective studies. Since the end of the 1960s, the usefulness of case-control studies has been widely demonstrated. Even so, over the course of their history, several challenges have been identified and different methodologies have been developed. The different methodological approaches discussed in this article included the traditional case-control study, the population-based, the nested with exclusive controls and the case-cohort study (time matched or not time matched).

In traditional case-control type studies, the most evident difficulty is adequate choice of controls and to preventing the introduction of selection bias. This is because these are studies that use a secondary base. This limitation is clearly overcome in the other study designs discussed in this article. Population-based case-control studies overcome, in part, this challenge relating to adequate identification of controls and prevention of selection bias. Notwithstanding, they are usually restricted to prevalent cases given that the base population is usually defined in a cross-sectional manner.

Nested case-control studies using exclusive controls and case-cohort studies (inclusive controls), in turn, are cohort-based studies and, thus, naturally allowing identification of incident cases. In contrast to cohort studies, they are justifiable owing to the gain in efficiency arising from reduction in cost and time spent, given that in these studies the data processed and analyzed relate only to selected cases and controls and not to the entire cohort. This gain in efficiency is desirable when measurements require costly procedures[22-24,27]. Another advantage is that these two types of designs retain several of the advantages of cohort studies, in particular those associated with adequate measurement of exposure variables, before the occurrence of the outcomes, and minimizing temporality bias and other types of measurement errors.

An advantage of case-cohort studies in relation to nested case-control using exclusive controls is that the same cohort sample can be used as a comparison group for studying various diseases in case-cohort studies, whereas this is not possible with nested case-control studies[23,24]. In non-time matched case-cohort, the choice of the reference sample (inclusive controls) is independent of the cases that will be identified later and can therefore be used for different types of comparison. In nested case-control studies with exclusive controls, the reference sample depends on prior identification of the cases, so that they are excluded from the control, and therefore they can only be used as comparison group for this specific outcome. As such, with regard to nested case-control design with exclusive controls, the limitation is that in order to conduct it the status of the cohort members needs to be known to allow to build the set of cases, and this is generally unknown before the potential subcohort is selected[28]. Moreover, case-cohort studies have the advantage that measurements can be started at any time after the original cohort has been configured, whereas in nested case-control studies with exclusive controls, cases need to be identified before the controls can be identified and assessed (as they are, by definition, exclusive controls)[24].

According to Sharp[24], the need exists to provide guidance on minimum requirements for applying and preparing/evaluating the case-cohort method, both for authors and for manuscript reviewers and editors of scientific journals[24]. This author noted that there was a relevant increase in the number of published articles using the case-cohort method since 2010, although he identified that a substantial part of the studies classified as case-cohort did not in fact apply this method[24].

In general, in the case of cohort-based case-control studies, attention must always be taken with the quality of the cohort identified (such as representativeness, losses to follow-up, quality of baseline data, among others). Finally, it is worth recalling that when data are available for the entire cohort, without additional cost, nothing is gained by only studying a subsample of

controls. In this case, the entire cohort should be used, and a traditional cohort study should be conducted[23].

Aspects related to analysis of the case-cohort study need to be covered in future publications and it is not part of the scope of this article. It should be mentioned, briefly, that as cases represent a sample of all the population-base cases and that controls represent a sample of the entire population-base, the traditionally calculated Odds Ratio (OR) will be a good estimate of Relative Risk (RR), whereas in traditional case-control studies the OR is usually an overestimation of RR[26].

We hope that this debate will add to reflection on the arsenal of methodological options of available retrospective observational studies called case-control studies and that it will encourage the most adequate choice for different contexts of epidemiological research, taking into account their limitations and advantages.

References

1. Descritores em Ciências da Saúde: DeCS [Internet]. ed.2017. São Paulo (SP): BIREME/OPAS/OMS. 2017 [atualizado 2017 Mai; citado 2017 Jun 13]. Disponível em:http://decs.bvsalud.org

2. Gordis L. Epidemiology. Philadelphia, USA: Elsevier Saunders; 5th ed.; 2014

3. Hennekens CH, Buring JE, Mayrent SL. Epidemiology in medicine. Boston, USA: Little, Brown&Co.; 1987

4. Kleinbaum DG, Sullivan KM & Barker ND. A pocket guide to Epidemiology. New York, USA:Springer; 2007

5. Kleinbaum DG, Kupper LL & Morgenstern H. Epidemiologic research: Priciples and quantitative methods. Belmont, CA: Lifetime Learning Publications; 1982

6. Pereira, MG. Epidemiologia: teoria e prática. Guanabara Koogan, 2012

7. Rêgo MAV. Aspectos históricos dos estudos caso-controle. Caderno de Saúde Pública, Rio de Janeiro, 17(4):1017-1024, jul-ago, 2001

8. Doll R, Hill AB. Smoking and carcinoma of the lung: Preliminary report. British Medical Journal, 2:739-748, 1950

9. Schrek R, Baker LA, Ballard GP, Dolgoff S. Tobacco smoking as an etiologic factor in disease. I. Cancer. American Association for Cancer Research, 10:49-5, 1950

10. Cole P. The evolving case-control study. Journal of Chronic Diseases, Vol.32:15-27, 1979

11. Toledo JA, Rodrigues MC. Teoria da mente em adultos: uma revisão narrativa da literatura. Bol. Acad. Paulista de Psicologia, São Paulo, Brasil-V.37, no92, p.139-156. 2017

12. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology – 3rd ed. p.; cm. Philadelphia, PA 19106 USA. Lippincott Williams & Wilkins. 2008

13. Miettinen, OS. The “case-control” study: valid selection of subjects. Journal of Chronic Diseases Vol.38, No.7, pp. 543-548, 1985

14. Fletcher RH, Fletcher SW. Epidemiologia clínica: elementos essenciais (4ª edição). Porto Alegre: Editora Artmed, 2006

15. Lima-Costa MF, Barreto SM. Tipos de estudos epidemiológicos: conceitos básicos e aplicações na área do envelhecimento. Epidemiologia e Serviços de Saúde. Volume12 - Nº4 - out/dez de 2003

16. Szklo M, Nieto FJ. Epidemioloy: beyond the basics (2ª edição). Jones and Bartlett Publishers, 2007

17. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of Controls in Case-Control Studies. I – Principles. American Journal of Epidemiology Vol.135 No.9, 1992

18. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of Controls in Case-Control Studies. II – Types of Controls. American Journal of Epidemiology Vol. 135 No.9, 1992

19. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of Controls in Case-Control Studies. III – Design options. American Journal of Epidemiology Vol.135 No.9, 1992

20. Conover WJ. Practical nonparametric statistics (3rd ed.). John Wiley & Sons, INC, 1999 ISBN 0-471-16068-7

21. Walter SD. Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. American Journal Epidemiology. 105(4):387–97, 1977. In Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Lippincott-WoltersKluwer; 2008

22. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika, Vol.73, No.1 Apr., pp.1-11.1986

23. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Delineando a pesquisa clínica (3ª edição). Porto Alegre: Editora Artmed, 2008

24. Sharp SJ, Poulaliou M, Thompson SG, White IR, Wood AM. A Review of Published Analyses of Case-Cohort Studies and Recommendations for Future Reporting. PLoS ONE 9(6):e101176, 2014. doi:10.1371/journal.pone.0101176

25. Ganna A, Reilly M, Faire U, Pedersen N, Magnusson P, Ingelsson E. American Journal of Epidemiology. 175(7):715–724, 2012. DOI:10.1093/aje/kwr374

26. Le Polain de Waroux O, Maguire H, Moren A. The case-cohort design in outbreak investigations. Euro Surveill. 2012;17(25):pii=20202

27. Cai T, Zheng Y. Evaluating prognostic accuracy of biomarkers in nested case–control studies. Biostatistics(2012), 13, 1, pp.89–100. doi:10.1093/biostatistics/kxr021

28. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the Whole Cohort in the Analysis of Case-Cohort Data. American Journal of Epidemiology Vol.169, No.11, 2009. DOI:10.1093/aje/kwp055

Declaration of conflict of interest

The authors declare that there is not conflict of interest to mention.

Authors' contributions: SOUSA AIA and DUARTE EC took part in the study conception and design, wiring and critically reviewing the intellectual content, and approving the final version of the manuscript.

This preprint was submitted under the following conditions:

- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.