Sampling: how to select participants in my research study?*

Jeovany Martínez-Mesa[1]
Rodrigo Pereira Duquia[3]
João Luiz Bastos[4]
David Alejandro González-Chica[2]
Renan Rangel Bonamigo[3]

DOI: http://dx.doi.org/10.1590/abd1806-4841.20165254

Abstract: Background: In this paper, the basic elements related to the selection of participants for a health research are discussed. Sample representativeness, sample frame, types of sampling, as well as the impact that non-respondents may have on results of a study are described. The whole discussion is supported by practical examples to facilitate the reader's understanding. Objective: To introduce readers to issues related to sampling.
Keywords: Dermatology; Epidemiology and biostatistics; Epidemiologic studies; Sample size; Sampling studies

INTRODUCTION

The essential topics related to the selection of participants for a health research are: 1) whether to work with samples or include the whole reference population in the study (census); 2) the sample basis; 3) the sampling process and 4) the potential effects nonrespondents might have on study results. We will refer to each of these aspects with theoretical and practical examples for better understanding in the sections that follow.

TO SAMPLE OR NOT TO SAMPLE

In a previous paper, we discussed the necessary parameters on which to estimate the sample size.[1] We define sample as a finite part or subset of participants drawn from the target population. In turn, the target population corresponds to the entire set of subjects whose characteristics are of interest to the research team. Based on results obtained from a sample, researchers may draw their conclusions about the target population with a certain level of confidence, following a process called statistical inference. When the sample contains fewer individuals than the minimum necessary, but the representativeness is preserved, statistical inference may be compromised in terms of precision (prevalence studies) and/or statistical power to detect the associations of interest.[1] On the other hand, samples without representativeness may not be a reliable source to draw conclusions about the reference population (i.e., statistical inference is not deemed possible), even if the sample size reaches the required number of participants. Lack of representativeness can occur as a result of flawed selection procedures (sampling bias) or when the probability of refusal/non-participation in the study is related to the object of research (nonresponse bias).[1,2]

Although most studies are performed using samples, whether or not they represent any target population, census-based estimates should be preferred whenever possible.[3,4] For instance, if all cases of melanoma are available on a national or regional database, and information on the potential risk factors are also available, it would be preferable to conduct a census instead of investigating a sample.

Received on 15.10.2015
Approved by the Advisory Board and accepted for publication on 02.11.2015
* Study performed at Faculdade Meridional - Escola de Medicina (IMED) – Passo Fundo (RS), Brazil.
Financial Support: None.
Conflict of Interest: None.

1 Faculdade Meridional (IMED) – Passo Fundo (RS), Brazil.
2 University of Adelaide – Adelaide, Australia.
3 Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA) – Porto Alegre (RS), Brazil.
4 Universidade Federal de Santa Catarina (UFSC) - Florianópolis (RS), Brazil.

However, there are several theoretical and practical reasons that prevent us from carrying out census-based surveys, including:

1. Ethical issues: it is unethical to include a greater number of individuals than that effectively required;

2. Budgetary limitations: the high costs of a census survey often limits its use as a strategy to select participants for a study;

3. Logistics: censuses often impose great challenges in terms of required staff, equipment, etc. to conduct the study;

4. Time restrictions: the amount of time needed to plan and conduct a census-based survey may be excessive; and,

5. Unknown target population size: if the study objective is to investigate the presence of premalignant skin lesions in illicit drugs users, lack of information on all existing users makes it impossible to conduct a census-based study.

All these reasons explain why samples are more frequently used. However, researchers must be aware that sample results can be affected by the random error (or sampling error).[3] To exemplify this concept, we will consider a research study aiming to estimate the prevalence of premalignant skin lesions (outcome) among individuals >18 years residing in a specific city (target population). The city has a total population of 4,000 adults, but the investigator decided to collect data on a representative sample of 400 participants, detecting an 8% prevalence of premalignant skin lesions. A week later, the researcher selects another sample of 400 participants from the same target population to confirm the results, but this time observes a 12% prevalence of premalignant skin lesions. Based on these findings, is it possible to assume that the prevalence of lesions increased from the first to the second week? The answer is probably not. Each time we select a new sample, it is very likely to obtain a different result. These fluctuations are attributed to the "random error." They occur because individuals composing different samples are not the same, even though they were selected from the same target population. Therefore, the parameters of interest may vary randomly from one sample to another. Despite this fluctuation, if it were possible to obtain 100 different samples of the same population, approximately 95 of them would provide prevalence estimates very close to the real estimate in the target population – the value that we would observe if we investigated all the 4,000 adults residing in the city. Thus, during the sample size estimation the investigator must specify in advance the highest or maximum acceptable random error value in the study. Most population-based studies use a random error ranging from 2 to 5 percentage points. Nevertheless, the researcher should be aware that the smaller the random error considered in the study, the larger the required sample size.[1]

SAMPLE FRAME

The sample frame is the group of individuals that can be selected from the target population given the sampling process used in the study. For example, to identify cases of cutaneous melanoma the researcher may consider to utilize as sample frame the national cancer registry system or the anatomopathological records of skin biopsies. Given that the sample may represent only a portion of the target population, the researcher needs to examine carefully whether the selected sample frame fits the study objectives or hypotheses, and especially if there are strategies to overcome the sample frame limitations (see Chart 1 for examples and possible limitations).

SAMPLING

Sampling can be defined as the process through which individuals or sampling units are selected from the sample frame. The sampling strategy needs to be specified in advance, given that the sampling method may affect the sample size estimation.[1,5] Without a rigorous sampling plan the estimates derived from the study may be biased (selection bias).[3]

TYPES OF SAMPLING

In figure 1, we depict a summary of the main sampling types. There are two major sampling types: probabilistic and nonprobabilistic.

NONPROBABILISTIC SAMPLING

In the context of nonprobabilistic sampling, the likelihood of selecting some individuals from the target population is null. This type of sampling does not render a representative sample; therefore, the observed results are usually not generalizable to the target population. Still, unrepresentative samples may be useful for some specific research objectives, and may help answer particular research questions, as well as contribute to the generation of new hypotheses.[4] The different types of nonprobabilistic sampling are detailed below.

Convenience sampling: the participants are consecutively selected in order of apperance according to their convenient accessibility (also known as consecutive sampling). The sampling process comes to an end when the total amount of participants (sample saturation) and/or the time limit (time saturation) are reached. Randomized clinical trials are usually based on convenience sampling. After sampling, participants are usually randomly allocated to the intervention or

Chart 1: Examples of sample frames and potential limitations as regards representativeness

Sample frames	Limitations
Population census	• If the census was not conducted in recent years, areas with high migration might be outdated • Homeless or itinerant people cannot be represented
Hospital or Health Services records	• Usually include only data of affected people (this is a limitation, depending on the study objectives) • Depending on the service, data may be incomplete and/or outdated • If the lists are from public units, results may differ from those who seek private services
School lists	• School lists are currently available only in the public sector • Children/ teenagers not attending school will not be represented • Lists are quickly outdated • There will be problems in areas with high percentage of school absenteeism
List of phone numbers	• Several population groups are not represented: individuals with no phone line at home (low-income families, young people who use only cell phones), those who spend less time at home, etc.
Mailing lists	• Individuals with multiple email addresses, which increase the chance of selection compared to individuals with only one address • Individuals without an email address may be different from those who have it, according to age, education, etc.

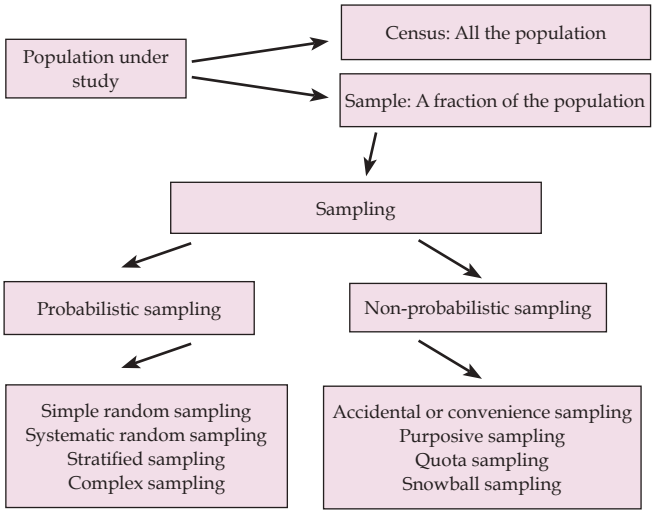

Figure 1: Sampling types used in scientific studies

control group (randomization).[3] Although randomization is a probabilistic process to obtain two comparable groups (treatment and control), the samples used in these studies are generally not representative of the target population.

Purposive sampling: this is used when a diverse sample is necessary or the opinion of experts in a particular field is the topic of interest. This technique was used in the study by Roubille et al, in which recommendations for the treatment of comorbidities in patients with rheumatoid arthritis, psoriasis, and psoriatic arthritis were made based on the opinion of a group of experts.[6]

Quota sampling: according to this sampling technique, the population is first classified by characteristics such as gender, age, etc. Subsequently, sampling units are selected to complete each quota. For example, in the study by Larkin et al., the combination of vemurafenib and cobimetinib versus placebo was tested in patients with locally-advanced melanoma, stage IIIC or IV, with BRAF mutation.[7] The study recruited 495 patients from 135 health centers located in several countries. In this type of study, each center has a "quota" of patients.

"Snowball" sampling: in this case, the researcher selects an initial group of individuals. Then, these participants indicate other potential members with similar characteristics to take part in the study. This is frequently used in studies investigating special populations, for example, those including illicit drugs users, as was the case of the study by Gonçalves et al, which assessed 27 users of cocaine and crack in combination with marijuana.[8]

PROBABILISTIC SAMPLING

In the context of probabilistic sampling, all units of the target population have a nonzero probability to take part in the study. If all participants are equally likely to be selected in the study, equiprobabilistic sampling is being used, and the odds of being selected by the research team may be expressed by the formula: P=1/N, where P equals the probability of taking part in the study and N corresponds to the size of the target population. The main types of probabilistic sampling are described below.

Simple random sampling: in this case, we have a full list of sample units or participants (sample basis), and we randomly select individuals using a table of random numbers. An example is the study by Pimenta et al, in which the authors obtained a listing from the Health Department of all elderly enrolled in the Family Health Strategy and, by simple random sampling, selected a sample of 449 participants. [9]

Systematic random sampling: in this case, participants are selected from fixed intervals previously defined from a ranked list of participants. For example, in the study of Kelbore et al, children who were assisted at the Pediatric Dermatology Service were selected to evaluate factors associated with atopic dermatitis, selecting always the second child by consulting order.[10]

Stratified sampling: in this type of sampling, the target population is first divided into separate strata. Then, samples are selected within each stratum, either through simple or systematic sampling. The total number of individuals to be selected in each stratum can be fixed or proportional to the size of each stratum. Each individual may be equally likely to be selected to participate in the study. However, the fixed method usually involves the use of sampling weights in the statistical analysis (inverse of the probability of selection or 1/P). An example is the study conducted in South Australia to investigate factors associated with vitamin D deficiency in preschool children. Using the national census as the sample frame, households were randomly selected in each stratum and all children in the age group of interest identified in the selected houses were investigated.[11]

Cluster sampling: in this type of probabilistic sampling, groups such as health facilities, schools, etc., are sampled. In the above-mentioned study, the selection of households is an example of cluster sampling.[11]

Complex or multi-stage sampling: This probabilistic sampling method combines different strategies in the selection of the sample units. An example is the study of Duquia et al. to assess the prevalence and factors associated with the use of sunscreen in adults. The sampling process included two stages.[12] Using the 2000 Brazilian demographic census as sampling frame, all 404 census tracts from Pelotas (Southern Brazil) were listed in ascending order of family income. A sample of 120 tracts were systematically selected (first sampling stage units). In the second stage, 12 households in each of these census tract (second sampling stage units) were systematically drawn. All adult residents in these households were included in the study (third sampling stage units). All these stages have to be considered in the statistical analysis to provide correct estimates.

NONRESPONDENTS

Frequently, sample sizes are increased by 10% to compensate for potential nonresponses (refusals/losses).[1] Let us imagine that in a study to assess the prevalence of premalignant skin lesions there is a higher percentage of nonrespondents among men (10%) than among women (1%). If the highest percentage of nonresponse occurs because these men are not at home during the scheduled visits, and these participants are more likely to be exposed to the sun, the number of skin lesions will be underestimated. For this reason, it is strongly recommended to collect and describe some basic characteristics of nonrespondents (sex, age, etc.) so they can be compared to the respondents to evaluate whether the results may have been affected by this systematic error.

Often, in study protocols, refusal to participate or sign the informed consent is considered an "exclusion criteria". However, this is not correct, as these individuals are eligible for the study and need to be reported as "nonrespondents".

SAMPLING METHOD ACCORDING TO THE TYPE OF STUDY

In general, clinical trials aim to obtain a homogeneous sample which is not necessarily representative of any target population. Clinical trials often recruit those participants who are most likely to benefit from the intervention.[3] Thus, the more strict criteria for inclusion and exclusion of subjects in clinical trials often make it difficult to locate participants: after verification of the eligibility criteria, just one out of ten possible candidates will enter the study. Therefore, clinical trials usually show limitations to generalize the results to the entire population of patients with the disease, but only to those with similar characteristics to the sample included in the study. These peculiarities in clinical trials justify the necessity of conducting a multicenter and/or global studiesto accelerate the recruitment rate and to reach, in a shorter time, the number of patients required for the study.[13]

In turn, in observational studies to build a solid sampling plan is important because of the great heterogeneity usually observed in the target population. Therefore, this heterogeneity has to be also reflected in the sample. A cross-sectional population-based study aiming to assess disease estimates or identify risk factors often uses complex probabilistic sampling, because the sample representativeness is crucial. However, in a case-control study, we face the challenge of selecting two different samples for the same study. One sample is formed by the cases, which are identified based on the diagnosis of the disease of interest. The other consists of controls, which need to be representative of the population that originated the

cases. Improper selection of control individuals may introduce selection bias in the results. Thus, the concern with representativeness in this type of study is established based on the relationship between cases and controls (comparability).

In cohort studies, individuals are recruited based on the exposure (exposed and unexposed subjects), and they are followed over time to evaluate the occurrence of the outcome of interest. At baseline, the sample can be selected from a representative sample (population-based cohort studies) or a non-representative sample. However, in the successive follow-ups of the cohort member, study participants must be a representative sample of those included in the baseline.[14,15] In this type of study, losses over time may cause follow-up bias.

CONCLUSION

Researchers need to decide during the planning stage of the study if they will work with the entire target population or a sample. Working with a sample involves different steps, including sample size estimation, identification of the sample frame, and selection of the sampling method to be adopted.❑

REFERENCES

1. Martínez-Mesa J, González-Chica DA, Bastos JL, Bonamigo RR, Duquia RP. Sample size: how many participants do I need in my research? An Bras Dermatol. 2014;89:609-15.
2. Röhrig B, du Prel JB, Wachtlin D, Kwiecien R, Blettner M. Sample size calculation in clinical trials: part 13 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010;107:552-6.
3. Suresh K, Thomas SV, Suresh G. Design, data analysis and sampling techniques for clinical research. Ann Indian Acad Neurol. 2011;14:287-90.
4. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. Int J Epidemiol. 2013;42:1012-4.
5. Krause M, Lutz W, Boehnke JR. The role of sampling in clinical trial design. Psychother Res. 2011;21:243-51.
6. Roubille C, Richer V, Starnino T, McCourt C, McFarlane A, Fleming P, et al. Evidence-based Recommendations for the Management of Comorbidities in Rheumatoid Arthritis, Psoriasis, and Psoriatic Arthritis: Expert Opinion of the Canadian Dermatology-Rheumatology Comorbidity Initiative. J Rheumatol. 2015;42:1767-80.
7. Larkin J, Ascierto PA, Dréno B, Atkinson V, Liszkay G, Maio M, et al. Combined vemurafenib and cobimetinib in BRAF-mutated melanoma. N Engl J Med. 2014;371:1867-76.
8. Goncalves JR, Nappo SA. Factors that lead to the use of crack cocaine in combination with marijuana in Brazil: a qualitative study. BMC Public Health. 2015;15:706.
9. Pimenta FB, Pinho L, Silveira MF, Botelho AC. Factors associated with chronic diseases among the elderly receiving treatment under the Family Health Strategy. Cien Saude Colet. 2015;20:2489-98.
10. Kelbore AG, Alemu W, Shumye A, Getachew S. Magnitude and associated factors of Atopic dermatitis among children in Ayder referral hospital, Mekelle, Ethiopia. BMC Dermatol. 2015;15:15.
11. Zhou SJ, Skeaff M, Makrides M, Gibson R. Vitamin D status and its predictors among pre-school children in Adelaide. J Paediatr Child Health. 2015;51:614-9.
12. Duquia RP, Menezes AM, Almeida HL Jr, Reichert FF, Santos Ida S, Haack RL, et al. Prevalence of sun exposure and its associated factors in southern Brazil: a population-based study. An Bras Dermatol. 2013;88:554-61.
13. Barrios CH, Werutsky G, Martinez-Mesa J. The global conduct of cancer clinical trials: challenges and opportunities. Am Soc Clin Oncol Educ Book. 2015:e132-9.
14. Victora CG, Barros FC. Cohort profile: the 1982 Pelotas (Brazil) birth cohort study. Int J Epidemiol. 2006;35:237-42.
15. Boing AC, Peres KG, Boing AF, Hallal PC, Silva NN, Peres MA. EpiFloripa Health Survey: the methodological and operational aspects behind the scenes. Rev Bras Epidemiol. 2014;17:147-62.

Mailing address:
Jeovany Martínez-Mesa
Faculdade Meridional - IMED
Escola de Medicina
R. Senador Pinheiro, 304
99070-220 - Passo Fundo - RS
Brazil
Email: jeovanymm@gmail.com

How to cite this article: Martinez-Mesa J, González-Chica DA, Duquia RP, Bonamigo RR, Bastos JL. Sampling: how to select participants in my research study? An Bras Dermatol. 2016;91(3):326-30.