Journal of Clinical Epidemiology 82 (2017) 71–78

Journal of Clinical Epidemiology

Predicting data saturation in qualitative surveys with mathematical models from ecological research

Viet-Thi Tran[a,b,c,*], Raphael Porcher[b,c,d], Viet-Chi Tran[e,f], Philippe Ravaud[b,c,d,g]

[a]Department of General Medicine, Paris Diderot University, 16 Rue Henri Huchard, 75018 Paris, France
[b]Centre de recherche en Epidémiologie et Statistiques (CRESS), INSERM U1153, Place du Parvis Notre Dame, 75004 Paris, France
[c]Centre d'Épidémiologie Clinique, Hôpital Hôtel-Dieu, Assistance Publique-Hôpitaux de Paris, 1 Place du Parvis Notre Dame, 75004 Paris, France
[d]Paris Descartes University, 12 Rue de l'Ecole de Medecine, 75006 Paris, France
[e]Laboratoire Paul Painlevé—UMR CNRS 8524, Bâtiment M2, Cité Scientifique, 59655 Villeneuve-d'Ascq, France
[f]Université des Sciences et Technologies de Lille, Cité Scientifique, 59650 Villeneuve-d'Ascq, France
[g]Department of Epidemiology, Columbia University Mailman School of Public Health, 116th St & Broadway, New York, NY, USA

Accepted 3 October 2016; Published online 24 October 2016

Abstract

Objective: Sample size in surveys with open-ended questions relies on the principle of data saturation. Determining the point of data saturation is complex because researchers have information on only what they have found. The decision to stop data collection is solely dictated by the judgment and experience of researchers. In this article, we present how mathematical modeling may be used to describe and extrapolate the accumulation of themes during a study to help researchers determine the point of data saturation.

Study Design and Setting: The model considers a latent distribution of the probability of elicitation of all themes and infers the accumulation of themes as arising from a mixture of zero-truncated binomial distributions. We illustrate how the model could be used with data from a survey with open-ended questions on the burden of treatment involving 1,053 participants from 34 different countries and with various conditions. The performance of the model in predicting the number of themes to be found with the inclusion of new participants was investigated by Monte Carlo simulations. Then, we tested how the slope of the expected theme accumulation curve could be used as a stopping criterion for data collection in surveys with open-ended questions.

Results: By doubling the sample size after the inclusion of initial samples of 25 to 200 participants, the model reliably predicted the number of themes to be found. Mean estimation error ranged from 3% to 1% with simulated data and was $<2\%$ with data from the study of the burden of treatment. Sequentially calculating the slope of the expected theme accumulation curve for every five new participants included was a feasible approach to balance the benefits of including these new participants in the study. In our simulations, a stopping criterion based on a value of 0.05 for this slope allowed for identifying 97.5% of the themes while limiting the inclusion of participants eliciting nothing new in the study.

Conclusion: Mathematical models adapted from ecological research can accurately predict the point of data saturation in surveys with open-ended questions.

Keywords: Sample size; Qualitative research; Data saturation; Open-ended questions; Surveys and questionnaires; Web-based questionnaires

1. Context

Surveys with open-ended questions are a simple design to explore the different aspects of a concept in a given population [1]. This design is popular in many fields, including health research, social science, and marketing. For example, in health research, surveys may help identifying the topics that should be addressed in items of patient-reported outcomes [2]. The use of open-ended questions allows respondents to describe with nuance and detail how they perceive the concept under study. By reading and reflecting on participant answers, researchers can identify the meaningful variations and relationships of aspects of the concept, which allows for developing theories on how a particular phenomenon "works."

Surveys with open-ended questions are related to qualitative research because they seek to describe the qualities of

Conflict of interest: None.

Funding: This study was funded by the French Health Ministry (PHRC AOM13127). Our team is supported by an academic grant from the program "Equipe espoir de la Recherche," Fondation pour la Recherche Médicale, Paris, France (no. DEQ20101221475). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article.

* Corresponding author. 1 place du Parvis Notre-Dame, 75181 Paris, France. Tel.: +33-1-42-34-89-87; fax: +33-1-42-34-87-90.

E-mail address: thi.tran-viet@htd.aphp.fr (V.-T. Tran).

http://dx.doi.org/10.1016/j.jclinepi.2016.10.001
0895-4356/

What is new?

Key findings

- Sample size in surveys with open-ended questions relies on the principle of data saturation. We used a mathematical model used in ecological research to describe and extrapolate the accumulation of themes during a survey using open-ended questions.
- Our model accurately predicted the point of data saturation in surveys with open-ended questions. We showed that the slope of the expected theme accumulation curve could be used as a stopping criterion in surveys using open-ended questions.

What this adds to what was known?

- Up to this date, sample size in surveys with open-ended questions was determined solely by the judgment and experience of researchers.
- Determining the number of themes or even describing exhaustively all themes is not an objective per se of qualitative research. However, our mathematical model can help researchers to estimate the number of participants to include and avoid small incomplete studies or needlessly large studies.

What is the implication and what should change now?

- We suggest the following analysis plan for determining sample size in surveys using open-ended questions:
  - Step 1: Invite a sample of 50 to 100 participants to respond and analyze their answers by using the researcher's preferred method.
  - Step 2: Organize findings as a matrix opposing observed themes and units of analysis.
  - Step 3: Use the model presented to compute the theme accumulation curve and the local slope of the curve at the point of data analysis.
  - Step 4: If the local slope of the theme accumulation curve is above the chosen stopping criterion, continue data collection and analysis and go back to step 2.

entities and phenomena that are not measured with numbers [3]. However, they also present differences from classic inquiry methods such as interviews or focus groups by the use of (1) structured questionnaires instead of free conversation, (2) a large sample of participants determined according to predefined criteria instead of purposeful recruitment, and (3) linear data collection and analysis instead of circular and iterative processes.

Despite these differences, the purpose of surveys, similar to other qualitative inquiry methods, is the comprehensive and thorough description of the topic of interest. To that end, data collection and analysis continue to the point when additional input from new participants no longer changes the researchers' understanding of the concept. This is the point of data saturation [4,5]. Determining the number of participants to be included to obtain data saturation is one of the most frequent questions in qualitative research [6], but no transparent reproducible method has been developed to verify researchers' claims of having "seen nothing in newly sampled units or feeling comfortable that a theoretical category has been saturated" [7].

The problem for researchers in assessing the point of data saturation, that is, estimating the "true"—and unknown—number of themes about a given topic, is similar to that faced by ecological researchers when trying to estimate the number of species in an area [8]. In these studies, researchers cannot count or observe every possible animal and therefore use sampling methods to determine species richness. They use quadrats, lures, or traps in the study area; observe individuals captured; and enumerate the species found. From their empirical sample, researchers then use mathematical models to extrapolate the accumulation of species and determine the "true" species richness of an area. Such models have been used, for example, to determine the number of ant species in a tropical rain forest in Costa Rica [9].

The objective of this study was to present how mathematical modeling established in ecological research may be used to describe and extrapolate the accumulation of themes during a study using open-ended questions to help researchers to determine the point of data saturation. The specific aims are to (1) adapt the models established in ecological research to themes retrieved from health research to estimate the number of different themes that could be discovered in a survey with open-ended questions, (2) assess the reliability of this method by using both real data collected during a survey of the burden of treatment that used open-ended questions [10] and simulated data sets, (3) present a method to estimate the point in data collection when the inclusion of new participants is not likely to lead to the identification of new themes, and (4) propose an analysis plan for (health) research involving open-ended questions based on our results.

2. Methods

We used mathematical modeling to determine the point of data saturation in surveys using open-ended questions. It is important to note that the aim of our work was not to predict the themes, ideas, and meanings that patients may elicit on the topic of interest but rather to estimate how these new