The second step in data analysis: Coding qualitative research data

Heather L. Stuckey

Department of Medicine and Public Health Sciences, Pennsylvania State University College of Medicine, Hershey, USA

A B S T R A C T

Coding is a process used in the analysis of qualitative research, which takes time and creativity. Three steps will help facilitate this process:
1. Reading through the data and creating a storyline;
2. Categorizing the data into codes; and
3. Using memos for clarification and interpretation.

Remembering the research question or storyline, while coding will help keep the qualitative researcher focused on relevant codes. A data dictionary can be used to define the meaning of the codes and keep the process transparent. Coding is done using either predetermined (a priori) or emergent codes, and most often, a combination of the two. By using memos to help clarify how the researcher is constructing the codes and his/her interpretations, the analysis will be easier to write in the end and have more consistency. This paper describes the process of coding and writing memos in the analysis of qualitative data related to diabetes research.

Key words: Coding, memos, methods, qualitative research

INTRODUCTION

In earlier issues, we discussed data collection through three types of interviews, and the first step in the analysis was transcribing and managing qualitative research data. Coding, or the process of organizing and sorting qualitative data, is the second step in data analysis. Some research methodologists believe that coding is merely technical, preparatory work for higher level thinking about the study, but coding is analysis.[1] Codes are usually used to retrieve and categorize data that are similar in meaning so the researcher can quickly find and cluster the segments that relate to one another. Depending on the size of the dataset, coding can take hours, weeks, or even months.

One of the largest datasets for which I created a coding dictionary was the second Diabetes Attitudes, Wishes and Needs (DAWN2) study. In that dataset, there were over 15,000 participants who responded to an average of three open-ended questions each to a 17-country survey about psychosocial needs in diabetes. A sophisticated coding dictionary and process was needed for this project.

I sat cross-legged on the floor of my bedroom to code data according to different colors of highlighted markers, each representing a different category. Pink represented one code, and yellow was coded to another, until all the colors of the 12-pack highlighter set were used. I hand-drew lines onto the transcription to organize the codes, until the pages were filled with color. Papers were lined across the floor, which was amusing to my cat, Oreo, wondering when I would be finished with all of the work.

Although I now use sophisticated qualitative data analysis and research software, such as Nvivo 10 (QSR International) and Atlas.ti 7, the general process of reading and coding the data remains much the same. The first step is always reading and knowing your data before you start to code.
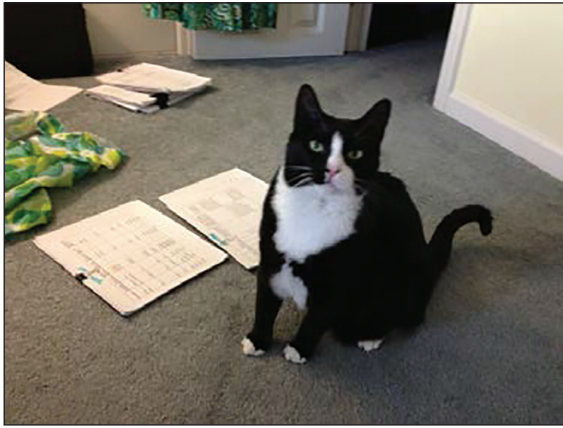
Access this article online	
Quick Response Code:	**Website:** www.joshd.net
	DOI: 10.4103/2321-0656.140875

Corresponding Author: Prof. Heather L. Stuckey, Department of Medicine and Public Health Sciences, Pennsylvania State University College of Medicine, Hershey, USA. E-mail: hstuckey@hmc.psu.edu

Reading the Data and Creating a Storyline

Before jumping into the process of coding data, it is important to think about your research question and the big picture, which some may refer to as "storyline" or "meta-narrative." One of the keys in coding your data, and in conducting a qualitative analysis more generally, is developing a storyline.[2] The story is directly related to your research question, such as "What are the data telling me that will help me understand more about the research question?" Your research question (the purpose of the study) is your guiding storyline. Keeping your purpose in mind will help you in later steps when you begin to develop themes in your data that link to this storyline. You need to become familiar with your data by reading through the transcripts at least 1 time (twice ideally). Then, you can start to think about the storyline that is told within the data.

One way to begin is to write down a sentence or short paragraph that summarizes what the data are saying in general terms. For example, if the research question is "what are the primary psychosocial challenges for people with diabetes (PWD) in India?," then you would read the data and write down your impressions or interpretations of the data. Your initial evaluation that contributes to the storyline might read something like this:

Although many PWD in India found little struggle in diabetes, some adults expressed psychosocial challenges in living with diabetes. Depression and feeling alone were two commonly stated challenges, which impacted their lives in terms feeling worried about future complications. There was a cultural impression that people wanted to keep diabetes hidden so that individuals would not discover they had diabetes, and to not complain about having diabetes. PWD did not complain of feeling stressed by having diabetes, or of being discriminated against, but rather a sense of feeling depressed and worried about the future.

Your coding scheme should be based on what you hope to convey to others. It is a good idea to start coding data with the purpose of your study, so your coding scheme is consistent. Developing a storyline will help you decide what concepts and themes you want to communicate in your analysis, and guides you in how your data could be organized and coded.[3] This does not mean, however, that every code needs to relate to the storyline, but the narrative will help to give focus to the purpose of the article.

Categorizing The Data Into Codes

One common myth that I frequently hear is that "the qualitative software program will code the data for me." This is not true. The software program helps to organize your data, but it does not code it. Software is a data management system, which is extremely helpful for large projects, or projects that require cross-analysis of variables such as demographics to specific codes (for example, running a report with "women" and "feeling worried about complications"). It is important that using software cannot be a substitute for learning data analysis methods because the researcher must know how to create codes, and analyze the data.[4] However, regardless of whether you choose to use data management software or code the data manually, you will follow the same process.

The process of creating codes can be predetermined — sometimes referred to as deductive or "a priori"[5] — or emergent,[6] or a combination of both. Predetermined coding may be based on a previous coding dictionary from another researcher or key concepts in a theoretical construct. They may derive from the interview guide or list of research questions. For instance, in the DAWN2 study, participants were asked multiple choice questions as well as open-ended questions. The open-ended questions were related to successes that helped in successful diabetes management (coded as "success"), wishes for improvement ("wishes") and needs that could improve self-management ("needs"). These began as the three a priori codes, because they were asked on the questionnaire. Other codes were emergent, which means that they were concepts, actions, or meanings, that evolved from the data and are different from the a priori codes. These included another main code called "advice" and all of the secondary codes that supported the main codes. Below is an excerpt of the codes from the DAWN2 study as entered in Nvivo 10:

As shown in Figure 1, the codes with a plus-sign to the left could be expanded, where you would see more emergent codes. The acronym PWD means "person with diabetes." There is also a code labeled "unsure of where to code" for codes that don't seem to fit or are too vague to code. The researcher can look at these codes after coding is completed to see if any further code can be extracted from these data. This helps to create a system to organize your data. Questions to ask yourself, while reading through the data are, "What does this mean? or "What does this exemplify?" The choice of writing a code manual (or data dictionary) is helpful if you are going to have a team of coders, or to help remember the meaning of the codes to assist in the interpretation. The use of a coding dictionary also provides a trail of rationale and evidence for the credibility of the study.[7] Figure 2 shows an example of codes for psychosocial challenges experienced by a PWD.

As you code your data, you will find other codes that may need to be created. Often, there are codes that need to be separated into two codes, for example, the initial code of "negative emotions" might need to be broken down to "depression" and "anxiety" as subcodes. Figure 2 shows all of the subcodes that describe "psychosocial challenges" and they are related to negative emotions. Now is the time to check with the storyline, or your research purpose, to see whether the codes you have created are in response to the purpose of the study.

USING MEMOS FOR CLARIFICATION AND INTERPRETATION

Coding also involves memos to write ideas or thoughts of how you arrived at the codes, and how are you using them to explain your storyline. These notes are often informal, kept for your insight and information only. One of the more practical uses of memos is to record how you are developing the codes and making decisions about coding.[8] This enhances the audit trail to demonstrate to the reader how decisions were made, and conclusions were reached. Memos can be written when you decide to combine or split codes, when you want to write conceptual notes about how the codes tell the storyline, or the context in which a certain code could be applied. For example, a memo about the DAWN2 is below:

I had a conference call with our French representative yesterday. He had a suggestion to think about the internal and external psychological processes that were being described. Perhaps consider this as a way of looking at depression and stress, versus lack of social support for PWD. I don't think I will change the codes to reflect this, but it will be good to remember when we look at thematic analysis.

Writing memos can help the researcher explore similarities and differences in the data, which is not done during

Figure 1: People with diabetes overall coding structure from second Diabetes Attitudes, Wishes and Needs

Guilt	Under this code, PWD state they feel guilt or a sense of blame about diabetes, such as "diabetes is my fault"
Hard and difficult	Having diabetes is hard and difficult, both mentally and/or physically
Lack of support	The PWD does not have psychological support to assist in diabetes management. Perhaps they feel lonely or distressed
Mentally hard to live with diabetes	PWD state that is hard on the mind or the soul to live with diabetes
Never accepted having diabetes	PWD talked about denial of diabetes, or not accepting the disease. This is a challenge for them in controlling their diabetes
People don't understand diabetes or PWD feels discrimination	The PWD thinks that there is a general lack of understanding about diabetes and its complications, or the PWD feels discrimination from having diabetes
Shame	PWD express shame due to their large body size, or because they have diabetes
Stress	Stress is challenge to the PWD. When coding, leave all the emotions of stress separate from one another; later, we can aggregate common emotions
Worry about complications	The PWD states that he/she is worried about complications or talks about the worry he/she has about the seriousness of diabetes

PWD: People with diabetes

Figure 2: An example of a coding manual (data dictionary) for code "psychosocial challenges"

the establishment of the codes, and write about initial relationships between certain codes, such as "type 1 diabetes" and "stressful diagnosis." Keeping detailed memos can help a researcher understanding why he/she made certain choices at the beginning of the study, and see how the thinking changes throughout the course of the data. Because the researcher is the primary instrument for analysis, it is important to be transparent about decisions made and reflective at the end of the study on how the process was mapped. These memos are kept separate from the original source, such as a transcript, to ensure the original comments are not confused with the researcher's interpretations of the data.

Qualitative research analysis is both a structured/linear and creative/iterative process and as any other practice, will improve and get easier in time. We are using the narratives of individuals with diabetes, their families, or healthcare providers in their own words to guide our coding structure. The process of coding breaks the data into parts so that the data are manageable, with the result of rebuilding the data to tell a storyline. This is related to the establishment of themes, which we will discuss in the next issue.

REFERENCES

1. Miles MB, Huberman AM, Saldaña J. Qualitative Data Analysis: A Methods Sourcebook. Thousand Oaks, CA: SAGE; 2014.
2. Physician Practice Connections — Patient Centered Medical Home Fact Sheet, 2010. Available from: http://www.ncqa.org/tabid/631/Default.aspx. [Last accessed on 2010 Jul 15].
3. Treiber J, Kipke R, Dizon C, Andrews J, Cassady D. Center for Program Evaluation and Research: Evaluation Tools Tip #18 Coding Qualitative Data 2013. Available from: http://www.tobaccoeval.ucdavis.edu/documents/Tips_Tools_18_2012.pdf. [Last accessed on 2014 Apr 13].
4. Denzin NK, Lincoln YS, editors. Collecting and Interpreting Qualitative Materials. Thousand Oaks, CA: SAGE; 2008.
5. Crabtree BF, Miller WL, editors. Doing Qualitative Research. 2nd ed. Thousand Oaks, CA: SAGE; 1999.
6. Boyatzis R. Transforming Qualitative Information: Thematic Analysis and Code Development. Thousand Oaks, CA: Sage Publications; 1998.
7. Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. Int J Qual Res 2006;5:10.
8. Birks M, Chapman Y, Francis K. Memoing in qualitative research: Probing data and processes. J Res Nurs 2008;13:9.