Organizing a breast cancer database: data management

Min Yi, Kelly K. Hunt

Department of Breast Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Contributions: (I) Conception and design: All authors; (II) Administrative support: KK Hunt; (III) Provision of study materials or patients: All authors; (IV) Collection and assembly of data: All authors; (V) Data analysis and interpretation: M Yi; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Min Yi, MD, PhD. Department of Breast Surgical Oncology, Unit 1434, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030, USA. Email: myi@mdanderson.org.

Abstract: Developing and organizing a breast cancer database can provide data and serve as valuable research tools for those interested in the etiology, diagnosis, and treatment of cancer. Depending on the research setting, the quality of the data can be a major issue. Assuring that the data collection process does not contribute inaccuracies can help to assure the overall quality of subsequent analyses. Data management is work that involves the planning, development, implementation, and administration of systems for the acquisition, storage, and retrieval of data while protecting it by implementing high security levels. A properly designed database provides you with access to up-to-date, accurate information. Database design is an important component of application design. If you take the time to design your databases properly, you'll be rewarded with a solid application foundation on which you can build the rest of your application.

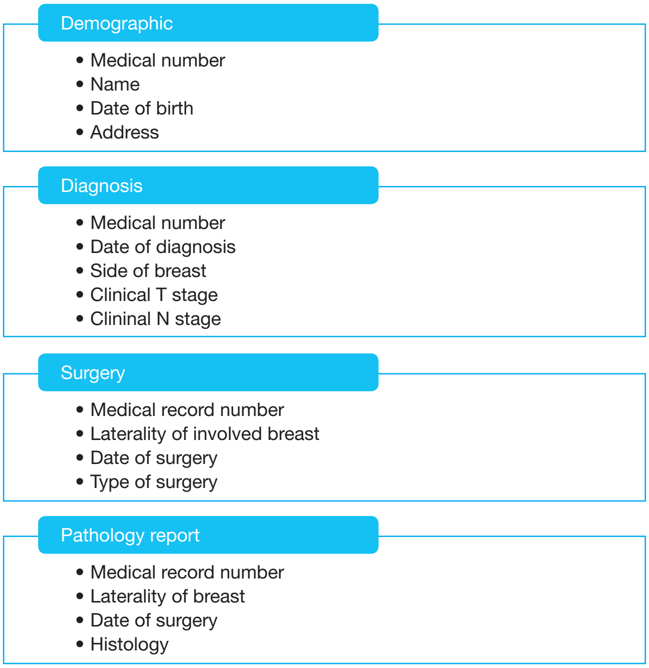
Keywords: Breast cancer; database; data management

Submitted Mar 09, 2016. Accepted for publication Apr 01, 2016.
doi: 10.21037/cco.2016.04.04
View this article at: http://dx.doi.org/10.21037/cco.2016.04.04

Introduction

Breast cancer is the most common cancer in women and the leading cause of cancer death among women in all ethnic groups worldwide (1). Reducing the world's breast cancer burden is a great and noble cause that involves many people, including physicians, researchers, epidemiologists, and others. All of these individuals appreciate and rely on breast cancer data. Physicians and researchers need breast cancer data to learn more about the causes of breast cancer and detect breast cancer earlier, thereby increasing the chance of finding a cure. Developing and organizing a breast cancer database can provide this type of data and serve as valuable research tools for those interested in the etiology, diagnosis, and treatment of cancer. Thousands of manuscripts on breast cancer clinical research are published every year. Our Breast Surgical Oncology Department at The University of Texas MD Anderson Cancer Center has a breast cancer research database with data on almost 30,000 patients. We have been able to provide many high quality publications to the breast cancer literature by using the data from this database. Some recent examples include, "Evaluation of the stage IB designation of the American Joint Committee on Cancer staging system in breast cancer" (*J Clin Oncol*) (2); "Evaluation of a breast cancer nomogram for predicting risk of ipsilateral breast tumor recurrences in patients with ductal carcinoma *in situ* after local excision" (*J Clin Oncol*) (3); "Novel staging system for predicting disease-specific survival in patients with breast cancer treated with surgery as the first intervention: time to modify the current American Joint Committee on Cancer staging system" (*J Clin Oncol*) (4); and "Classification of ipsilateral breast tumor recurrences after breast conservation therapy can predict patient prognosis and facilitate treatment planning" (*Ann Surg*) (5). High quality publications can guide clinicians for better decision making.

Depending on the research setting, the quality of the data can be a major issue. Assuring that the data collection process does not contribute inaccuracies can help to assure the overall quality of subsequent analyses. Data management is work that involves the planning, development, implementation, and administration of systems for the acquisition, storage,


Demographic

- Medical number
- Name
- Date of birth
- Address

Diagnosis

- Medical number
- Date of diagnosis
- Side of breast
- Clinical T stage
- Clininal N stage

Surgery

- Medical record number
- Laterality of involved breast
- Date of surgery
- Type of surgery

Pathology report

- Medical record number
- Laterality of breast
- Date of surgery
- Histology

Figure 1 The preliminary list of major tables in the breast cancer database.

and retrieval of data while protecting it by implementing high security levels.

Developing a database structure

The database structure is the manner in which investigators intend to store the data for the study so that it can be readily accessed for subsequent data analyses. Designing a database is in fact fairly easy, but there are a few rules to abide by. It is important to know what these rules are, but more importantly is to know why these rules exist, otherwise mistakes will be made. There are different types of breast cancer databases: by purpose (patient care: electronic medical record; research: breast cancer surgical research database) by software (Microsoft Access, Visual FoxPro, Oracle, Microsoft SQL Server and others). The design process consists of the following steps: determine the purpose of your database; find and organize the information required; divide the information into tables; turn information items into columns; specify primary keys; set up the table relationships; and standardize the data input (6-8).

Determining the purpose of your database

First determine the purpose of the database: define research questions and determine what measurements are needed to answer them. Talk to the individuals who will use the database. Brainstorm about the questions the database will be intended to answer. Sketch out the reports needed to produce the data. Produce the forms that will be used to record the data. Examine well-designed databases similar to the one that is being designed.

Finding and organizing the required information

To find and organize the information required, start with your existing information. For example, is the patient information in the hospital electronic medical record for breast cancer patients? If there is no such existing information, one will have to design a form to capture that information. What information should be included on the forms at each patient encounter? Identify and list each of these items. Ask the following data collection questions: Where are the data sources? Will it be retrieved from the patient medical record or collected by clinicians manually? How will the data collection forms be accessed by those individuals recording the data (clinic setting, operating, inpatient setting, pathology suite). Who will enter the data into the database (data entry personnel, physicians, nurses)? For example, a form intended to capture information from the first visit of a newly diagnosed breast cancer patient will contain the date of visit, demographic information, height and weight, symptoms at diagnosis, menstrual history and menopausal status, family history, radiographic studies obtained at diagnosis and breast cancer clinical staging information. Next, consider the types of reports researchers might want to produce from the database. For instance, clinicians planning a clinical trial might want to know the number of stage I patients diagnosed in the last month. Giving thought to the reports you might want to create helps you identify items you will need in your database.

Dividing the information into tables

To divide the information into tables, choose the major entities, or subjects. For example, after finding and organizing information for a breast cancer database, the preliminary list might look like *Figure 1*.

Turning information items into columns

To determine the columns in a table, decide what information you need to track about the subject recorded in

Table 1 Example of a data dictionary

Variable name	Variable description	Variable type	Variable width	Values/notes
MRN	Medical record number	Numeric	6.0	000001–900000
Treatment	Treatment group	Numeric	1.0	1=treated, 2=control
AGE	Age in months	Numeric	2.1	6.0–59.9
DOB	Date of birth	Date	–	MM/DD/YYYY
SEX	Sex	Numeric	1.0	1=male, 2=female
Height	Height (cm)	Numeric	3.1	–
Weight	Weight (kg)	Numeric	3.1	–
Proc_date	Procedure date	Date	–	MM/DD/YYYY
systolic_BP	Systolic blood pressure	Numeric	3.0	–

the table. For example, for the demographic table variables including medical record number, name, date of birth, address, race, height and weight and contact phone number comprise a good starting list of columns. Each record in the table contains the same set of columns, so you can store those information for each record. For example, the address column contains patients' addresses. Each record contains data about one patient, and the address field contains the address for that patient. Once you have determined the initial set of columns for each table, investigators can further refine the columns. A tip for determining the columns: don't include calculated data; store the information in its smallest logical parts.

Data dictionary

In every research project, investigators should generate a data dictionary (*Table 1*) that will contain a list of variables in the database as well as the assigned variable names and a description of each type of variable (e.g., character, numeric, dates). The data dictionary should also include the values accepted for each variable and any helpful comments such as important exclusions and skip patterns. The data dictionary is used primarily for data analysis and is an indispensable tool for the analysis team. Together with the database, it should provide comprehensive documentation that enables other researchers who might subsequently want to analyze the data to do so without any additional information.

When developing the database variable names, try to limit the name to 50 characters (shorter is better). Use a letter as the first character of the name (don't start names with numbers) and avoid using spaces in the names even if the system allows it.

Specifying primary keys

The data in tables requires keys (primary key) for identification of rows. Two requirements for primary keys are (I) every row must have a value in the primary key, empty fields are not allowed and (II) the primary key values can never be duplicated. Different tables will have different primary keys and some of them may need to use a combined key. For example, for each patient with breast cancer, there will be at most two unique records for diagnosis information (left and right) and the primary key for the diagnosis table will be a combination of medical record number and laterality of breast, however, for surgical procedures, there will be multiple records and the primary key will be the combination of the medical record number, laterality of breast and surgery date. There is only one unique record for patient demographic information and the primary key will be the patient medical record number only.

Creating the table relationships

After you have divided the information into tables, you will need a process to bring the information together again in meaningful ways. For example, clinicians may want to have a report about how many mastectomies were performed for clinical stage I breast cancer patients with invasive ductal carcinoma in 2012. This report includes information from several tables: diagnosis table (clinical stage); surgery table (date of surgery and surgery type) and pathology table (histology). There are three types of relationships between these tables: one-to-one, one-to-many and many-to-many. Consider this example with the diagnosis and surgery tables in the breast cancer database. A patient may have one diagnosis but can have multiple surgeries. It follows that for any diagnosis represented in the diagnosis table, there can be many surgeries represented in the surgery table. The relationship between the diagnosis table and surgery table is, therefore, a one-to-many relationship.

Standardize data input

Data standardization is a key part of ensuring data quality. Standardizing things like abbreviations and formatting during data entry can save many hours when analyzing the data in the future. When considering how to standardize the database, there will be many decisions to make. Should you use all caps, or normal capitalization? Should you store systolic and diastolic blood pressure in separate fields? There is no right or wrong answer to many of these questions. It depends on the priorities and how one intends to use the data. Put some thought into it, choose one way to enter data, and enter it that same way every single time. This is so important: be consistent. What follows are some tips. Use a separate column for each piece of information. Don't enter data such as "120/80" for blood pressure. Enter systolic blood pressure as one variable and diastolic blood pressure as another variable. Use the same unit throughout one column. The data value should not contain the unit. The unit can be noted in the header or in the data key. Don't use color to separate different patient groups since color is not readable by statistical software. Identify patient groups by adding a column.

Statistical consultant

The most effective way to work with a statistical consultant is to include them from the very beginning of the project. A statistician who is also knowledgeable in your area of research can be of great value in helping to define and focus the research effort into an efficient and successful project. The statistician can contribute relevant expertise in decisions about data management from the earliest stages. Decisions about how to code measures, and what to computerize, directly affect the ease, even the feasibility of subsequent analyses. When in doubt, ask the statisticians. Be sure the effort you are putting forth is necessary and will yield the results you are hoping to achieve.

In conclusion, a properly designed database provides you with access to up-to-date, accurate information. Database design is an important component of application design. If you take the time to design your databases properly, you'll be rewarded with a solid application foundation on which you can build the rest of your application.

Acknowledgements

None.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2012 Incidence and Mortality Webbased Report. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute, 2015. Available online: www.cdc.gov/uscs
2. Mittendorf EA, Ballman KV, McCall LM, et al. Evaluation of the stage IB designation of the American Joint Committee on Cancer staging system in breast cancer. J Clin Oncol 2015;33:1119-27.
3. Yi M, Meric-Bernstam F, Kuerer HM, et al. Evaluation of a breast cancer nomogram for predicting risk of ipsilateral breast tumor recurrences in patients with ductal carcinoma in situ after local excision. J Clin Oncol 2012;30:600-7.
4. Yi M, Mittendorf EA, Cormier JN, et al. Novel staging system for predicting disease-specific survival in patients with breast cancer treated with surgery as the first intervention: time to modify the current American Joint Committee on Cancer staging system. J Clin Oncol 2011;29:4654-61.
5. Yi M, Buchholz TA, Meric-Bernstam F, et al. Classification of ipsilateral breast tumor recurrences after breast conservation therapy can predict patient prognosis and facilitate treatment planning. Ann Surg 2011;253:572-9.

6. Churcher C. Beginning database design: from novice to professional. 2nd edition. New York, NY: Apress, 2012:9-24.
7. Stephens R. Beginning database design solutions. Indianapolis: Wiley Publishing Inc., 2009:63-224.
8. Database design basics. Retrieved 1 February, 2016. Available online: https://support.office.com/en-US/article/Database-design-basics-EB2159CF-1E30-401A-8084-BD4F9C9CA1F5

Cite this article as: Yi M, Hunt KK. Organizing a breast cancer database: data management. Chin Clin Oncol 2016;5(3):45. doi: 10.21037/cco.2016.04.04