Published in final edited form as:

Acad Radiol. 2015 April ; 22(4): 527–533. doi:10.1016/j.acra.2014.12.002.

How I do it: A Practical Database Management System to Assist Clinical Research Teams with Data Collecting, Organization, and Reporting

Howard Lee, B.S.¹, Julius Chapiro, M.D.¹, Rüdiger Schernthaner, M.D.¹, Rafael Duran, M.D.¹, Zhijun Wang, M.D., Ph.D.¹, Boris Gorodetski, B.S.¹, Jean-François Geschwind, M.D.¹, and MingDe Lin, Ph.D.²

Howard Lee: howard.lee1722@gmail.com; Julius Chapiro: j.chapiro@googlemail.com; Rüdiger Schernthaner: rschern1@jhmi.edu; Rafael Duran: rduran4@jhmi.edu; Zhijun Wang: wangzj301hospital@163.com; Boris Gorodetski: boris.gorodetski@charite.de; Jean-François Geschwind: jfg@jhmi.edu; MingDe Lin: ming.lin@philips.com

¹Russell H. Morgan Department of Radiology and Radiological Science, Division of Vascular and Interventional Radiology, The Johns Hopkins Hospital, Sheikh Zayed Tower, Ste 7203, 1800 Orleans St, Baltimore, MD, USA 21287

²U/S Imaging and Interventions (UII), Philips Research North America, 345 Scarborough Road, Briarcliff Manor, New York 10510

Introduction

With the growing amount of clinical research studies in the field of interventional oncology, selective patient data is becoming more difficult to store and organize effectively. Existing hospital EMR (electronic medical record) systems store patient data in the form of reports and data tables. Our institution's EMR system placed our researchers in a position where time consuming methods are needed to search for suitable patients for clinical studies. Researchers had to manually read through the reports and data tables to filter patients and gather data. For most studies, spreadsheet programs such as Microsoft Excel[®] (Microsoft, Washington, USA) are often used as a data repository similar to a database to record and organize patient data for research. Once the spreadsheet is populated, it is manually filtered by set study parameters and then pushed to statistical analysis software for further analysis. For statistical analysis, columns containing text are translated into binary values (1 or 0) to be in a format acceptable by statistical analysis software. For example, each tumor entity is assigned a new column. Patient histological reports are read manually to assign a 1 or 0 to each tumor entity column, 1 for positive, 0 for negative. Under a tumor entity column, researchers would write a 1 for all patients with the tumor and a 0 for all patients without the tumor.

© 2014 AUR. All rights reserved.

Correspondence to: Jean-François Geschwind, jfg@jhmi.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

This method of data storage has limitations in the organization and the quality of the data. Data input and analysis without a database run a higher risk of incorrect data entry, patient exclusion, and a higher risk of introducing duplicates. Furthermore, data selection and calculation is time consuming. An alternative could be the clinical research database that Meineke et. al. proposed (1). However, it is too unspecific for interventional oncology research and would need additional optimization, for example, the capability to automatically calculate various variables such as tumor staging systems and to record information about multiple treatment sessions.

The purpose of this study was to provide an improved workflow efficient tool through the use of a clinical research database management system (DBMS) optimized for interventional oncology clinical research.

Materials and Methods

This was a single-institution prospective study. The study was compliant with the Health Insurance Portability and Accountability Act (HIPAA) and was waived by the Institutional Review Board.

Database and Query Interface Design

The presented database management system has two distinct parts, the database server and client interface, illustrated in Figure 1. The database is run by software (MySQL, Oracle Corporation, California, USA and phpMyAdmin, The phpMyAdmin Project, California, USA) on a central computer server within the department (2, 3). Authorized users were granted access to this password protected and encrypted secured server (HIPAA compliant). Multiple users concurrently add, edit, and query data remotely through a customized graphical user interface (GUI) utilizing Microsoft Access® (Microsoft, Washington, USA). Any data changes are immediately logged for others to see. The database performed automatic calculations using queries, user-defined search criteria. Queries were saved, rerun, and exported to spreadsheets. Queries aid in data analysis and increase study productivity (4). They are powerful tools for filtering and sorting datasets. Figure 2 illustrates the query interface and an example of request from the database.

Graphical User Interface Design and Utility

In our research environment, the database GUI was created to facilitate patient data input. This was done by using custom user-friendly interface forms that contain textboxes and labels including demographic data, treatment information (e.g. conventional transarterial chemoembolization (TACE)), tumors types, dates and types of radiological exams, etc. The GUI is used to view patient data and allows users to add/edit data (Figure 3). The database interface is not limited to one form. It can have multiple forms, shown as tabs, to assist grouping various medical data. Figure 4 shows an example of multiple tabs for groups of related data.

Automatic Calculations

Automatic calculations may be run between values, such as dates. For example, the database may calculate the time between baseline imaging, follow-up imaging, treatment dates, pre- and post-treatment dates, date of diagnosis, and patient's date of death in relation to a particular treatment or event (e.g. randomization), essential for survival studies. Using these queries, the database can also calculate the median overall survival automatically. The database does also automatically calculate clinical scores such as Child-Pugh score and Barcelona Clinic Liver Cancer (BCLC) stage as shown in Figure 5 (5). For our purposes, the Child Pugh score and BCLC were calculated using baseline data before a patient's first embolization as is typically done for staging. The illustrated calculators can be revised as needed. Once patient blood data is available, queries are run to produce a list of all patients with Child-Pugh scores. Researchers can then quickly retrieve them.

Statistical Output

Another powerful feature of the database is its ability to provide a first tier of statistical information. Using this GUI, the user defines the search criteria and runs queries to obtain immediate statistical information about a particular set of parameters. With this feature, the database can quickly output an accurate summary of patient data such as, for example, how many patients have colorectal carcinoma and undergo conventional TACE.

Questionnaire Assessment

A questionnaire (15 questions) was designed and distributed to 21 board-certified interventional radiologists who conduct clinical research at our academic hospital that include Phase I, II, and III clinical trials, and retrospective studies. The questionnaire determined how data is controlled in retrospective studies and the likelihood to use the database. The questionnaire is shown in Table 1. The purpose of the questionnaire was to 1) illustrate the general scope of where researchers were having problems within Excel and data organization, such as wasted effort working with duplicate patients and unintentional failure to include available patients, and 2) to gauge how receptive they would be to a database system. Using this information, the database system was constructed. There were weekly progress updates with the clinical research team to ensure that the original goals set out to address the deficiencies of Excel were being resolved.

Results

Questionnaire Results

All 21 interventional radiologists completed the questionnaire. Self evaluation results are shown in Figure 6. In data collection and analysis, over 50% (11/21) spent most of the time searching, filtering, and/or categorizing data. However, about 50% (10/21) spent little to no time calculating the data. 67% of respondents (14/21) realized at some point that there were erroneously included patients who should have been excluded and there were patients who were erroneously not included. Over 85% (18/21) were very receptive to using software that produces group summaries such as totals of each tumor type with minimal effort, calculates

clinical staging and score systems automatically, and also allows remote access for multiple users to add/edit data in a central server with data modification logs.

Query Interface Output

In Figure 7, the query of male patients, over 40 years old, with HCC is run. Figure 8 shows a query result of patients with TACE and Child-Pugh score A calculated by the database. Figure 9 illustrates an interval of time between two events as a query that can be calculated automatically (e.g. time elapsed between two embolization procedures). The output of the queries as described above is shown in a structured and concise list, which can be exported for further research study specific analysis.

Discussion

The main finding of this study is that there is a need for a much more time efficient and accurate way to store, retrieve, and analyze patient data for clinical research studies. The database management system presented here fulfills these needs. This was achieved through the use of automatic calculations, interface forms, queries, etc. With a personalized interface, data access, entry, organization, queries, calculations, and export processes are seamlessly performed to assist clinical research with data and statistical analysis. Furthermore, the database is a unified repository of clinical research information and a shared resource among the clinical research team. This allows for a multi-user level experience where there can be simultaneous access to the data and where the efforts of each individual in adding/appending new information can be used by the entire team.

With the presented database put into use, the effort for clinical studies can truly focus on conducting various statistical analysis and data interpretation rather than preparing data for analysis (6). All retrospective data can be merged into this database, enabling a centrally maintained and shared resource. Our clinical research team now has access to a customized database of patients with a large number of clinical parameters, allowing a vast combination of queries to form or support study hypotheses. The user defined GUI-connected interface is invaluable for anyone collecting data as it facilitates data entry and minimizes data entry errors.

In previous data collection and analysis, converting spreadsheet data to binary/numeric format was time consuming and impractical. The database presented in this study relieves the inconvenience of manually searching, organizing, and calculating data. Processing calculations, especially more complex calculations such as clinical staging scores, can now be done automatically. Prior to implementing the presented database system, a typical Excel spreadsheet for the clinical studies at our institution would have over 100 columns. These columns included patient demographics, repeat treatment dates and types (new columns per TACE session), and repeated pre-/post-imaging dates and types (new columns per multi-modular scan). Tracking medical data is frequently difficult due to the large amount of columns in the spreadsheet. Compared to a typical Excel spreadsheet with many columns, browsing and adding prospective data through the database interface presented here is more organized and practical with ten defined tabs for data groups, ranging from a patient's basic information to treatments to survival status. In addition, the database interface lists all

repeating treatments and imaging per patient as rows instead of columns, facilitating comparisons between multiple treatments of a patient. Combining the database's ability to calculate statistical analysis with automatic calculation queries, reports can be generated with virtually any parameter. This is not only helpful in radiology, but also beneficial for other studies and hospital information systems.

The database management system in this study has some limitations. A database system may not be suitable for all kinds of research teams. There are several factors that may illustrate the need of a database. In a previous report on data collection, applicable examples and guidelines were addressed to determine whether or not implementing a database is feasible in the current environment (7). Depending on the environment and context, a database may not be implemented right away as it needs additional testing. Furthermore, the database will need a dedicated server to host the database along with the data. In order to use the database interface, training is required. Someone who specializes in databases, such as a database administrator, needs to teach researchers and other potential users how to use the database interface and query interface for filtering patients and obtaining statistics. This is especially needed in more advanced queries and in developing additional GUIs. It should be noted that Microsoft Access is being used in this work as a "front-end" interface that communicates with the SQL database to query (filter) data, and for input/appending to existing data. Other software such as FileMaker Pro (FileMaker, Santa Clara, CA) and REDCap would serve a similar function (8). The need for the SQL database is so that multiple users can access the stored data at the same time, increased level of security, stability, and performance, and serving as a unified repository of clinical research information that can be shared by the research team (9, 10). Also, the database administrator has to not only construct a database on a server with input from clinicians and other end users, but in addition would need to maintain the database (11, 12). Typical maintenance includes routine backups, altering database structure and interface for new data types, and updating database and client software. A server can be hosted on a PC or online, both of which all parties involved can access in the same network locally or remotely. Furthermore, databases can be enabled to communicate with other databases. While the initial setup and learning curve is high, the database allows for fluid data entry in an organized fashion, querying results including calculations, and storing data while supporting simultaneous user access. With the variety of research teams and departments, ideally each suitable team should have their own database. This is not necessarily only for interventional oncology but also for any specific area of research, for example, studies with patients undergoing ablation, percutaneous abscess drainage (PAD), etc. These databases can be connected for interdisciplinary research to provide a broader scope of data and facilitate data search (13).

Conclusion

The current database implementation and interface allows a much faster and more detailed retrospective analysis of patient cohorts. In addition, it facilitates data management and a standardized information output for ongoing prospective clinical trials. The database management system with an interface is a work efficient and robust tool that provides a significant edge over manual retrieval of patient records by filtering data and assisting statistical analysis in a study-relevant fashion.

Acknowledgments

Funding and support has been provided by NIH/NCI R01 CA160771, P30 CA006973, and Philips Research North America, Briarcliff Manor, NY, USA.

References

1. Meineke FA, Stauber S, Lobe M, Winter A. A comprehensive clinical research database based on CDISC ODM and i2b2. *Studies in health technology and informatics*. 2014; 205:1115–9. [PubMed: 25160362]
2. Stobart, S.; Vassileiou, M. *MySQL Database and PHPMyAdmin Installation PHP and MySQL Manual*. Springer; London: 2004. p. 461-73.
3. Kuenz, D. *Book Manage data for free with MySQL*. City: Element K Journals; 2001. Manage data for free with MySQL; p. 7-10.
4. Coronel, CMS.; Rob, P. *Database systems: design, implementation, and management*. 9. Boston, Massachusetts: Cengage Learning; 2009.
5. Llovet JM, Di Bisceglie AM, Bruix J, et al. Design and endpoints of clinical trials in hepatocellular carcinoma. *Journal of the National Cancer Institute*. 2008; 100(10):698–711. [PubMed: 18477802]
6. Kanas G, Morimoto L, Mowat F, O'Malley C, Fryzek J, Nordyke R. Use of electronic medical records in oncology outcomes research. *ClinicoEconomics and outcomes research : CEOR*. 2010; 2:1–14. [PubMed: 21935310]
7. Schmier JK, Kane DW, Halpern MT. Practical applications of usability theory to electronic data collection for clinical trials. *Contemporary clinical trials*. 2005; 26(3):376–85. [PubMed: 15911471]
8. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*. 42(2):377–81. [PubMed: 18929686]
9. *MySQL Database Provides Full Transactional Support*. *Worldwide Databases*. 2002; 14(11) 0-N/A.
10. *Oracle Improves Database Performance with Latest Development Milestone Release for MySQL 5.7; New Release of the World's Most Popular Open Source Database is 2x Faster than MySQL 5.6 and Over 3x Faster than MySQL 5.5 in Benchmark Tests*. Book *Oracle Improves Database Performance with Latest Development Milestone Release for MySQL 5.7; New Release of the World's Most Popular Open Source Database is 2x Faster than MySQL 5.6 and Over 3x Faster than MySQL 5.5 in Benchmark Tests*. City2014.
11. Xie SX, Baek Y, Grossman M, et al. Building an integrated neurodegenerative disease database at an academic health center. *Alzheimer's & dementia : the journal of the Alzheimer's Association*. 2011; 7(4):e84–93.
12. Parkes, D.; Lowman, M.; Andres, C., et al., editors. *Pro Python System Administration*: Apress. 2010. Automatic MySQL Database Performance Tuning; p. 329-48.
13. Piriyaongsa J, Bootchai C, Ngamphiw C, Tongsima S. microPIR: an integrated database of microRNA target sites within human promoter sequences. *PloS one*. 2012; 7(3):e33888. [PubMed: 22439011]

Figure 1. The Dataflow Chart

This chart shows a general layout of the database server and its clients. It illustrates how the database management system performs queries (orange circle) such as statistical analysis. Multiple computers are granted access to the database. The blue rectangles represent the database management system software. Researchers can utilize the database client graphical user interface (GUI) to import data without needing to format. Researchers also control data through the GUI. Queries are usually run through the GUI to provide wanted results. Once

the results are obtained, researchers export the query to a spreadsheet, illustrated by the green rectangle.

Field:	MRN:	Last Name:	First Name:	Age:	Gender:	Type:
Table:	PATIENT_DATA	PATIENT_DATA	PATIENT_DATA	PATIENT_DATA	PATIENT_DATA	PATIENT_DATA
Sort:						
Show:	☒	☒	☒	☒	☒	☒
Criteria:				>40	"m"	"HCC"
or:						

Figure 2.

This figure illustrates the query interface. In this example query, a list of male patients over the age of 40 with hepatocellular carcinoma (HCC) is wanted. The user inputs search criteria for age, gender, and tumor type, ">40", "m", and "HCC" respectively. MRN: medical record number.

a Patient Form

PID: 4000
 EPR MRN: 123456
 Last Name: Test
 First Name: Test
 DOB: 1/1/1960
 Ethnicity: Other
 Gender: m
 Comments:
 LAST MODIFIED: 3/31/2014 3:20:52 PM
 MODIFIED BY:
 Save

b Tumor:

Tumor ID: 3377
 PID: 4000
 Type: HCC
 Comments:
 LAST MODIFIED: 3/31/2014 3:20:57 PM
 MODIFIED BY:
 Save

c Embolization Procedures

EMBO ID: 3524
 PID: 4000
 Type: DEB TACE
 Comments:
 Pre Scan Date: 12/27/2012
 Pre Scan: MRI
 EMBO Date: 1/1/2013
 Post Scan Date: 1/2/2013
 Post Type: MRI
 LAST MODIFIED: 3/31/2014 3:21:18 PM
 MODIFIED BY:
 Save

a1 Record: 1 of 1
 Record: 1591 of 1591
 No Filter Search

b1 Record: 1 of 1
 No Filter Search

c1 Record: 1 of 1
 No Filter Search

Figure 3.

This form illustrates how users input data to the database. The form is divided into three parts:

- (a) Patient Form – Data consists of basic patient information. Patient Identification (PID) is a unique number generated by the database to uniquely identify patients. LAST MODIFIED is a timestamp of when the data was most recently updated or added. MODIFIED BY is a text box that records who updated/added data. (a1) shows the total amount of patients in the database.
- (b) Tumor – Data consists of a patient’s primary and secondary tumors in the liver. The dropdown allows users to select a tumor or add new tumor types (e.g. metastatic disease). (b1) shows how many tumors types the patient has in the liver.
- (c) Embolization Procedures – Data consists of intra-arterial therapies (IATs) sessions. (c1) shows how many IATs sessions a patient has went through.

Patient ID	P_ID	Comments	
MRN	P_MRN	P_COMMENTS	
First Name	P_FNAME		
Last Name	P_LNAME		

Basic Information	Diagnostic	CBCT	TACE	MRI	MRI Post Target Lesion	Target Lesion	Non Targeted Lesion	Curative Treatment	Survival
-------------------	------------	------	------	-----	------------------------	---------------	---------------------	--------------------	----------

Target Lesion 1

RECIST Large (cm):		
RECIST Small (cm):		qEASL %: <input type="text"/>
EASL %:		qEASL Vol: <input type="text"/>
mRECIST (cm):		qDWI cc: <input type="text"/>
vRECIST (cm3):		qADC: <input type="text"/>

Figure 4.

This figure illustrates the tabular form where each group of related data is shown as individual tabs to assist user navigation. The display of patient identification information and comments are maintained while the user navigates to different tabs to preserve the scope and field of view for each patient.

Classification

Child-Pugh **IF N/A LEAVE BLANK**

Encephalopathy: Absent

Ascites: Severe

Albumin: 1.9

Bilirubin (mg/dL): 3.9

PT/INR: 1.9

Total Score: 12

Classification: C

Calculate

BCLC (2012)

Child Pugh: C

PS #: ☒ 0 ☐ 1-2 ☐ >2

Okuda #: ☒ 1-2 ☐ 3

Portal Invasion? ☒

Metastases? ☒

Lymph node involved? ☒

Nodules: ☐ Single <2cm
☐ Single or 3 nodules <3cm
☒ Multinodular

Stage: D Calculate Clear

Figure 5.

This form shows a patient's Child-Pugh score and Barcelona Clinic Liver Cancer (BCLC) stage. They are automatically calculated when provided with pertinent patient data. The "Calculate" buttons are used to refresh the form should any patient data value change. PT/INR: Prothrombin Time/International Normalized Ratio; PS: Performance Status.

Figure 6.
The self evaluation results are from Table 1.

MRN	Last Name	First Name	Gender	Type
			m	HCC
			m	HCC
			m	HCC
			m	HCC
			m	HCC

Figure 7.

This figure illustrates the output of a query for male patients with hepatocellular carcinoma (HCC). The interface outputs a list of all patients matching the search criteria.

✔ Showing rows 0 - 5 (6 total, Query took 0.0019 sec)				
P_ID	P_LNAME	P_FNAME	P_PROC_DATE	Classification
			2006-02-27	A
			2006-05-02	A
			2006-12-06	A
			2006-10-16	A
			2006-10-04	A
			2006-04-06	A

Figure 8.

This is the output of a query for patients who had undergone TACE in 2006 (P_PROC_DATE column) with Child Pugh Class A, here labeled as “Classification”. The automatically calculated Child-Pugh Class can be used also for querying.

The screenshot shows a database interface with a sidebar on the left listing various tables and queries. The 'TimeBetweenEMBOs' query is circled in red. The main window displays a table with the following data:

PATIENT	EMBODate	Next_EMBO	Elapsed_Days
	7/21/2005	1/16/2007	544
	1/16/2007		
	1/11/2005	3/11/2005	59
	3/11/2005		
	7/3/2007	7/17/2007	14
	7/17/2007	8/14/2007	28
	8/14/2007	10/1/2007	48
	10/1/2007		
	9/21/2011	11/14/2011	54

Figure 9.

The database automatically calculates the days between TACE sessions for each patient as a query (red circle). The current treatment “EMBODate,” is subtracted from the next treatment, “Next_EMBO.” Empty fields indicate that the patient has undergone only one treatment or the session is the latest treatment. Because the query is saved, double clicking the query indicated by the red circle refreshes the calculation for the entire database of patients.

Table 1**Questionnaire Assessment**

Response: Yes No
Question: I searched and filtered data manually <i>Example: Sorting and copying relevant data</i>
Question: I inputted formulas and Excel functions to calculate scores, response rates, or statistics in my Excel spreadsheet
Question: I summarized my Excel data in a report <i>Example: Total number of Child Pugh A patients</i>
Question: I converted non-binary data (<i>volume measurements, numeric values, occurrence rates of symptoms</i>) into binary data (0/1) by defining a cut-off point to differentiate <i>Example: Between responder and non-responder to a given therapy for statistical analysis</i>
Question: I have done statistical analysis myself
Question: I unknowingly produced duplicate data that I later found out was already collected by another colleague
Response: 0–20% 21–40% 41–60% 61–80% 81–100%
Question: From the beginning of data collection to finishing analysis, about what percentage of the total time spent for a single retrospective study did you spend on:
Question: Querying/filtering/categorizing data? <i>Example: Defining subsets of patients with certain criteria such as patients treated only with cTACE or only with DEB-TACE</i>
Question: Calculating data? <i>Example: Min, Max, Mean, Sum, Clinical Scores such as Child-Pugh</i>
Response: Very Unlikely Unlikely Neutral Likely Very Likely
Question: If given the opportunity, how likely will you use software that:
Question: Produces group summaries with minimal effort? <i>Example: Total number of Child Pugh A patients</i>
Question: Calculates clinical staging and score systems automatically?
Question: Allows multiple users to add and edit data into the same database so that redundant collection of the same patients by different colleagues can be avoided?
Question: Allows users to track data modifications?
Question: Stores data in a centralized location with remote access?