MASSACHUSETTS GENERAL HOSPITAL, DEPARTMENT OF SURGICAL ONCOLOGY
CENTER FOR QUANTITATIVE MEDICINE

2008 Partners Healthcare Breast Cancer Database

Features

Matthew Nolan
Research Assistant

James S. Michaelson, PhD.
Principal Investigator

5/12/2008

This document details the procurement and construction of the Partners Breast Cancer Database during 2007-2008. In addition to the information on the creation of the database, this reports also outlines key facts and figures about the data, itself.

Source of this document: Partners_BrCaDB_Features.docx

Table of Contents

Introduction

The Partners Healthcare Retrospective Breast Cancer Database is comprised of over 24,771 women and men diagnosed with one or more breast carcinomas. These patients were diagnosed between 1968 and 2007 at either the Massachusetts General Hospital or the Brigham and Women's Hospital, and follow-up time for some patients exceeds 30 years.

Partners Tumor Registry Data – Date last contact up to 10/2007, Date initial diagnosis up to 6/2007

RPDR Notes/Full-text reports data – Up to 3/2008

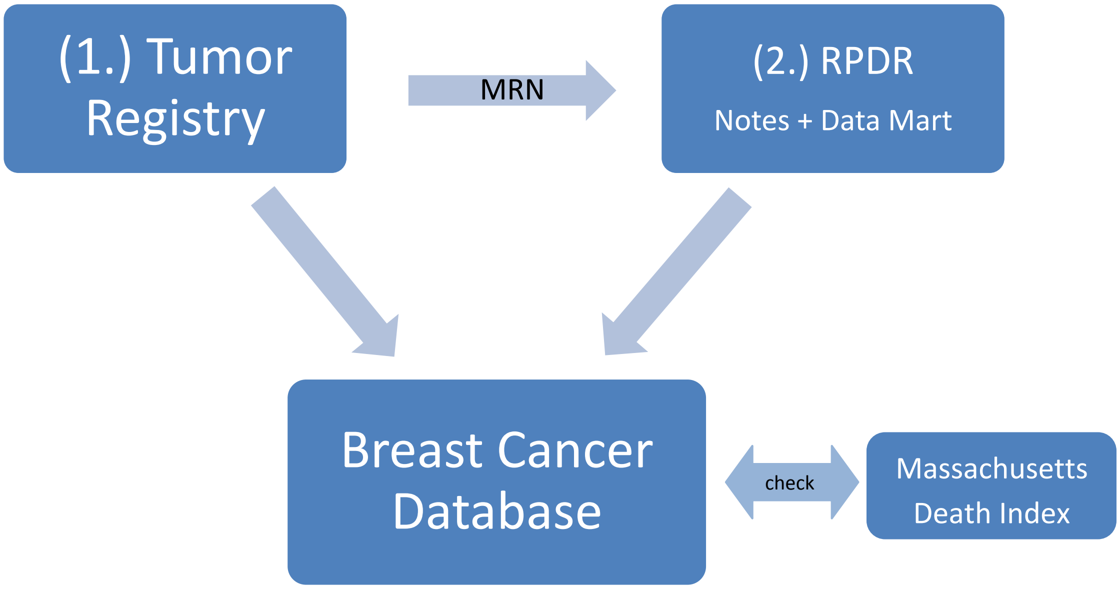
RPDR Data Mart – Up to 11/2007

Sources of Information

The information for the database comes principally from 2 sources: the Partners Healthcare Tumor Registry (managed by Carol Venuti), and the Partners Research Patient Data Registry (RPDR). The Tumor Registry is maintained by the tumor registrars, who create a record for every patient seen within the Partners Healthcare system who had been diagnosed with a neoplasm. The registrar adheres to very specific guidelines concerning the classification of tumors, and sorts through detailed information in the pathological and clinical reports to record essential markers of the disease like tumor size, number of lymph nodes examined/positive, hormone receptor status, grade, stage, etc.

The data from the Tumor Registry is very powerful because it gives the succinct information of the disease. However, this brevity necessitates that the registrar exclude some information about the neoplasms. In order to create the most comprehensive breast database possible, the Michaelson group uses the Medical Record Numbers for each patient listed in the Tumor Registry data to request additional patient data from the RPDR (pathological reports, clinical reports, operative notes, radiology reports, etc.). In addition to the RPDR reports/notes data, we also have a private RPDR Data Mart that gives even more information on the breast cancer patients from our original Tumor Registry data request. The data mart hosts every lab ever ordered for a patient, every medicine administered, and many other pieces of discreet data (*not* full-text reports).

Finally, we also have data from the Massachusetts Death Index from the MA Registry of Vital Records. We use this information to corroborate vital status for the breast cancer patients in our database. If any source identifies a patient as deceased, we conclude that she or he is deceased.



First, one obtains the disease-specific data from the Tumor Registry. Then, using the Tumor Registry data, one submits the MRNs for each patient to the RPDR to get the full-text reports. Death information is corroborated with the Massachusetts Death Index.

We received the data for our breast cancer database in several forms, among which were MS Access files, MS Excel files, flat txt files, and SQL databases. We cleaned the data, and compiled it into a locally-hosted MS SQL 2005 database. Details about how the data were cleaned and configured can be found in Appendix A, “Constructing the Database.”

Structure of the Database

All tables are located in a single MS SQL database, **Partners_BrCaDB**, hosted on the machine JSMTOWER1. The tables are named such that the first part of the string identifies the source, and the latter part identifies the information within the table (e.g. “TumReg_Patient” or “RPDR_Pathology” or “MADC_05” [Massachusetts Death Certificate data] or “DM_Patient_Dimension” [RPDR Data Mart]).

Our core list of patients comes from the Tumor Registry, and the RPDR data merely contain supplemental information about the Tumor Registry patients; it does not contain additional patients diagnosed with breast cancer. Therefore, we use the **TumReg_Patient** table as our main table when making patient-centric queries, from which most other tables are radially related. The Tumor Registry contains data on 24,763 patients, but only 23,015 of these patients have corresponding RPDR data. While EMPI (Enterprise Master Patient Index – a unique numeric value for unique patients, irrespective of facility) is the ideal key for patients, we only have EMPI numbers for those cases listed in *both* the RPDR and the Tumor Registry data. Thus 1,748 patients in the Tumor Registry data are without an EMPI. Therefore, when joining between tables in the Tumor Registry data, we use the field “Composite_ID” (which equals EMPI if one exists, and a concatenation of Last_name+First_name+DOB if not). When querying between the Tumor Registry and RPDR, we use “EMPI.” Finally, “Patient_Num” is the unique identifier for the patients in the Data Mart. It is simply a de-identified version of “EMPI” with identical functionality: each “Patient_Num” is mapped to a single “EMPI.”

To perform complex queries, one can use the MS SQL Server Management Studio (there is a query-design wizard much like the Access query designer for novice users). We have also developed an MS Access 2007 front-end for viewing the records. The form displays the information about a patient, along with all corresponding diagnoses, and full-text reports about that patient.

Tables and Views with their Primary Keys and Indexes

TumReg_Patient – Composite_ID (PK); the main patient table
TumReg_Diagnostic – MRN_SeqNum[1] (PK); the main table for breast cancer diagnoses
TumReg_Treatment – Composite_ID (index); describes treatments administered
TumReg_FollowUp – MRN_SeqNum + Flw_Date (PK)
TumReg_CollStage – MRN_SeqNum (PK)

RPDR_LMRNote – EMPI (IX); full-text clinical notes from the Longitudinal Medical Records
RPDR_Radiology – EMPI (IX); full-text radiology
RPDR_RadiologyImages – EMPI (IX)
RPDR_OperativeReports – EMPI (IX)
RPDR_Pathology – EMPI (IX)
RPDR_DischargeSummaries – EMPI (IX)
RPDR_Cardiology – EMPI (IX)
RPDR_Endoscopy – EMPI (IX)
RPDR_Pulmonary – EMPI (IX)
RPDR_ContactInformation – EMPI (IX)
RPDR_MRN – EMPI (IX)

MRN_Mapping – Mrn + Company_Cd (PK)

MADC_05_data – Last_Name + First_Name + DOB (IX)
MADC_70to04_data – Last_Name + First_Name + DOB (IX)
MADC_ICD10
MADC_ICD9
MADC_ICD8
MADC_master – Last_Name + First_Name + DOB (IX)

DM_Concept_Dimension (had some trouble creating PKs on the data mart tables because they are huge)
DM_Demographics
DM_Diagnosis
DM_DRG
DM_Encounter_Mapping

[1] The “MRN_SeqNum” is the unique identifier for cancer diagnoses. It is a concatenation of the patient’s MRN and Sequence Number. The Sequence Number identifies the order in which a diagnosis was made for an MRN.

DM_Encounters
DM_HealthHistory
DM_HPCGG
DM_LabTests
DM_Medications
DM_MicroBiology
DM_Observation_Fact
DM_Patient_Dimension
DM_Patient_Mapping
DM_Procedures
DM_Processing
DM_Provider_Dimension
DM_Providers
DM_Transfusion
DM_Visit_Dimension

TumReg_CoreTable (view) – aggregates the core information about each breast cancer patient and their first relevant (malignant) diagnosis into a single view

Diagram of Tumor Registry relationships

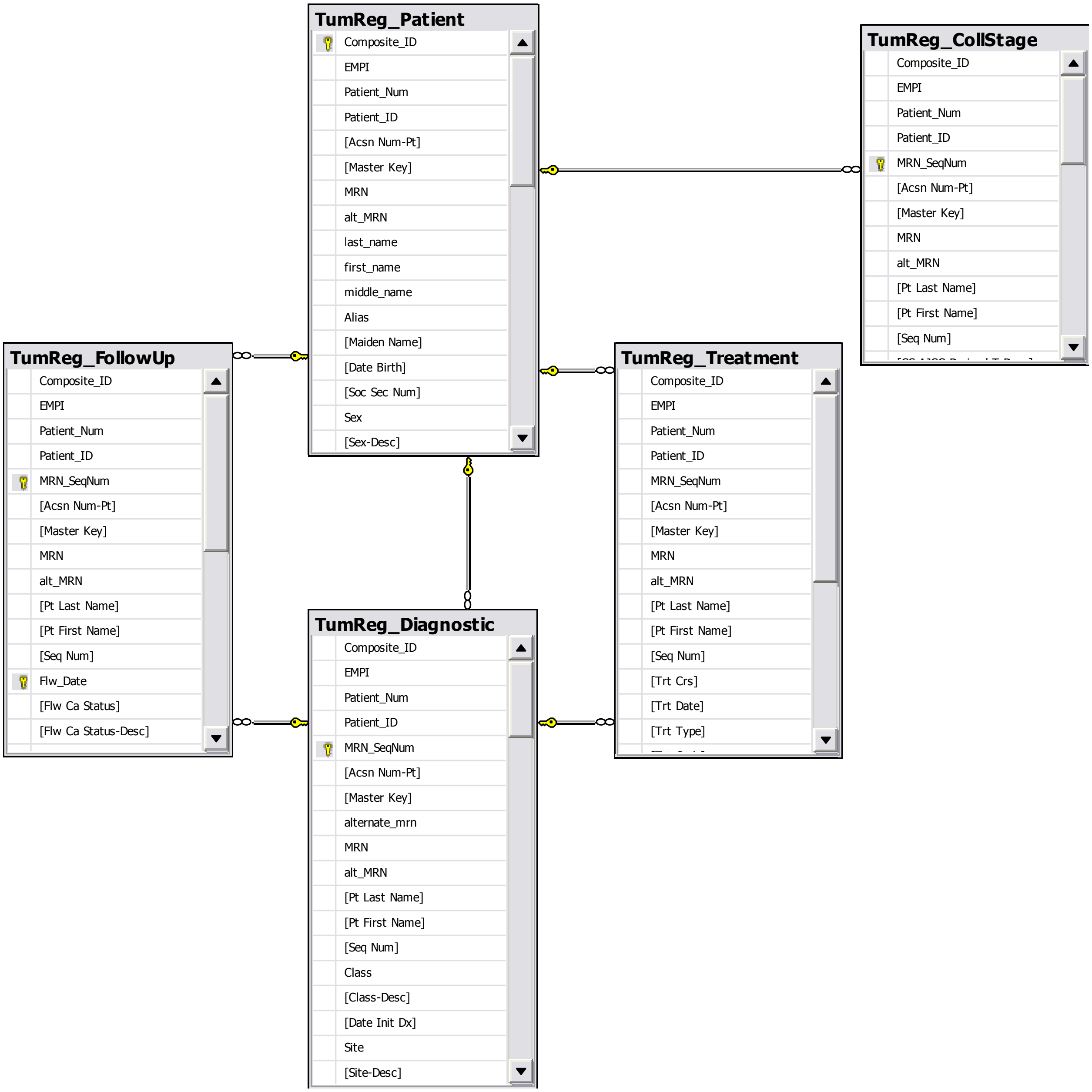
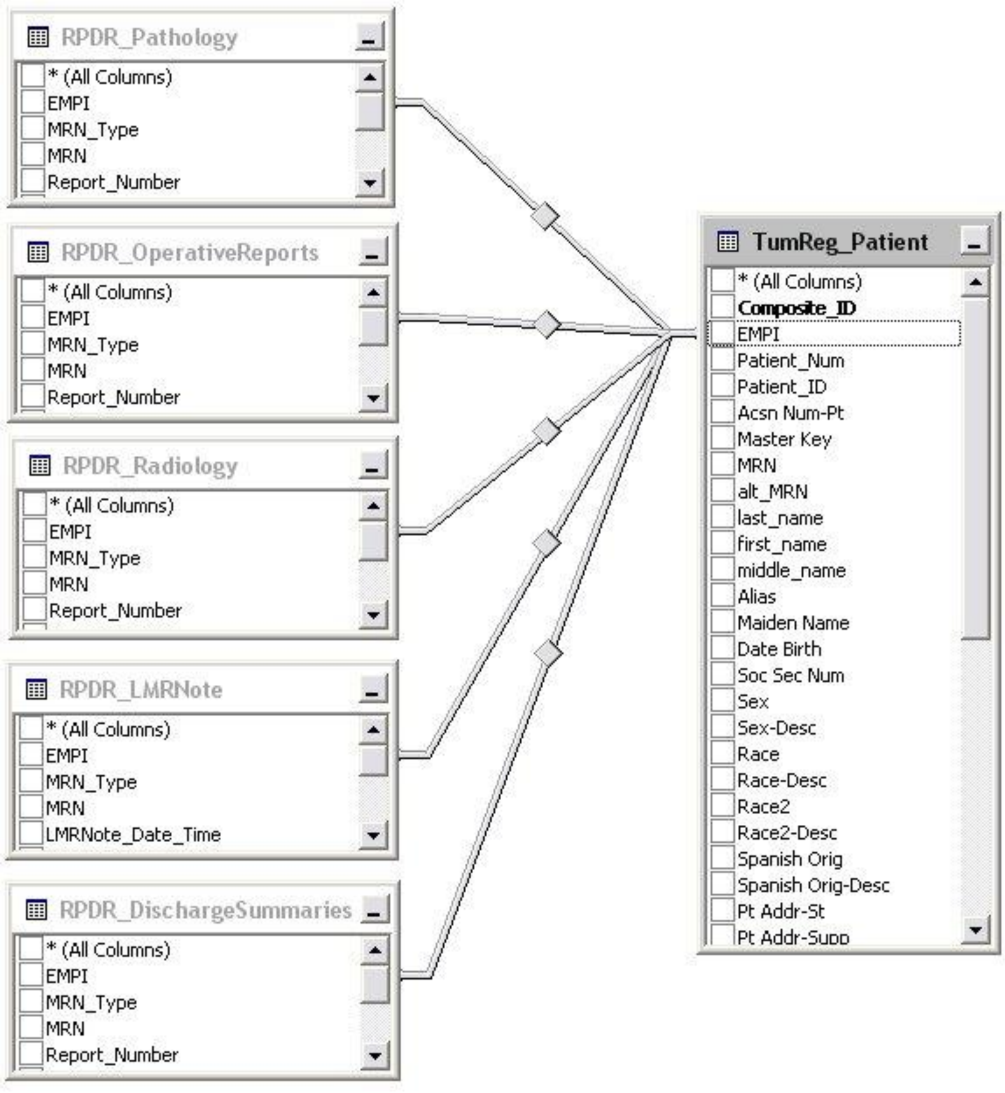
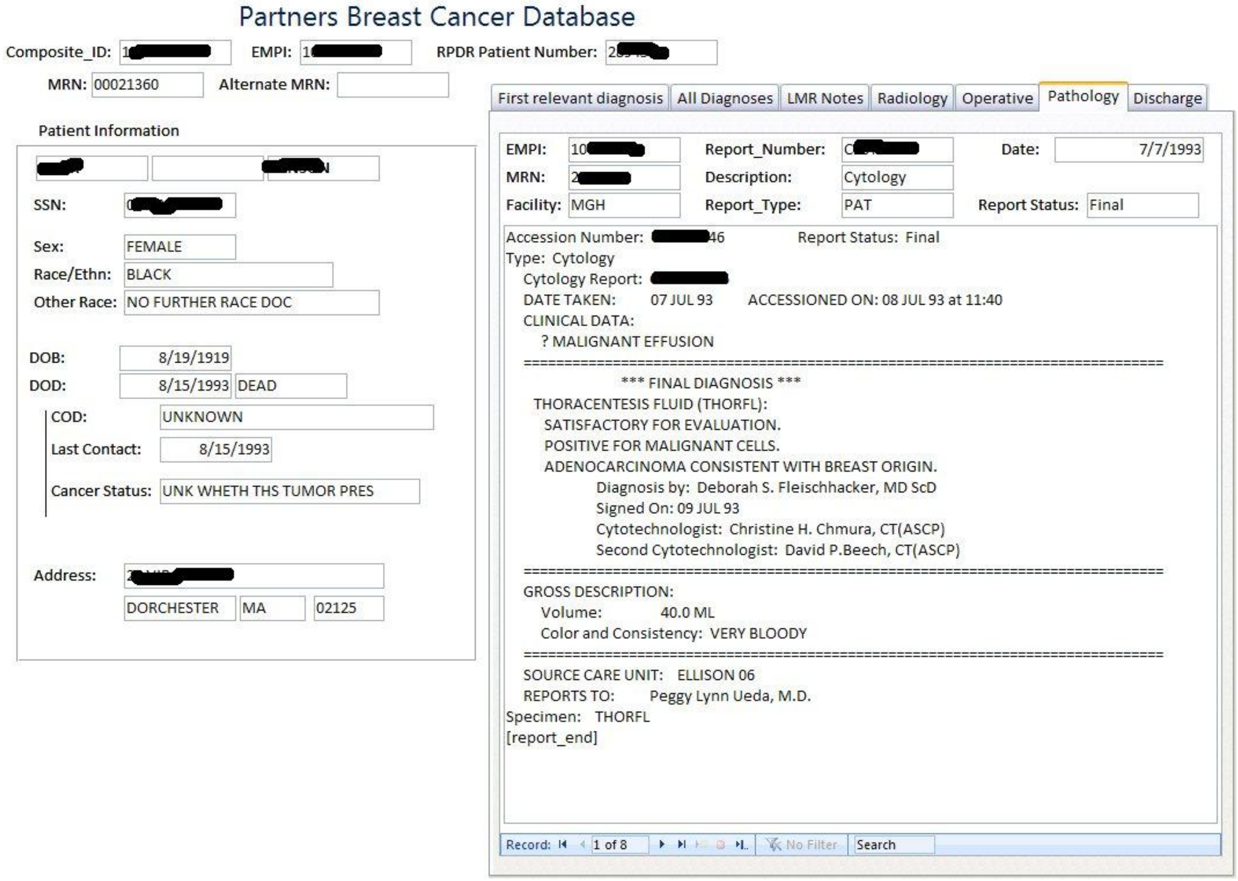


Diagram of Tumor Registry – RPDR Notes Relationships



Note that EMPI is not a Primary Key in the TumReg_Patient table, because not every patient from the Tumor Registry has been assigned an EMPI. However, all EMPIs in the Patient table are unique.

Main Form for Reviewing Individual Records



Partners Breast Cancer Database

Composite_ID: 1[redacted] EMPI: 1[redacted] RPDR Patient Number: 2[redacted]

MRN: 00021360 Alternate MRN:

Patient Information

[redacted] [redacted]

SSN: [redacted]

Sex: FEMALE

Race/Ethn: BLACK

Other Race: NO FURTHER RACE DOC

DOB: 8/19/1919

DOD: 8/15/1993 DEAD

COD: UNKNOWN

Last Contact: 8/15/1993

Cancer Status: UNK WHETH THS TUMOR PRES

Address: [redacted]

DORCHESTER MA 02125

EMPI: 10[redacted] Report_Number: C[redacted] Date: 7/7/1993

MRN: 2[redacted] Description: Cytology

Facility: MGH Report_Type: PAT Report Status: Final

```
Accession Number: [redacted]46          Report Status: Final
Type: Cytology
   Cytology Report: [redacted]
   DATE TAKEN:       07 JUL 93      ACCESSIONED ON: 08 JUL 93 at 11:40
   CLINICAL DATA:
      ? MALIGNANT EFFUSION
   ===========================================================================
                    *** FINAL DIAGNOSIS ***
     THORACENTESIS FLUID (THORFL):
       SATISFACTORY FOR EVALUATION.
       POSITIVE FOR MALIGNANT CELLS.
       ADENOCARCINOMA CONSISTENT WITH BREAST ORIGIN.
              Diagnosis by:  Deborah S. Fleischhacker, MD ScD
              Signed On: 09 JUL 93
              Cytotechnologist:  Christine H. Chmura, CT(ASCP)
              Second Cytotechnologist:  David P.Beech, CT(ASCP)
   ===========================================================================
   GROSS DESCRIPTION:
      Volume:           40.0 ML
      Color and Consistency:  VERY BLOODY
   ===========================================================================
   SOURCE CARE UNIT:   ELLISON 06
   REPORTS TO:        Peggy Lynn Ueda, M.D.
Specimen:   THORFL
[report_end]
```

The Core Table

The Core Table exists as a 'view' called "TumReg_CoreTable" in the Partners Breast Cancer Database. It aggregates the core information about each patient and their first malignant diagnosis (or first in situ if they have never had a malignant tumor) into a single view. The table contains one record for each patient in the database and can be easily exported into Excel, Access, or a flat file for further manipulation or review. This CoreTable is useful because it offers a simple, concise, patient-centric record of every digitally-captured first primary malignant (or benign) diagnosis of breast cancer to occur at the Massachusetts General Hospital and Brigham and Women's Hospital.

There is one record for every patient in the TumReg_Patient table, but each patient may have more than one diagnosis, as recorded in the TumReg_Diagnostic table. A great deal of research dealing with cancers focuses on selecting the first primary malignant tumor. This allows researchers to assess the earliest point at which cancer was detected in a patient, and yields a better understanding of how the disease progresses (rather than considering a secondary diagnosis, when cancer may have lingered in the body from the initial onset of disease). The follow-up time is then calculated from the time of first malignant diagnosis until the last date of contact with the patient or the patient's recorded death.

However, the process of selecting the first "relevant" diagnosis is actually rather complicated. It involved a cascade of queries to identify different classes of diagnoses, and then marking the record of interest. For example, patients with only 1 record in the TumReg_Diagnostic table had that record flagged for inclusion. Patients with exactly two diagnoses where only one was malignant had the malignant record flagged. Patients with exactly two malignant diagnoses with significantly different dates of diagnosis had the record of the earlier diagnosis flagged. The process continued in this way until each patient had exactly one TumReg_Diagnostic record flagged for inclusion in the CoreTable. Note that while we did filter the Diagnostic table to select exactly one case per patient, the strictly unambiguous cases are flagged in the field "CT_base," in which we captured the reliable, first-primary records for 24,007 of the 24,763 patients. The other "CT_..." fields contain flagged records for the remaining patients, but these cases were difficult to select because of multiple or similar diagnoses and are less reliable. Therefore, the TumReg_CoreTable query includes only those 24,007 patients for whom we could easily determine the first diagnosis of breast cancer. Full explanation of the filtering process can be found in Appendix B – Determining the CoreTable.

Appendix A - Constructing the Database

Data from the Tumor Registry

On 10/4/07, MN requested the Tumor Registry data for all patients diagnosed with breast disease, for as far back as the registry dates. On 10/17/07, MN received the data on a CD from Carol Venuti. The CD contained the following 12 files:[2]

- Breast_Coll_Stage_BWH_Oct07.xls
- Breast_Coll_Stage_MGH_Oct07.xls
- Breast_Diagnostic1_BWH_Oct07.xls
- Breast_Diagnostic1_MGH_Oct07.xls
- Breast_Diagnostic2_BWH_Oct07.xls
- Breast_Diagnostic2_MGH_Oct07.xls
- Breast_Patient_BWH_Oct07.xls
- Breast_Patient_MGH_Oct07.xls
- Breast_Treatment_BWH_Oct07.xls
- Breast_Treatment_MGH_Oct07.xls
- Breast_FollowUp_MGH_Oct07.xls
- Follow_up_BWH_Oct07.txt

[2] Location of these files: "\source_data\"

11 of the files are Excel 97-03 format, while 1 is a simple comma-delimited txt. The data from the txt file contained too many records for the older Excel file format, and could only be saved as an ASCII file.

Essentially, the Tumor Registry gave us 5 tables of data: Collaborative Staging, Diagnostic, Patient Demographics, Treatment, and Follow-Up. The tables were subdivided by hospital (MGH for Massachusetts General Hospital, BWH for Brigham and Women's Hospital), but the fields are exactly the same for each institution. The Diagnostic data fields were, furthermore, split into two separate files, but the case fields are meant to report on the same diagnosis.

Fields in the Data Files from the Tumor Registry

Italicized fields indicate fields that will be used to join records between different tables.

Collaborative Staging

Acsn Num-Pt
Master Key
Med Rec Num
Pt Last Name
Pt First Name
Seq Num
CS AJCC Derived T-Desc
CS AJCC Derived T
CS AJCC Derived N-Desc
CS AJCC Derived N
CS AJCC Derived M-Desc
CS AJCC Derived M
CS AJCC Derived Stage
CS Derived SS2000
CS Tumor Size
CS Extension
CS Size/Ext Eval
CS Lymph Nodes
CS Reg Nodes Eval
CS Mets at Dx
CS Mets Eval
CS SS Factor 1
CS SS Factor 2
CS SS Factor 3
CS SS Factor 4
CS SS Factor 5
CS SS Factor 6

Diagnosic 1

Acsn Num-Pt
Master Key
Med Rec Num
Pt Last Name
Pt First Name
Seq Num
Class
Class-Desc
Date Init Dx
Site
Site-Desc
Laterality
Laterality-Desc
ICDO3 Histo
ICDO3 Histo-Desc
ICDO3 Behav
ICDO3 Behav-Desc
Grade
Age At Dx
Tum Mark 1
Tum Mark 1-Desc
Tum Mark 2
Tum Mark 2-Desc
Grade-Desc
AJCC Ed
Tumor Size
Reg LN Inv
Reg LN Ex
Gen Sum Stg
Gen Sum Stg-Desc
Path T
Path N
Path M
Path Descriptor
Path Descriptor-Desc
Path Stgd By
Path Stgd By-Desc
Clin T
Clin N
Clin M
Clin Descriptor
Clin Descriptor-Desc
Clin Stgd By
Clin Stgd By-Desc
Dist Site 1
Dist Site 1-Desc
Dist Site 2
Dist Site 2-Desc
Dist Site 3
Dist Site 3-Desc

Diagnostic 2

Acsn Num-Pt
Master Key
Med Rec Num
Pt Last Name
Pt First Name
Seq Num
Surg Margins
Surg Margins-Desc
Date 1st Rec
1st Rec Type
1st Rec Type-Desc
Date Last Contact
Ca Status

Ca Status-Desc
Cause Expir
Cause-Expir-Desc
Fam Hx
Fam Hx-Desc
Smoke Hx
Smoke Hx-Desc
Alcoh Hx
Alcoh Hx-Desc
Marital Status
Marital Status-Desc
DGX_XD1PERHX
QA-Remarks

Treatment

Acsn Num-Pt
Master Key
Med Rec Num
Pt Last Name
Pt First Name
Seq Num
Trt Crs
Trt Date
Trt Type
Trt Code

Trt This Fac
Trt Fac Code
Trt I-O
Trt I-O-Desc
Trt Src
Trt RLNS
Trt RLNS-Desc
Trt SRDS
Trt SRDS-Desc

Patient

Acsn Num-Pt
Master Key
Med Rec Num
Pt Last Name
Pt First Name
Pt Middle Name
Alias
Maiden Name
Date Birth
Soc Sec Num
Sex
Sex-Desc

Race
Race-Desc
Race2
Race2-Desc
Spanish Orig
Spanish Orig-Desc
Pt Addr-St
Pt Addr-Supp
Pt City
Pt State
Pt Zip
Pt Phone Num

Current Occup
Current Industry
Employer
Longest Occup
Longest Industry
Date Expir
Death Match Indctr
Death Loc
Autopsy
Autopsy-Desc
Vital

Follow-Up

Acsn Num-Pt
Master Key
Med Rec Num
Pt Last Name
Pt First Name
Seq Num
Flw Date
Flw Ca Status
Flw Ca Status-Desc
Flw Cnt Mth
Flw Cnt Mth-Desc
Flw Cnt Phys
Flw Cnt Phys Last Name
Flw Cnt Phys First Name
Flw Comments
Flw Rec Date
Flw Rec Type
Flw Rec Type-Desc
Flw Rec Dist Site 1
Flw Rec Dist Site 1-Desc
Flw Rec Dist Site 2
Flw Rec Dist Site 2-Desc
Flw Rec Dist Site 3
Flw Rec Dist Site 3-Desc

Cleaning and Importing the Tumor Registry Data

As noted previously, the Tumor Registry sent the data in 11 Excel and 1 comma-delimited txt file. In order to turn this data into a more useful format, we needed to clean it and import into a database (we will use Access 2007).

The raw data records contained lots of space-filling characters (e.g. record for Soc Sec Num field entered as “000-00-000 ”). Importing all these extraneous spaces would unnecessarily increase the size of our database and make it inefficient. In order to make the data more manageable, we needed to remove these superfluous spaces. MN created a new Excel spreadsheet named “TumReg_data_trimmed_consolidated.xlsx.” He created a worksheet in the new spreadsheet for each one of the corresponding Tumor Registry data files. Then, in the new worksheet, using the =TRIM(“ some text or a pointer ”) function to remove leading and trailing spaces, MN pointed the TRIM function for each cell in the new worksheet at its corresponding cell in the original data worksheet. The new worksheets now contain trimmed versions of the original data, but they were only pointers. In order to make the trimmed data physically exist in the new worksheet, he used copy all → paste special → values right back into the new worksheet. The spreadsheet

TumReg_data_trimmed_consolidated.xlsx then contained all the source data, trimmed and consolidated.[3]

MN changed the "Data Type" of all records in "TumReg_data_trimmed_consolidated.xlsx" to "Text." This step will prevent Access from encountering importation errors, as Access may try to import the "Date Birth" field in date format, but not all records in "Date Birth" are in the correct date format. Access will now import all records from the Tumor Registry data as "text" data types, and we will then reformat the fields once they are part of the Access DB.

MN created an Access 2007 database, which is the main DB file for the Tumor Registry data:

TumReg_breast_08.accdb

MN imported the data from the trimmed and consolidated spreadsheet "TumReg_data_trimmed_consolidated.xlsx." All fields from all tables were imported as "text." First, MN used the Excel → Access data import tool in Access to import the data from the "Coll_Stage_BWH" worksheet of "TumReg_data_trimmed_consolidated.xlsx," naming the new Access table "Coll_Stage." Then, MN then created a new field named "Hospital," and updated the value of this field to BWH. Then MN used the Excel → Access data import tool to import the data from the Excel worksheet "Coll_Stage_MGH." This time, instead of creating a new table, MN appended this data to the newly-created Access table "Coll_Stage." Then, for each record with a Null value for "Hospital," MN updated those records to "MGH." These same steps were repeated for the excel worksheets Diagnostic 1, Diagnostic 2, Patient, Treatment, and FollowUp.[4] The two Diagnostic tables were then merged together into a single table using a maketable query, joined on fields "Acsn Num-Pt," "Master Key," "Med Rec Num," and "Seq Num."[5] There are now 5 main tables in the DB, "CollStage," "Diagnostic," "FollowUp," "Treatment," and "Patient." Once every new table was complete, MN verified that the number of records from the Tumor Registry source data agreed with the number of records in the new tables.

At this point, 5 tables existed in the new Tumor Registry breast database. As noted previously, MN imported all fields as "text" to avoid importation errors in Access. In order to make the data more useful, MN recoded the data type of some fields to match the appropriate format. If the data had to be modified before changing the data type, MN made a note in the field comments (e.g. all dates coded as "0000/00/00" were recoded to Null so that the data type could be successfully changed to a date). Also, because many fields in several tables contained duplicate fields – one with a numeric code and one

[3] Note that Follow_up_BWH_Oct07.txt was first imported into the new Excel worksheet, and the same steps were then used to trim the data. However, upon import, MN noticed that some records were corrupted. After examining the source txt file, MN saw that a few records contained line breaks/delimiters in the middle of the record, causing field-mismatches. MN corrected the improper line breaks on lines 24022, 28192, 28228, 38203, 44132, and 54574 in Follow_up_BWH_Oct07.txt and re-imported the data into "TumReg_data_trimmed_consolidated.xlsx."

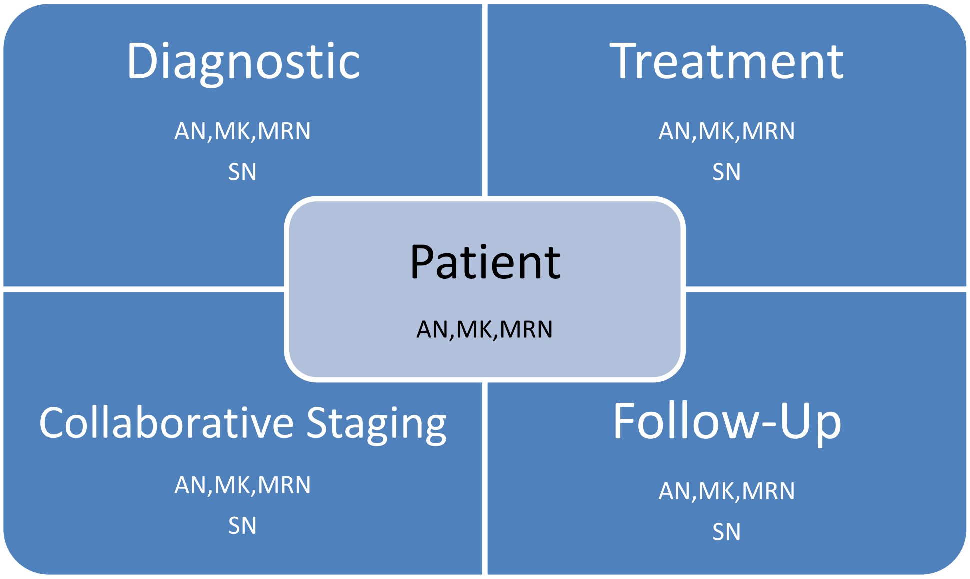
[4] Note that some of the field names from the worksheets in "TumReg_data_trimmed_consolidated.xlsx" needed to be slightly modified (manually) in order for Access to import them. For example, there cannot be slashes in an Access field name.

[5] MN deleted the preliminary tables "Diagnostic1" and "Diagnostic2" because all their data is now contained in the merged table "Diagnostic," and importing the tables from the source is trivial, should the need arise.

with the text description – MN deleted the numeric corresponding numeric code fields to eliminate superfluous data.

Relationships in the Tumor Registry Data

In order to appropriately join records between Tumor Registry data tables, we needed to understand how the tables were related. The 5 main tables are linked by several identifying fields: "Acsn Num-Pt," "Master Key," "Med Rec Num," "Pt Last Name," "Pt First Name," and "Seq Num." In theory, each patient in the database has a unique Med Rec Num. Therefore, there *should* not be any duplicates of this field for records in the Patient table, which records simple patient demographics, and does not record information about (multiple) diagnoses (i.e., there is no "Seq Num" in the Patient table). For each record of a tumor, the Seq Num (for "sequence number") indicates the order of records for patients who may have had more than one tumor in the registry. The numbering begins at "00" for patients who have only 1 record in the Tumor Registry, and "01" for patients who have more than 1 record in the Tumor Registry. [6]



EXPECTED Keys for Relationships between tables: AN = Acsn Num-Pt, MK = Master Key, MRN = Med Rec Num, SN = Seq Num

When analyzing the patient table, MN found that there were 3 duplicates of "Med Rec Num": there were five Nulls, two "10964393", and two "3631374." MN deleted these 9 records from the Patient table, because the Patient table must contain unique IDs in order for the relationships to function correctly between the tables ("Med Rec Num" must be a primary key for the Patient table). MN then deleted all records from other tables with MRN = Null, "10964393", or "3631374" because the Patient table must have referential integrity to the other tables for which it serves as a foreign key. Moreover, we needed to have unique MRNs when requesting further data from the RPDR.

[6] Note that some patients may have more than one tumor in the Tumor Registry, but for different diseases. Thus, it is possible that a patient with only one record in our *breast* cancer database has a corresponding sequence number of "03" because she had two prior tumors that were not breast-related.

In addition to duplicate MRNs in the Patient data, there also exists a problem with patients who visited both the Mass General Hospital and the Brigham and Women's Hospital. After 1978, there are 445 patients who fall into this category (445/25287 ~ 2% of all patients recorded in Patient table).[7] These patients are essentially listed twice in the patient table, with unique MRNs, ANs, and MKs for each hospital they visited. Additionally, there is 1 patient who was assigned two MRNs but was seen only at the Mass General. The duplicate entries carry into the CollStage, FollowUp, Treatment, and Diagnostic tables.

In order to rectify the problem, we needed to find a way to link the records that contain information on one unique patient, for which MRN is no longer useful because 445 patients are assigned two MRNs in the data. MN found that the identifiers 'Pt Last Name', 'Pt First Name', and 'Date Birth' provide a useful key that identifies unique patients.[8] MN created a few field, 'Patient_ID', that will serve as the primary key for Patient, concatenating 'Pt Last Name', 'Pt First Name', and 'Date Birth.' This new field was then added and updated as a foreign key for the other four tables in the database.[9] MN deleted records from the Patient table for the duplicate with the "Min" of 'Med Rec Num'; i.e., the smallest (starting with value of first character) of MRNs was removed from the patient table. This was a somewhat arbitrary way of removing half of the entries, as MN observed that the data recorded about patients at each hospital was almost identical. To reiterate, when queried by last name, first name, and DOB, there were originally 446 people with more than one MRN in the Patient table, meaning there were 892 total records containing duplicates. For each pair of records, MN removed the record with the smaller MRN, thereby removing 446 records from the Patient table. Patient_ID was then set as the primary key for the Patient table. The Patient table now contains records for 24841 patients.

Finally, there are also several records in the Patient table where a data entry error occurred in the name fields. In these 25 cases, the query to catch duplicate patients by first name, last name, and DOB failed because the name varied slightly (e.g. first name of 'Cindy' and 'Cynthia' received separate MRNs when it is obvious, based on other fields, that they are the same person). In order to catch these duplicates, MN queried the Patient table for a count >1 of 'Med Rec Num', grouped by 'Date Birth' and 'Soc Sec Num.' This query revealed the patients with identical social security numbers and dates of birth who have different values for MRN *and* who have different values for first and last name. There are 25 persons in the database like this, who represent a total of 50 records in Patient. Because it would be difficult to distinguish which case contained the "correct" information, we decided to delete these 50

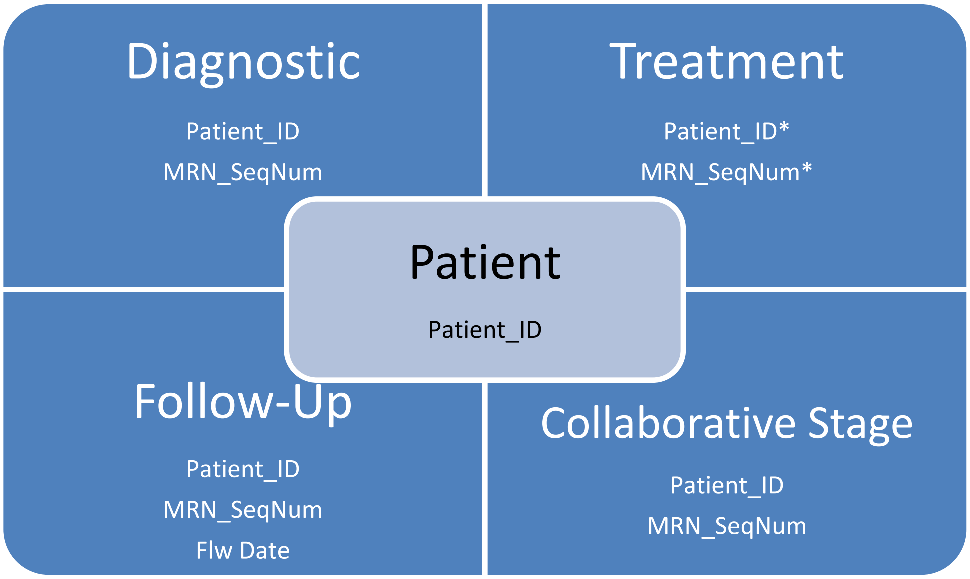
[7] These duplicates were discovered by using a "Totals" query, selecting count of 'Med Rec Num' >1 while grouping by 'Pt Last Name,' 'Pt First Name,' and 'Date Birth.'

[8] While there may still exist some duplicated patient entries in the database, it would be difficult to find another way to identify unique individuals. For instance, if we wanted to try a person's street address as a possible unique key, it turns out that the records for each hospital entered the information differently, e.g. Whitmer Rd. and Whitmer Road. After some deliberation, MN found the last name-first name-DOB fields to be the most efficient at finding unique patients. Additionally, SSN information is not as complete as DOB.

[9] MN created the a new concatenated field, rather than simply use 3 fields as a primary key, because the 'Date Birth' field is not recorded in any table besides Patient. Additionally, it will probably be more useful that way if tables are imported into programs like Excel.

records from the database.[10] The Patient table now contains 24791 unique patients, verified by First Name, Last Name, Date of Birth, and Social Security Number.

Because we had found that MRN was not able to be used as a unique patient identifier, we will discuss the new relationships in our database. 'Patient_ID' is the primary key for the Patient table, and will be used as a foreign key in the other tables to identify records that belong to the same patient (whether seen at the MGH or BWH). Initially, we sought to identify records between tables that described the same visit using the 'Seq Num' in combination with the MRN. Because we rendered the MRN data defunct, we will link records that describe the same visit using a concatenation of MRN and Seq Num, called "MRN_SeqNum."[11] Our new schema looks like the following:



ACTUAL Primary Keys for Relationships between tables; * Patient_ID and MRN-SeqNum are NOT primary keys for the Treatment table

Therefore, when querying data on the same patient between different tables, we will join the tables on Patient ID. When querying records on a particular diagnosis from one patient between tables, we will join the tables on Patient ID, and MRN_SeqNum.

[10] Once we had set the primary key for the Patient table to 'Patient_ID,' we enforced referential integrity for cascading updates and deletes when establishing relationships between the 'Patient_ID' foreign keys in the other tables and the primary key in the Patient table. Therefore, when we delete records from the Patient table, Access deletes corresponding records from the 4 peripheral tables.

[11] The Seq Num is always relative to the MRN, and although we now identify patients based on Patient_ID, the Seq Num must be referenced to the original facility at which the patient had an encounter (the MRN was unique for each patient at each hospital). To really understand the sequence of patient records, one needs to consider the date information for encounter records.

Note that the FollowUp table contains more than one record for each diagnosis. Thus, Patient_ID and MRN_SeqNum did not suffice as a primary key for the table. In order to differentiate between FollowUp records on the same diagnosis, we must also consider the Flow Date (“Flw Date”). Similarly, the Treatment table contained more than one record per diagnosis, and consequently, Patient_ID and MRN_SeqNum cannot function as a primary key. Furthermore, while the Treatment table does contain fields on the Treatment Date (“Trt Date”) and the Treatment Code (“Trt Code”) , neither field is useful as a Primary Key because the data are incomplete. Therefore, the Treatment table will have no primary key, but can return all the Treatment records corresponding to a single diagnosis when joined with another table on Patient_ID and MRN_SeqNum.

UPDATE: After attempting to assign EMPIs to the patients (we receive EMPIs from the RPDR, based on the MRNs), we realized that the Tumor Registry Patient table still contained some

Sources of Key Field Information in the Tumor Registry

Tumor Size

CollStage: 'CS Tumor Size' (in mm) – removed leading zeros, converted to number; 'CS Tumor Size' contains information only for cases diagnosed after 1/1/2004

Diagnostic: 'Tumor Size' (in cm) – removed decimal point (now in mm), removed leading zeros, converted to number; in some cases Diagnostic contains tumor size info when CollStage does not, and vice versa; in some cases Diagnostic tumor size information does not agree with CollStage information

Lymph Nodes

CollStage: 'CS Lymph Nodes' – identifies if positive lymph nodes present, but not a specific number; 'CS Reg Nodes Eval' – identifies whether nodes were examined, but not a specific number

Diagnostic: 'Reg LN Inv' – number of positive lymph nodes; 'Reg LN Ex' – number of lymph nodes biopsied

RPDR Data

The Research Patient Data Registry is the single most comprehensive source of patient data within the Partners Healthcare system. Our main contact at the RPDR is Vivian Gainer, and the Director of the entire RPDR project is Shawn Murphy, MD, PhD, both of whom are exceptionally helpful. The RPDR manages both discreet data (fields with singular values) and long text data in the form of "notes" (e.g. radiology reports, operative notes, etc.).

The RPDR has two main methods for obtaining the discreet data. One is to use their web-based utility to request specific data about a set of patients via MRNs (enter http://rpdr/ in a browser on a Partners machine). However, there is no easy way to request *all* possible data (one much manually select each desired field, and there are thousands of lab results). Therefore, for large data requests, the RPDR can create a private "data mart" that acts like a mini version of the entire RPDR, but contains only the information that pertains to a set of MRNs. We chose this option in an effort to be as thorough as possible in generating our Partners Breast Database, using the MRNs we received from the Tumor Registry.

RPDR Data Mart

The MS SQL data mart resides on an RPDR server, and we have full read/write access to our particular mart.[12] Our list of MRNs comes from the data we received from the Tumor Registry. After eliminating a few duplicate MRNs from the original source files "Breast_Patient_BWH_Oct07.XLS" and Breast_Patient_MGH_Oct07.XLS," MN generated the txt file "RPDR_MRN_request.txt" that contained 25,289 MRNs.[13] Vivian needed about a week to make arrangements for creating the database.

Once the data mart was ready to be populated, we first had to upload our MRN list to the table "incomingmrns" in the processing database BreastCancer_Mart on server RPDRutil.mgh.harvard.edu and flag our request as 'ready to be processed' by changing a few values in the "Processing" table (Vivian actually did this step for us).

Two days later, our data mart was ready, populated with all available data about the MRNs we submitted. The name of our data mart is

"BreastCancer_Mart" on MS SQL Enterprise Server "phssql251.mgh.harvard.edu"

Initially, our login username (not listed in this document) did not have adequate permissions to access the database, but Vivian promptly fixed that issue once notified.

We also had to meet with Vivian Gainer and Shawn Murphy to discuss security issues about the data mart. All data in the mart follow the standard HIPPA regulations – only IRB-approved study staff may look at the data, and all methods of accessing the data must be secured physically (locked room) and digitally (password/encryption). We had to provide the IRB protocol, 2003 P 000606 "An Analysis of Breast Cancer Survival," to Vivian for their records at the RPDR.

While the data mart contains de-identified patient information (only a patient number), those patient numbers can be mapped to MRNs and EMPIs using the MRN_mapping table on the RPDRutil server. The patient mappings are kept on a separate server from the data mart to create another layer of security and privacy.

Tables and Fields in the RPDR Breast Data Mart

The RPDR breast data mart (BreastCancer_Mart) contained 21 tables when it was first created:

Concept_Dimension
Demographics
Diagnosis
DRG

[12] At first, the permissions weren't right for our mart. We could connect to the server, but we didn't have Admin privileges for our database. Vivian quickly fixed this issue once I brought it to her attention.

[13] Note that this number exceeds the number of actual Patients in the Core Table we created from the Tumor Registry

dtproperties
Encounter_Mapping
Encounters
HealthHistory
HPCGG
LabTests
Medications
MicroBiology
Observation_Fact
Patient_Dimenion
Patient_Mapping
Procedures
Processing
Provider_Dimension
Providers
Transfusion
Visit_Dimension

The main table, Observation_Fact, contains the actual values of the "observations" in the database. This table has almost 28 million cases.

RPDR 'Notes' Data

While the data mart provides exhaustive discreet data, it does not include any 'notes'; there are no operative notes, radiology notes, discharge summaries, etc. While the RPDR does manage this data, it is not available in the data mart. Therefore, we had to request all long-text notes in a separate RPDR request, using the normal RPDR web-utility, available at http://rpdr/ from a Partners workstation. We used "create a detailed data request" and uploaded the same MRNs as submitted for the data mart, from the file RPDR_MRN_request.txt. In order to make the data returned manageable, we actually submitted two requests, one for each hospital – BWH and MGH. We chose the following fields, to obtain all possible notes:

A list of patient medical record numbers
Cardiology Reports
Discharge Summaries
Endoscopy Reports
Operative Notes
Pathology Reports
Pulmonary Reports
Radiology Reports

Identifying Patient Information - not available for Limited Data Sets
LMR outpatient notes- not available for Limited Data Sets

When one requests the data in this manner, the RPDR returns several executable files, which, when run, demand a password, and return the decrypted version of the data in flat txt format. We obtained the following decrypted files:

Filename	**Description**
BWH data:	
men18_032708120912282466_Car.txt	Cardiology reports
men18_032708120912282466_Con.txt	Contact data
men18_032708120912282466_Dis.txt	Discharge summaries
men18_032708120912282466_End.txt	Endoscopy reports
men18_032708120912282466_Lno.txt	LMR outpatient notes
men18_032708120912282466_Log.txt	Logfile for processing of request
men18_032708120912282466_Mrn.txt	List of MRNs
men18_032708120912282466_Opn.txt	Operative notes
men18_032708120912282466_Pat.txt	Pathology reports
men18_032708120912282466_Pul.txt	Pulmonary reports
men18_032708120912282466_Rad.txt	Radiology reports
men18_032708120912282466_Rnd.txt	Radiology image data
processing_log.txt	Another processing log
Report for Detailed Data - MRN.doc	Report on any errors encountered
RPDR_men18_032708120912282466_Let.txt	Explanation of data returned
MGH data:	
men18_032708122629887017_Car.txt	Cardiology reports
men18_032708122629887017_Con.txt	Contact data
men18_032708122629887017_Dis.txt	Discharge summaries
men18_032708122629887017_End.txt	Endoscopy reports
men18_032708122629887017_Lno.txt	LMR outpatient notes
men18_032708122629887017_Log.txt	Logfile for processing of request
men18_032708122629887017_Mrn.txt	List of MRNs
men18_032708122629887017_Opn.txt	Operative notes
men18_032708122629887017_Pat.txt	Pathology reports
men18_032708122629887017_Pul.txt	Pulmonary reports
men18_032708122629887017_Rad.txt	Radiology reports
men18_032708122629887017_Rnd.txt	Radiology image data
processing_log.txt	Another processing log
Report for Detailed Data - MRN.doc	Report on any errors encountered
RPDR_men18_032708122629887017_Let.txt	Explanation of data returned

Upon inspecting the files, and several attempts to import the data into MS SQL, MN found that the data were not in a very useful form. The main text report field for each data file contained no text delimiters, and MS SQL Server 2005 encountered many errors during attempted importation. MN contacted Laurie

Bogosian at the RPDR, who indicated that they could re-process our request and return Access files instead of flat files (the fields are already formatted in Access, so this option was much more desirable). From the time we notified her of the problem, it took about 3 days for Laurie to get back to us with the Access files. We received the following databases with noted tables:

BWH:

- men18_032708120912282466.mdb
    - Contact Information, MRN list, Operative Reports, Pathology Reports, Pulmonary Reports, Radiology Reports
- men18_0327081209122824661.mdb
    - Cardiology Reports, Discharge Summaries, Endoscopy, LMR Notes, MRN list (duplicate)

MGH:

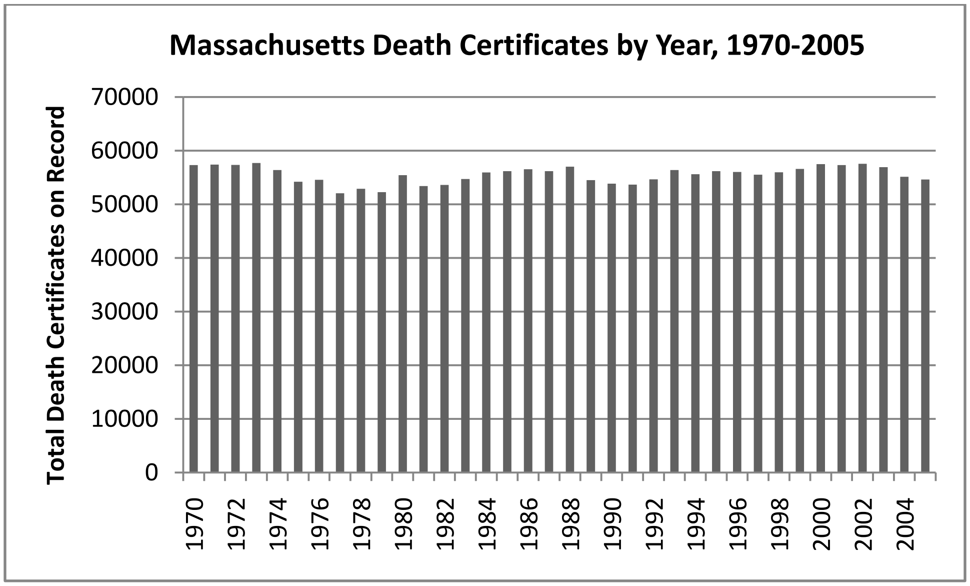
- Access_File_1.mbd – Discharge Summaries, Endoscopy, MRN list
- Access_File_2.mbd – Operative Reports, MRN list
- Access_File_3.mbd – LMR Notes, MRN list
- Access_File_4.mbd – Cardiology Reports, Pathology Reports, MRN list
- Access_File_5.mbd – Pulmonary Reports, Radiology Reports, Radiology Image data, MRN list

The Access files contain the exact same data as the flat (txt) files, but the Access files are easier to work with, so we used them. MN noticed that the BWH Access data did not contain a table for the Radiology Image data, however, so he imported the flat file men18_032708120912282466_Rnd.txt into the table “RadiologyImages” in the BWH database men18_0327081209122824661.mdb (the other Access DB, containing the “Radiology” table, was full). Additionally, the Access version of the MGH notes data did not contain a table for ContactInformation. MN imported the flat file containing the MGH Contact Information data, men18_032708122629887017_Con.txt, into the table ‘ContactInformation’ in the file Access_File_4.mdb. Now all the data for the MGH and BWH are located in Access databases with fully complementary information (all tables in Access format for the BWH also exist in Access format for the MGH).

Massachusetts Death Registry

In order to verify the death information as best we can, we corroborate date-of-death information between the Tumor Registry, the Massachusetts State Death Index, and the Social Security Master Death Index. Our contact for the Massachusetts death records is Charlene Zion (Charlene.Zion@state.ma.us). It only takes a few weeks from the time we make our request to receive the death data. Our data from the MA Death Registry contains information on deaths that occurred between 1970 and 2005, comprising almost 2 million deaths overall.

We originally had a complete version of the MA Death Index from 1970 to 2004 (MN converted the old file MA_Death_Reg.mdb to Access 2007 format, MA_Death_Reg_70-04.accdb). We requested the most up-to-date information from Charlene, which was 2005. Charlene sent us the 2005 death data, which is located in the flat file MA_Death_Index_05.TXT. The field dictionary is named MA_death_fields.doc.



Importing Local Data into MS SQL Server 2005

Once we had obtained all the data for our various databases, we imported the data into a full-scale Microsoft SQL Server 2005 (Service Pack 2) database. The database is hosted locally on the machine JSMTOWER1. The host machine is a Dell Precision 490 with an Intel Xeon dual-core processor – 2.00GHz/core with 4MB L2 cache and a 1333MHz FSB. The system has 2 GB of RAM running at 667MHz (4 DIMMS), and the main hard disk, on which the DBs are stored, runs at 10K RPM with 16MB databurst caching. The OS is Windows XP, Service Pack 2. The databases are managed in Microsoft SQL Server Management Studio, and some data retrieval interfaces use Microsoft Access 2007.

In order to get MS SQL Server 2005 on a Windows machine, we needed to install Microsoft IIS (Internet Information Systems – bundled with Windows, but not installed by default), and the MS .NET Studio Environment. We then installed MS SQL Server 2005 and the MS SQL Management Studio, which we downloaded from the Microsoft website using our MSDN license.

The MS SQL server name is JSMTOWER1, and any computer administrator may logon via Windows NT authentication. The actual path to the DB files is C:\Program Files\Microsoft SQL Server\MSSQL.1\MSSQL\DATA\, but one should not manipulate anything in this directory.

RPDR Data Mart (BreastCancer_Mart)

The breast Data Mart is not hosted locally. It is available, with full write/read access, on the server phssql251.mgh.harvard.edu.

MN imported the table MRN_Mapping from the RPDRUtil server into the local SQL database Partners_BrCaDB. This table contains the MRNs and related EMPIs (Enterprise Master Patient Index) for the RPDR data. The "Enterprise Master Patient Index" is the single numeric value that identifies a single patient moving through the Partners Healthcare system. Although one patient may have received different MRNs, one for the Brigham and one for the MassGeneral, the EMPI remains constant. Not all patients were assigned an EMPI, but every patient with 'notes' data from the RPDR has an EMPI.

Tumor Registry Data (DB: TumorRegistry_breast)

We first logged into our local SQL server (server name is computer name – JSMTOWER1) using the MS SQL Management Studio and Windows NT credentials. We then created a new database (right-click Databases → New Database) named TumorRegistry_breast. Once we created our SQL database, we wanted to import our Tumor Registry data into it. Right-clicked database name → Tasks → Import Data. Because our Tumor Registry data is in Access 2007 format, we chose "Microsoft Office 12.0 Access …" as the data source. We then needed to enter the full path to our Access DB in the "Data Source" field of the "Properties" dialogue. We continued to select the five most pertinent tables from the Tumor Registry to import: Diagnostic, Patient, CollStage, FollowUp, Treatment. We needed to "Edit mapping" on several tables because the import wizard did not identify the Date/Time fields properly. In the "Edit mapping" dialogue for each table, we had to make sure every field to be imported had a data type, and the cases which did not happened to all be explicitly set to "datetime". We finished the import wizard and all data imported successfully into our SQL database, TumorRegistry_breast. We then manually copied the SQL code from the access query "CoreTable" and saved it as a "View" in our SQL database.[14]

MN added a field for EMPI to all tables in the Tumor Registry data. MN used the table MRN_Mapping and the query assign_EMPIs_TumReg.sql.

The database TumorRegistry_breast is 185MB large.

RPDR Notes (DB: RPDRNotes_breast)

We imported the RPDR notes data from the MS Access tables given us (we did not use the txt files due to importation issues). MN created a new database named RPDRNotes_breast. He imported the Access tables Cardiology, ContactInformation, Discharge_Summaries, Endoscopy, LMRNote, Mrn, Operative_Reports, Pathology, Pulmonary, Radiology, and RadiologyImages, making sure that all date fields were imported in datetime format (they were not by default). The BWH data were imported first, and the MGH data were appended to the BWH data.

The database RPDRNotes_breast is 11.2GB large.

[14] Instead of importing data in the MS SQL Management Studio as described, it is also possible to export data into SQL from Access 2007. From inside Access 2007, there is also functionality to export Access queries as views into SQL. We choose to use the import method because the Management Studio Import Wizard seemed more robust.

Massachusetts Death Index (DB: MA_DeathIndex)

MN started with the previous MA death data from 1970-2004, MA_Death_Reg_70-04.accdb. He then imported the tables "Death Registry:1970-2004", "ICD 10", "ICD9", and "ICD8" into a new local SQL database, MA_DeathIndex. In the table MA_deaths_70to04, the field 'date_death' was originally stored as a text field, containing 8 and 6-digit strings for the date of death (yyyymmdd and yymmdd). MN created a new datetime-format field, 'DOD', and transferred all the values of date_death into the new field using the series of queries found in convert_DOD_to_datetime.sql.[15]

The 2005 Massachusetts death data is contained in the file MA_Death_Index_05.TXT. The data is width-delimited, and the dictionary for this data is named MA_death_fields.doc. MN imported this data into the database MA_DeathIndex, in the table MA_Death_Index_05. The name fields were initially split between different columns, so MN concatenated them into single fields using the query concat_names_05data.sql. The date-of-death and date-of-birth data were converted into proper date format using the queries located in convert_DOD_DOB_datetime_05data.sql. The data for sex was extracted using the query isolate_sex_info_05data.sql.

One can query against the MA death data using the view MA_death_index_master in the database MA_DeathIndex. It contains the fields for last name, first name, social security number, sex, date of death, and date of birth (if known, and in string format – can be parsed for month/day information).

The database MA_DeathIndex is 2.7GB large.

Contents

Tumor Registry (TumorRegistry_breast)

Tables (Primary Key)

- Patient (Patient_ID)
- Diagnostic (MRN_SeqNum)
- FollowUp (MRN_SeqNum, Flw_date)
- Treatment

[15] This conversion process proved to be significantly more complicated than expected. Even after reformatting the date_death values into a date-like format, some cases still contained bad data. After multiple "Arithmetic overflows," causing the conversion update to fail, MN found that 2 records contained equivalent date values of April 31st, a date that does not exist. These value of date_death for these records was reset to null, and conversion proceeded unhindered.

- CollStage (MRN_SeqNum)

Fields

Patient

Patient_ID	[Pt Addr-Supp]
[Acsn Num-Pt]	[Pt City]
[Master Key]	[Pt State]
[Med Rec Num]	[Pt Zip]
[Pt Last Name]	[Pt Phone Num]
[Pt First Name]	[Current Occup]
[Pt Middle Name]	[Current Industry]
Alias	Employer
[Maiden Name]	[Longest Occup]
[Date Birth]	[Longest Industry]
[Soc Sec Num]	[Date Expir]
Sex	[Death Match Indctr]
[Sex-Desc]	[Death Loc]
Race	Autopsy
[Race-Desc]	[Autopsy-Desc]
Race2	Vital
[Race2-Desc]	NoName
[Spanish Orig]	NoName1
[Spanish Orig-Desc]	NoName2
[Pt Addr-St]	Hospital

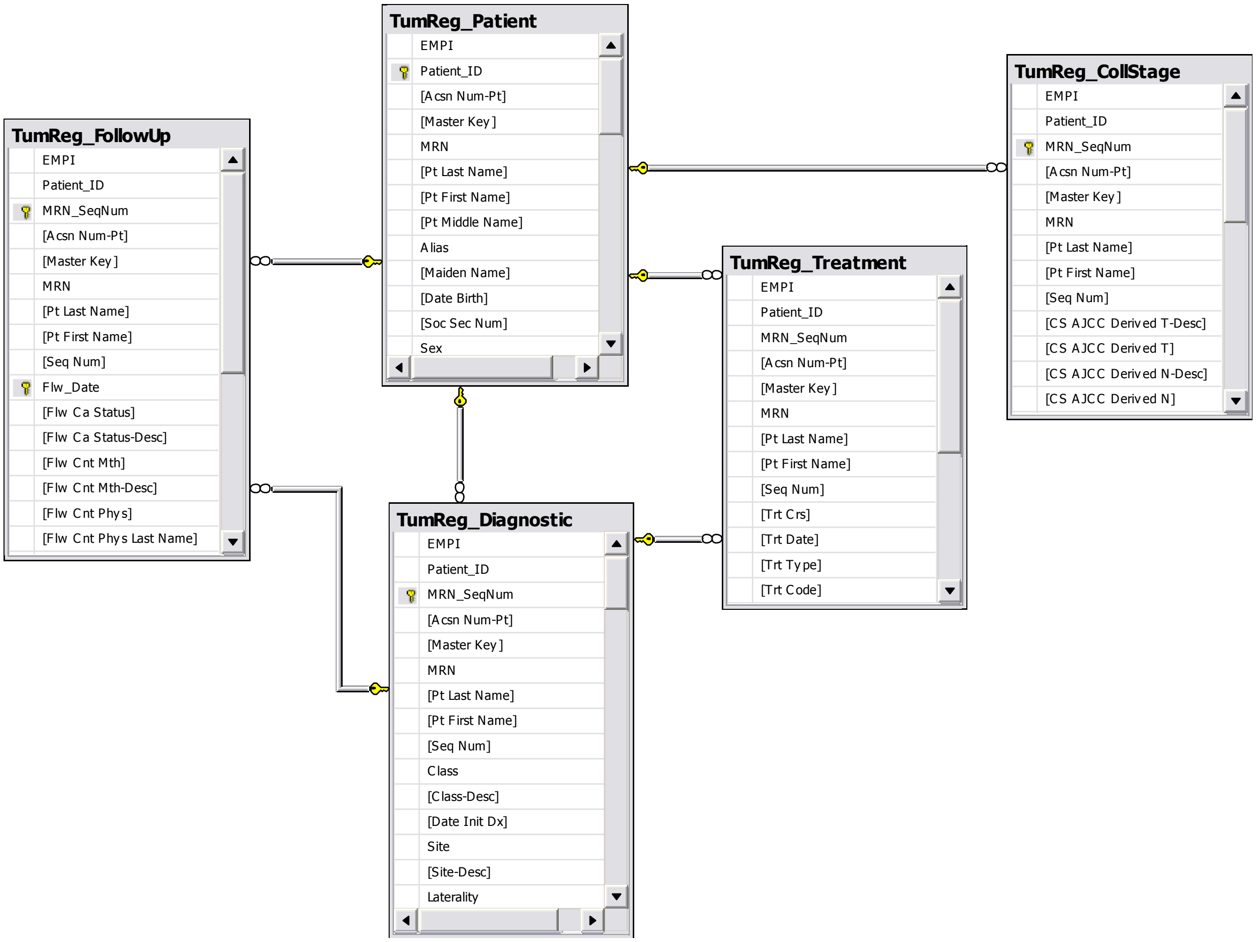
Diagnostic

Patient_ID	[Tumor Size]	[Date 1st Rec]
MRN_SeqNum	[Reg LN Inv]	[1st Rec Type]
[Acsn Num-Pt]	[Reg LN Ex]	[1st Rec Type-Desc]
[Master Key]	[Gen Sum Stg]	[Date Last Contact]
[Med Rec Num]	[Gen Sum Stg-Desc]	[Ca Status]
[Pt Last Name]	[Path T]	[Ca Status-Desc]
[Pt First Name]	[Path N]	[Cause Expir]
[Seq Num]	[Path M]	[Cause-Expir-Desc]
Class	[Path Descriptor]	[Fam Hx]
[Class-Desc]	[Path Descriptor-Desc]	[Fam Hx-Desc]
[Date Init Dx]	[Path Stgd By]	[Smoke Hx]
Site	[Path Stgd By-Desc]	[Smoke Hx-Desc]

[Site-Desc]
Laterality
[Laterality-Desc]
[ICDO3 Histo]
[ICDO3 Histo-Desc]
[ICDO3 Behav]
[ICDO3 Behav-Desc]
Grade
[Age At Dx]
[Tum Mark 1]
[Tum Mark 1-Desc]
[Tum Mark 2]
[Tum Mark 2-Desc]
[Grade-Desc]
[AJCC Ed]

[Clin T]
[Clin N]
[Clin M]
[Clin Descriptor]
[Clin Descriptor-Desc]
[Clin Stgd By]
[Clin Stgd By-Desc]
[Dist Site 1]
[Dist Site 1-Desc]
[Dist Site 2]
[Dist Site 2-Desc]
[Dist Site 3]
[Dist Site 3-Desc]
[Surg Margins]
[Surg Margins-Desc]

[Alcoh Hx]
[Alcoh Hx-Desc]
[Marital Status]
[Marital Status-Desc]
DGX_XD1PERHX
[QA-Remarks]
Hospital
CT_base
CT_morethan2records
CT_2records_similarDateDx
CT_2records_sameDateDx_sameTumSz
CT_2records_sameDateDx_diffTumSz
CT_misfits
manual_revision_notes



RPDR Notes

RPDR Data Mart

MA Death Index

Appendix B – Determining the CoreTable

The Core Table

The Core Table contains one case per patient with the most relevant data about breast disease. When creating this table, we primarily considered the Diagnostic table, as it contains the information on the dates of diagnosis and the nature of the disease (benign/malignant) for a given tumor. For patients who had multiple malignant diagnoses, we choose the diagnosis with the earliest “Date Init Dx.” For patients with multiple in situ diagnoses, we chose the diagnosis with the earliest “Date Init Dx.” For patients with both malignant and in situ diagnoses, we choose the earliest diagnosis of malignant disease. Thus, based on the minimum date of diagnosis, we used the “MRN-SeqNum” from that case to join similar cases from the other tables.

Technical details for creating the Core Table

In each of the following cases, the queries generated a new table[16], which could be joined with the Diagnostic table on the relevant fields to update the Diagnostic field ‘CT_base’ to “1” for the record of interest.

1. For patients with only 1 record in the Diagnostic table, updated CT_base field to “1” (query: CoreTable_CASES_1PatientID)
2. For patients with exactly 2 records, where only 1 of the 2 records is “malignant,” updated CT_base field to “1” for the malignant records (query: CoreTable_CASES_1insitu+1malig)

Malignant:

3. For patients with exactly 2 malignant records (but could have other in situ records) with unique dates of diagnosis more than 60 days apart, updated CT_base field to “1” for the earliest malignant record (query: CoreTable_CASES_2maligs_uniqueDateDx)[17]
4. For patients with exactly 2 malignant records (but could have other in situ records) with the same date of diagnosis and 1 record has a Null tumor size, updated CT_base field to “1” for the record containing the non-null tumor size. (query: CoreTable_CASES_2maligs_sameDateDx_1nullTumSz)

In Situ:

5. For patients with exactly 2 in situ records (and no malignant records) with unique dates of diagnosis, updated CT_base field to “1” for the earliest in situ record (query: CoreTable_CASES_2insitus_uniqueDateDx)

[16] Depending on the dependencies of the queries, MS Access will not allow update queries between a table and a query, so we had to SELECT INTO a new table with the results of our queries.

[17] MN manually selected the pertinent record for patients with malignant tumors <60 apart because they often represent the same diagnosis, but one record – not necessarily the earlier one – will contain much more relevant information on the disease. When this was the case, MN chose the case containing more information.

6. For patients with exactly 2 in situ records (and no malignant records) with the same date of diagnosis and 1 record has a Null tumor size, updated CT_base field to “1” for the record containing the non-null tumor size. (query: CoreTable_CASES_2insitus_sameDateDx_1nullTumSz)

After performing these queries to identify conclusively relevant records for inclusion in the Core Table, we had selected 24,035 patients. There are a total of 24,791 unique patients (according to the Patient table), which means that 24791-24035= 756 patients do not yet have a recorded selected for the Core Table. These patients represent the more complex cases of breast disease, and comprise such things as multi-focal and bilateral cases, patients seen at two institutions for the same tumor, etc. The following table gives a summary of the residual patients to be selected:

Condition (for a single patient)	Patients	Field Updated upon Manual Review	Query Used to Find Patients Matching Condition
More than 2 malignant records	49	CT_morethan2records	CoreTable_CASES_2+maligs
More than 2 in situ records[18]	6	CT_morethan2records	CoreTable_CASES_2+insitus
2 malignant records, Date Dx < 60 days apart	191	CT_2records_similarDateDx	CoreTable_CASES_2maligs_similarDateDx
2 in situ records, Date Dx < 60 days apart	51	CT_2records_similarDateDx	CoreTable_CASES_2insitus_similarDateDx
2 malignant records, same DateDx, same TumSz[19]	205	CT_2records_sameDateDx_sameTumSz	CoreTable_CASES_2maligs_sameDateDx_sameTumSz
2 in situ records, same DateDx, same TumSz	64	CT_2records_sameDateDx_sameTumSz	CoreTable_CASES_2insitus_sameDateDx_sameTumSz
2 malignant records, same DateDx, different TumSzs[20]	166	CT_2records_sameDateDx_diffTumSz	CoreTable_CASES_2maligs_sameDateDx_diffTumSz
2 in situ records, same DateDx, different TumSzs	1	CT_2records_sameDateDx_diffTumSz	CoreTable_CASES_2insitus_sameDateDx_diffTumSz
Total	733		

For these 8 separate conditions, MN manually reviewed the records for the 733 patients. During record review, he looked for a single record for each patient that best characterized the disease. Some patients with similar dates of initial diagnosis had little to no information in the first record. Therefore, MN would choose the record that was not a "blank" although the date of diagnosis may have been a few weeks after the very first useless record. For patient with multiple diagnoses because they presented with multi-focal disease (and bilateral), MN chose the earliest record for the Core Table, and made a note in the "manual_revision_notes" field of the Diagnostic table. The key to the "manual_revision_notes" field is:

1 : Ambiguous records
2 : Bilateral disease
3 : Multi-focal disease
4 : Records from different hospitals

For records of different sized tumors on the same date of diagnosis, MN chose the record with the larger tumor and made a note for either multi-focal or bilateral disease in the "manual_revision_notes" field. For particularly ambiguous cases, MN made a note in "manual_revision_notes" that the case should be revisited. For patients who have records of two different tumors with dates of diagnosis less than 60 days apart, MN made a note in the "manual_revision_notes" field. For cases with two diagnoses with dates of diagnosis less than 60 days apart at different facilities, MN chose the latter record and noted

[18] Note that this query does not check to see that patients with more than 1 in situ do not also have malignant records. When manually reviewed, 1 patient was found to have 3 in situs and 1 malignant. The malignant record was chosen for the Core Table.

[19] 122 of these cases were selected using a query after MN realized that many records were exact duplicates for each hospital; 1 for MGH and 1 for BWH. (query: CoreTable_CASES_2maligs_sameDateDxTumSzLNInvLNex_diffHosp). The case from the BWH was selected for the Core Table, and the field "manual_revision_notes" was updated to "4," indicating information about the same tumor between the two facilities.

[20] 111 of these cases were selected using a query after MN realized that many records contained clear indications of bilateral cancer: when "Hospital" was the same for each record while "Laterality" was different. (query: CoreTable_CASES_2maligs_sameDateDx_diffTumSz_bilat) Additionally, 23 other cases (of the 166 total) were selected using a query after MN realized that some were clear cases of multi-focal disease: when "Hospital" was the same for each record and "Laterality" was the same – but with different tumor sizes, as the parent query captured. (query: CoreTable_CASES_2maligs_sameDateDx_diffTumSz_multFoci)

the Hospital duplicate in "manual_revision_notes," as the patient likely continued her treatment at the second facility. For patients with 2 records on the same date of diagnosis, at different facilities, with different tumor sizes, MN chose the larger tumor size for the Core Table (if tumor size difference <=5mm, updated "manual_revision_notes" to 4 for 'seen at different hospitals,' or if size difference >5mm, updated notes to "1," indicating ambiguous case). During the manual review process, MN periodically ran the query CoreTable_CASES_checkDups to ensure no more than 1 record per patient was selected for the Core Table.

At this point in the selection process, there were a total of 24,035+733=24,768 patients for whom we had selected a single record for the Core Table, meaning that 24,791-24,768=23 patients had yet to be selected. Using the query "CoreTable_CASES_excludedPatiendIDs," MN manually reviewed each outlying record to determine the best one for inclusion in the Core Table based on completeness of data, ticking the field "CT_misfits" and noting ambiguity of records when necessary.[21]

Table	Total Records	Unique Patients
Patient	24,791	24,791*
Diagnostic	27,104	24,791
Treatment	71,423	24,505
FollowUp	112,924	23,385
CollStage	4,820	4,526

* Total unique patients in Tumor Registry database

[21] Many of these patients had 3 records in the Diagnostic table, 2 in situ and 1 malignant, which the previous queries did not catch.