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

SEER Breast Cancer Database, 2007

Creation and Characterization

Matthew Nolan
Research Assistant

James S. Michaelson, PhD.
Principal Investigator

4/4/2008

Updated 5/12/08

This document describes the methods used in creating the SEER Breast Cancer Database at the MGH Center for Quantitative Medicine. It also provides a characterization of the data, giving important statistics on frequencies and trends within the population.
Source of this document: SEER_BreastDB07_report_creation-characterization.docx

Table of Contents

Introduction

This database is a rebuilding of the previous SEER Breast Cancer Database, performed in MS Access. The previous version contained data only until 2002, but this version contains data through 2004. This rebuild was necessary because the older version was two years out of date, and accurate, complete follow-up information is crucial for our survival analyses of breast cancer. Moreover, certain markers like ER and PR status did not start to be collected until 1990, so those data are just beginning to mature for our 15-year survival time point.

The Surveillance, Epidemiology, and End Results (SEER) program is a national registry for cancers that is commissioned by the National Cancer Institute (NCI). In 1973, SEER began to maintain records of patients with all types of cancer from 9 select populations around the country. Over the last 30+ years, SEER has added more populations to the SEER "watch-list," and there are now millions of reported cancer incidences in the database, making SEER the most extensive and definitive registry of cancers in the United States. Its vast size provides incredible statistical power, which is vitally important for meaningful survival analysis. In SEER, the resolution of the data is impressive, and one can calculate survival rates for very narrow sub-populations of patients. Certain survival analyses, like hazard rates, are especially demanding because they essentially represent the derivative of the survival rate –the Kaplan-Meier survival plot, although it is a step function, must approximate a smooth curve in order to compute the corresponding hazard function. Only very large datasets behave in this manner, and thus SEER is very useful.

Because SEER freely distributes their data for medical and demographic research purposes, all patient data is de-identified. Thus, while the data is typically complete, we are unable to cross-check such factors as vital status and follow-up time with resources like the Social Security Master Death Index. In order to use the SEER data, we must make our first assumption; that SEER maintains high standards in their record-keeping methodology. However, a cursory look at the data will show that most fields contain high percentages of significant data (i.e. not "unknown"), which makes us believe that the cancer registrars are meticulous and probably accurate.

For more information about the SEER program, visit their website at http://seer.cancer.gov/ .

This report provides the methods and details for the creation of the SEER Breast Cancer Database at the Center for Quantitative Medicine of the Massachusetts General Hospital.

The database can be found at D:\+WORK+\SEER_9-07\SEER_BrCaDB_73-04_MASTER.accdb

Ways to Access the SEER Data

SEER makes their cancer data available in several different ways, each having their own merit.

Fast Stats – The SEER website includes an online applet for users to generate their own data summary tables for cancers in population subsets. It is a simple way to obtain desired information without having to download or sort the data on one's own computer. http://seer.cancer.gov/faststats/ (also see http://seer.cancer.gov/canques/)

SEER*Stat – This program may be downloaded from the SEER website (http://seer.cancer.gov/seerstat/software/). The program allows a user to login to the SEER database servers and request data through a streamlined GUI. To obtain access to the system, the user must have an approved data-request form, found here: http://seer.cancer.gov/data/request.html. The program automatically decodes the data, which makes all of it immediately available for analysis. SEER*Stat can export the requested data in the form of text files, but these files are massive and inefficient because they contain the full-length, decoded field values (instead of encoded numeric values). Another disadvantage is that some fields of the data have no documentation, so using the data from these fields is not advised.

CD Images – SEER also offers to send an approved user a set of CDs that contains the SEER*Stat program, and the data-source text files for all cancers, to be analyzed on one's own software. The advantage of this method is that the source text files contain encoded, width-delimited, numeric data, which keeps the file size as small as possible. One disadvantage is that the CD data do not contain some demographic data (but only those fields for which there is no documentation), and the encoded values necessitate tedious transformation in some cases.

The reason that database size is an issue to our group is that MS Access files (.mdb, .accdb) have a maximum size of 2 GB. If we wanted to include all breast cancer data in a single Access database (instead of using linked tables, which become useless when the directory structure of the DB changes), then the verbose, redundant data from SEER*Stat is far too large for Access to handle in a single file.

Our Methods for Obtaining the Data

Because the data exists in several forms, and because we wanted to be thorough in obtaining all the possible breast cancer data that SEER offers, we acquired two versions of the SEER data: one from SEER*Stat, and one from the CDs. While the datasets contain practically the same number of records, the SEER*Stat data contains extra fields on population demographics, but these fields do not have documentation. Note that, instead of actually getting a hard copy of the CDs from SEER, we downloaded the CD images.

One would expect to receive the same data from the online database as is available on the CDs, but this is not the case. If one selects all available variables for review in the SEER*stat online program, there are 260 fields (compared to 115 fields on the CD). There are, however, about 70 fewer patients in the SEER*stat system than on the CD (799999 on CD, 799929 online). During the last update of our SEER database, YM used the SEER*stat program to obtain data. Her set included 257 fields and 717,810 records. We are hesitant to use this online data because there are no detailed explanations of the additional fields not included on the CD (as we find in the dictionary seerdic.pdf). Nonetheless, for the sake of completeness, we also acquired this data.

Data from the SEER*Stat online program

*Fetching the Data from the SEER*Stat online program*

Opened SEER*stat, and did 'File->Client server login.'

Clicked on the table icon to perform a case-listing session. Used "Incidence - SEER 17 Regs Limited-Use, Nov 2006 Sub (1973-2004 varying)" database.[1]

For "Selection" tab, we did not want JUST Malignant behavior, so we deselected that option. All patients must, however, have a known age for our analyses, so we left that option checked.

In the same tab, clicked "Edit" in "Selection statement," chose "Site+Morphology->Site rec B with Kaposi and mesothelioma" variable (from left pane) and "Breast" (from right pane), such that the selection statement read " {Site and Morphology.Site rec B with Kaposi and mesothelioma} = ' Breast' " MN determined that this brief population description appeared to represent the most up-to-date and comprehensive group of patients with breast cancer.

In the "Table" tab we added the "Patient ID" and "Record Numer" variables to "Column" for every case-listing session to ensure unique identities of records. For the first case-listing

[1] Note that these are linked databases, and the reason for the additional fields available online is probably the county record information, which is not included on the CDs.

session, the variables located in the folders "Age at Diagnosis" through "Therapy" were added to "Column."

In the "Output" tab we labeled our output matrix "SEER_BrCaDB_1." We saved the case-listing as "PATH:\case-listings\SEER_BrCaDB_caselisting1.sl."

The case-listing was executed by clicking on the lighting bolt icon, and a matrix containing our data was produced. The SEER*stat matrix was saved as "PATH:\case-listings\SEER_BrCaDB_1.slm." We chose "Matrix->Export" with the following options: data file was "PATH:\case-listings\SEER_BrCaDB_1.txt," dic file was "PATH:\case-listings\SEER_BrCaDB_1.dic," selected "Output Variable Names Before Data" checkbox, selected "Labels Without Quotes" radio button,[2] and selected "Tab" as field delimiter. There are 63 total fields in this data query.

The previous steps were completed in a similar manner for the other case-listing sessions.

Case-listing session 2 contained all variables in folders "Extent of Disease" through "Other,"[3] which contained a total of 85 variables.

Case-listing session 3 contained all variables in folders "County Attributes" through "County Attributes 2000s," which contained a total of 86 variables.

Case-listing session 4 contained all variables in folders "County Attributes 1990s" through "County Attributes 1970s," which contained a total of 32 variables.

*Some Issues with the SEER*Stat data*

When considering how to best obtain the online SEER data, it became apparent that if one were to perform a case-listing session with *all* available fields, the file would be too large to import into Access. So, as YM had done, we needed to create multiple case-listing sessions in order to obtain all the available fields. However, it was not certain that each case-listing session would return the additional fields *in the same record order*. In order to solve this problem, MN searched for a field that uniquely identified each case, which could be used to link the various field tables to the same patient. It was determined that there are two important fields for this sort of analysis: "Patient ID" and "Record Number." "Patient ID" contains a unique ID for each patient in the SEER registry, while "Record Number" contains a number (01-07) that tracks multiple records for a particular patient. That is, one patient can have multiple records in the SEER registry (possibly for different tumors, recurrences). **A quick analysis revealed that over**

[2] This option actually doesn't work in SEERstat 6.3.5. All field names are contained within quotation marks, and there is no way to prevent this.

[3] The only fields excluded were some of the "Site specific Sequence Numbers." We chose the fields which contained "most detail," and did NOT select the fields with "mid" and "least" detail. MN believed this information would be redundant and unnecessary.

10% of the database is comprised of multiple records for the same patient (which calls into question that accuracy of previous calculations performed on the SEER data). We will discuss this problem more thoroughly during outcome analysis, but for now we will simply use the Patient ID and Record Number to ensure that each case listing record corresponds to the other case listing records.

Importing into MS Access

When importing the SEER*stat txt exports, you must use the "delimited" option. We chose to tab-delimit our fields, and the "First row contains field names." Let Access create a primary key for your tables. Set the "Data Type" for all fields equal to "Text" to avoid importation errors.

Each database, named sequentially "PATH:\SEER_BrCaDB_73-04_online_#," contains one table, named "SEER_BrCaDB_#."

MN inspected each of the 4 databases the ensure that the primary ID that Access had assigned to the cases corresponded to the same patient in each database, which it did.

Compiling the 4 datasets

We created a new database, named "PATH:\ SEER_BrCaDB_73-04_online_MASTER.mdb." Within this database, we added table links to the 4 other online-data databases we had just created.

Data from the CDs

Obtaining the data from the CDs

Requested SEER data via http://seer.cancer.gov/data/request.html. Data is registered to James Michaelson, PhD. Username: 11401-Nov2006 ; Password: QPX41G

Downloaded the "CD2" image in late August, 2007.

There are 3 breast cancer databases in the SEER dataset (all named "BREAST.TXT," but in their respective folders), specified as follows:

YR1973_2004.SEER9

> This directory contains the SEER November 2006 Limited-Use Data files from nine SEER registries for 1973-2004. The SEER 9 registries are Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco-Oakland, Seattle-Puget Sound, and Utah. Data are available for cases diagnosed from 1973 and later for these registries with the exception of Seattle-Puget Sound and Atlanta. The Seattle-Puget Sound and Atlanta registries joined the SEER program in 1974 and 1975, respectively.

YR1992_2004.SJ_LA_RG_AK

> This directory contains the SEER November 2006 Limited-Use Data files from the San Jose-Monterey, Los Angeles, Rural Georgia and Alaska Natives SEER registries for 1992-2004.

YR2000_2004.CA_KY_LO_NJ

> This directory contains the SEER November 2006 Limited-Use Data files from the Greater California, Kentucky, Louisiana, and New Jersey SEER registries for 2000-2004."[4]

[4] Source: "PATH:\Readme.txt"

Importing into MS Access

Consulted "Readme.txt" for info on the populations and "seerdic.pdf" for info on the field names/lengths. Created spreadsheet "SEER_field_names.xls" in "PATH:\" which specifies the fields for each element in the database.

When importing the DB txt file, you must use the "fixed-width" option. However, due to a bug in the program, the "specification" wizard in the "advanced" section will not properly save your updated field listings if you have loaded previously saved field listings, so *you must enter all the field start points and width information in one sitting*. You can copy and paste one column at a time from excel into the specifications window, but the specs window must contain enough pseudo-blank entries for the number of fields you want to paste (you must create these blank field entries manually). Be sure that all fields are of type "Text" to avoid importation errors. Let Access create a primary key for your tables.

We now have 3 tables in "PATH:\SEER_BrCaDB_73-04_RAW_CD2.mdb" named 'BreastDB_00-04_CA-KY-LO-NJ' (146,303 records), 'BreastDB_73-04_SEER9' (529,263 records), and 'BreastDB_92-04_SJ-LA-RG-AK' (104,433 records).

Compiling the 3 datasets

We then wanted to compile these three populations into one master population. First, we copied and pasted the SEER9 population table into a new table, named 'SEER_MASTER_BrCaDB_RAW'. Then, we opened a new query named 'Master table append,' of type "Append query." Show tables 'SEER_MASTER...' and 'BreastDB_00-04…' In 'BreastDB_00-04…,' ctrl+shift select fields 'Patient_ID' through 'Vital_status_recode,' making sure not to select 'ID,' since appending the primary key will cause duplicates and the query will fail. Drag and drop selected fields into 'Field' and run the query (the "!" button). Do the same process for the 3rd and final table.

The table 'SEER_MASTER_BrCaDB_RAW' now contains the records from all the breast cancer patients in the SEER dataset. **There are 779,999 records in the raw master table**.

** Be sure to check 'File->Database properties->General' frequently to monitor the size of your database. An Access mdb/accdb file cannot exceed 2GB. If you get close, run 'Tools->Database utilities->Compact and repair database.' This process can be slow, but it drastically reduces the database file size.

Note that we will use the CD data as our main SEER Breast Cancer Database, and all transformations to the data fields described in this report were performed on the dataset from the CDs. In addition to the convenience of having the whole database stored in one file (impossible for the SEER*Stat data due to verbose, decoded values), we are more confident in this data because we have the source files of the database, and we will know exactly how the data was modified to make it accessible.

Location of the SEER Breast Cancer DB

The SEER master BrCaDB raw table from PATH:\SEER_BrCaDB_73-04_RAW_CD2.mdb (table: SEER_MASTER_BrCaDB_RAW) was copied into a new database so as to keep the original copy untouched. We made all transformations and field addendums to the table SEER_BrCaDB_MASTER in the new database.

File of the SEER Master Breast Cancer Database:

PATH:\SEER_BrCaDB_73-04_MASTER.accdb

(original file name SEER_BrCaDB_73-04_MASTER.mdb, upgraded to accdb format)

Structure of the SEER Breast Cancer Database

The SEER data obtained from the CD2 image contained encoded values for each field (e.g. values of 1 or 2 describing male or female, instead of the actual string "male" or "female"). This method of data archiving *drastically* reduces the size of the DB, but it renders the data incomprehensible for immediate analysis. In order to make the data useful, we created many foreign key lookup tables, which decode the numeric values into their actual meaning. We did not perform this method on every field, but only on those fields which we deemed most relevant to our studies. However, also note that not all fields needed to be decoded, as their numeric value is significant in itself, such as tumor size. The following table lists all of the fields in the database, a flag if they are available for analysis (have a definition table/intrinsic meaning), remarks about the meaning of the field, indications for the amount and quality of the data, and any other pertinent comments about the sources of the fields. We adapted this table from the following coding manuals (found in "PATH:\manuals\"):

Seerdic.pdf from the CD data at http://seer.cancer.gov/manuals/CD2.SEERDic.pdf

ICD-O-3 manual (icdo3.sitetype.d08152007 .pdf and .xls) from http://seer.cancer.gov/icd-o-3/

ICD-0-2 to ICD-O-3 conversion manual from http://seer.cancer.gov/tools/conversion/ICDO2-3manual.pdf

EOD 10-digit manual (1988+) from http://seer.cancer.gov/manuals/EOD10Dig.pub.pdf

EOD 4-digit manual (1983-1987) from http://seer.cancer.gov/manuals/historic/EOD_1984.pdf

EOD 13-digit manual + EOD 2-digit manual (1973-1982) from http://seer.cancer.gov/manuals/historic/EOD_1977.pdf

SEER Coding and Staging (CS) manuals SPM_2004_maindoc.r1.pdf, SPM_AppendixA.pdf, SPM_AppendixB_r1.pdf, SPM_AppendixC_Part3_r1.pdf, and SPM_AppendixD_r1.pdf from http://seer.cancer.gov/tools/codingmanuals/historical.html

Note that all update queries used in creating the new consensus data fields have been saved in the database. Every transformation has been catalogued according to the field it modified and/or a description of the modification. To see these queries and transformations, open the queries section of the SEER Master BrCaDB, located in PATH:\SEER_BrCaDB_73-04_MASTER.accdb

Fields in the SEER Breast Cancer Database

Available	Field Name	Data Description	% cases w/ Data; Data Notes	Remarks
●	**Patient_ID_number**	unique for each patient in SEER registry (see "Record_number")	100%; detailed	
●	**Registry_ID**	defines the SEER population of patient (NJ, Puget Sound, etc.)	100%; detailed	
	Marital_Status_at_DX	single, married, separated, divorced, etc.	100%; detailed	
	Race_Ethnicity	over 30 different ethnicity values	100%; detailed	
	Spanish_Hispanic_Origin	even includes info if surname is hispanic	100%; detailed	
	NHIA_Derived_Hispanic_Origin	uses fancy algorithm for determining hispanic ethnicity	100%; detailed	
●	**Sex**	male or female	100%; detailed	
●	**Age_at_diagnosis**		100%; detailed	
●	**Year_of_Birth**		100%; detailed	
	Birth_Place	consulted SPM_AppendixB_r1.pdf (SEER geocodes)	100%; ~30% "unknown"	
●	**Sequence_Number--Central**	whether a case represents a single primary, 2nd of 3 primaries, etc.	100%; detailed	
●	**Month_of_diagnosis**		100%; detailed	
●	**Year_of_diagnosis**		100%; detailed	
●	**Primary_Site**	defined in seerdic.pdf as ICD-O-3 topography codes, but MN used SEER Coding+Staging Manual SPM_AppendixC_Part3_r1.pdf	100%; ~20% "breast, NOS"	
●	**Laterality**	left, right, or bilateral	100%; detailed	

	Histology_92-00_ICD-O-2	earlier/later classifications converted, some years may be inaccurrate	100%; detailed	
	Behavior_92-00_ICD-O-2	earlier/later classifications converted	100%; detailed	
•	**Histologic_Type_ICD-O-3**	earlier classifications converted, ICD-O-2 inaccurracies may affect this classification	100%; detailed	
	Behavior_Code_ICD-O-3	defines in situ/malignant; earlier classifications converted; identical to Behavior_recode_for_analysis	100%; detailed	
•	**Grade**	grade 1-3, 4=undifferentiated	100%; ~35% "unknown"	
	Diagnostic_Confirmation	describes method for confirmation of carcinoma	100%; detailed	
	Type_of_Reporting_Source	hospital, lab, autopsy, etc.	100%; detailed	
•	**EOD—Tumor_Size**	in mm (1988+); 001 means microscopic, 002 means ≤ 2mm; consult EOD10Dig.pub.pdf	~80%; ~10% "unknown"	
	EOD—Extension	farthest tumor extension (1988+); consult EOD10Dig.pub.pdf	~80%; detailed	
	EOD—Extension_Prost_Path	only applicable to prostate cancer	0%;	
	EOD—Lymph_Node_Involv	highest lymph node chain involved (1988+)	~80%; ~5% "unknown"	
•	**Regional_Nodes_Positive**	# pos nodes (1988+); 97 = node pos., 98 = none examined	~80%; ~25% unknown/nonspecific	
•	**Regional_Nodes_Examined**	# nodes examined (1988+); 96-98 = some nodes examined	~80%; ~3% unknown/nonspecific	
	EOD—Old_13_Digit	old EOD codes	(not investigated)	
	EOD—Old_2_Digit	old EOD codes	(not investigated)	

	EOD—Old_4_Digit	old EOD codes	(not investigated)	
	Coding_System_for_EOD	how SEER determined current EOD info	(not investigated)	
•	**Tumor_Marker_1**	ER+/- (1990-2003)	~92%; ~44% definitive +/-	
•	**Tumor_Marker_2**	PR+/- (1990-2003)	~92%; ~43% definitive +/-	
	Tumor_Marker_3	only applicable to testis	(not investigated)	
•	**CS_Tumor_Size**	discrete size in mm (2004+)	~8%; ~7% detailed	for Collaborative Staging info, consult SPM_AppendixC_Part3_r1.pdf
	CS_Extension	tumor extension (2004+)	~8%; detailed	website: http://www.cancerstaging.org/cstage/index.html
	CS_Lymph_Nodes	type of lymph nodes involvement (2004+)	~8%; detailed	
	CS_Mets_at_Dx	location of mets at Dx (2004+)	~8%; detailed	
	CS_Site-Specific_Factor_1	ER+/- (2004+)	~8%; ~6% definitive +/-	
	CS_Site-Specific_Factor_2	PR+/- (2004+)	~8%; ~6% definitive +/-	
	CS_Site-Specific_Factor_3	# pos nodes, ipsilateral axillary (2004+)	~8%; ~6% definitive	
	CS_Site-Specific_Factor_4	IHC of regional lymph nodes (2004+)	~8%; ~1.5% definitive	
	CS_Site-Specific_Factor_5	Molecular studies of reg. lymph nodes (2004+)	~8%; ~0.3% definitive	
	CS_Site-Specific_Factor_6	Degree of tumor invasiveness (2004+)	~8%; ~7%	

			definitive	
	Derived_AJCC_T	AJCC T component, derived from CS fields (2004+)	~8%; (not investigated)	
	Derived_AJCC_N	AJCC N component, derived from CS fields (2004+)	~8%; (not investigated)	
	Derived_AJCC_M	AJCC M component, derived from CS fields (2004+)	~8%; (not investigated)	
	Derived_AJCC_Stage_Group	AJCC Stage Group component, derived from CS fields (2004+)	~8%; (not investigated)	
	Derived_SS1977	SEER Summary Stage 1977, derived from CS fields (2004+)	~8%; (not investigated)	
	Derived_SS2000	SEER Summary Stage 2000, derived from CS fields (2004+)	~8%; (not investigated)	
	Derived_AJCC—Flag	indicates whether AJCC stage was derived from CS or EOD fields (2004+)	~8%; (not investigated)	
	Derived_SS1977—Flag	indicates whether SEER Summary Stage 1977 was derived from CS or EOD fields (2004+)	~8%; (not investigated)	
	Derived_SS2000—Flag	indicates whether SEER Summary Stage 2000 was derived from CS or EOD fields (2004+)	~8%; (not investigated)	
	CS_Version_1st	indicates version number initially used to code CS fields (2004+)	~8%; (not investigated)	
	CS_Version_Latest	indicates most recent version used to code CS fields (2004+)	~8%; (not investigated)	
	RX_Summ—Surg_Prim_Site	describes any surgical procedure performed on primary site (1998+)	~48%; detailed	for RX_Summ info, consult SPM_AppendixC_Part3_r1.pdf
	RX_Summ—Scope_Reg_LN_Sur	describes any biopsy, removal, aspiration of regional lymph nodes (2003+)	~15%; detailed	

	RX_Summ—Surg_Oth_Reg_Dis	describes any other surgeries to distant lymph nodes/tissues/organs (2003+)	~15%; only 0.2% have significant data	
	RX_Summ—Reg_LN_Examined	# regional lymph nodes examined (1998-2002)	~35%; ~1% "unknown"	
	RX_Summ—Reconstruct_1st	describes type of reconstruction for breast cancer surgery patients (1998-2002)	~30%; ~4% "unknown"	
•	**Reason_for_no_surgery**		100%; ~1% "unknown"	
•	**RX_Summ—Radiation**	method of radiation therapy	100%; ~3% "unknown"	
	RX_Summ—Rad_to_CNS	only for lung and leukemia	(not investigated)	
•	**RX_Summ—Surg___Rad_Seq**	order of treatments for radiation and surgery	100%; ~1% "unknown"	
	RX_Summ—Surgery_Type	(1983-1997); consult page D-17 of AppendD.pdf in seer.cancer.gov/manuals/historic/	~50%; ~10% unspecific	
	RX_Summ—Surg_Site_98-02	describes any surgical procedure performed on primary site (1998-2002); consult page C-63 of AppendC.pdf in seer.cancer.gov/manuals/historic/	~30%; ~1% unspecific	
	RX_Summ—Scope_Reg_98-02	describes any biopsy, removal, aspiration of regional lymph nodes (1998-2002); consult page C-63 of AppendC.pdf in seer.cancer.gov/manuals/historic/	~30%; detailed	
	RX_Summ—Surg_Oth_98-02	describes any other surgeries to distant lymph nodes/tissues/organs (1998-2002); consult page C-63 of AppendC.pdf in seer.cancer.gov/manuals/historic/	~30%; detailed	
•	**SEER_Record_Number**	patients can have multiple SEER records, beginning with record number 01	100%; detailed	
	Over-ride_age_site_morph	reviewed: certain cases with spurious age/site/morph data	0%;	
	Over-ride_seqno_dxconf	reviewed: certain cases with multiple primaries w/out microscopic confirmation	~0.5%;	

	Over-ride_site_lat_seqno	reviewed: certain cases with multiple primaries w/ same histology in same site	~0.5%;	
	Over-ride_surg_dxconf	reviewed: certain cases with reported cancer surgery but unable to analyze tissue removed	0%;	
	Over-ride_site_type	reviewed: spurious site/histological combinations	~0.2%;	
	Over-ride_histology	reviewed: spurious histology/behavior/confirmation combinations	~0.1%;	
	Over-ride_report_source	reviewed: certain cases with reported second primaries were verified as independent primaries	~0.01%;	
	Over-ride_ill-define_site	reviewed: certain cases with second primaries with ill-defined first primaries were verified as second primaries	~0.01%;	
	Over-ride_Leuk,_Lymph	only for leukemia and lymphoma	(not investigated)	
	Over-ride_site_behavior	reviewed: certain cases of in situ cancer w/ little information about primary site	~0.01%;	
	Over-ride_site_eod_dx_dt	reviewed: certain cases with "localized" disease w/ little information about primary site	~0.01%;	
	Over-ride_site_lat_eod	reviewed: certain cases with spurious laterality coding	~0.01%;	
	Over-ride_site_lat_morph	reviewed: certain cases with other spurious laterality codings	~0.01%;	
	ICD-O-2_conversion_flag	defines method for ICD-O interconversions	100%; detailed	
	SEER_Type_of_Follow-up	expected follow-up type (autopsy, active..)	100%; detailed	
	ICD-O-3_conversion_flag	defines method for ICD-O interconversions	~95%; detailed	
	Age_Recode_<1_Year_olds	5-year age bins	100%; detailed	
	Site_Recode	all cases have same value (2600), which indicates breast	100%; detailed	consult seer.cancer.gov/siterecode/icd_01272003/

	Site_Re_with_Kaposi_and_Mesothelioma	all cases have same value (2600), which indicates breast	100%; detailed	consult seer.cancer.gov/siterecode/icd_01272003/
	Recode_ICD-O-2_to_9	conversion of primary site and morphology codes	100%; detailed	
	Recode_ICD-O-2_to_10	conversion of primary site and morphology codes	100%; detailed	
	ICCC_site_recode_ICD-O-2	ICCC code based on primary site and ICD-O-2	100%; ~15% "unknown"	consult seer.cancer.gov/iccc/iarciccc.html
	SEER_modified_ICCC_site_recode_ICD-O-2	modified: ICCC code based on primary site and ICD-O-2	100%; ~15% "unknown"	
	ICCC_site_recode_ICD-O-3	ICCC code based on primary site and ICD-O-3	100%; ~15% "unknown"	
	ICCC_site_recode_extended_ICD-O-3	modified: ICCC code based on primary site and ICD-O-3	100%; ~15% "unknown"	
•	**Behavior_Recode_for_Analysis**	defines tumor as in situ or malignant (identical to Behavior_Code_ICD-O-3)	100%; definitive	
	ICD-O_Coding_Scheme	indicates original coding (ICD-O 2 or 3) of case	100%; detailed	
	Histology_Recode—Broad_Groupings	squamous, epithelial, basal, granular, etc.	100%; ~95% detailed	
	Histology_Recode—Brain_Groupings	not applicable	(not investigated)	
	CS_Schema	all cases have same value (58), which indicates breast	(not investigated)	
	Race_recode_White,_Black,_Other		100%; ~95% detailed	consult http://seer.cancer.gov/seerstat/variables/seer/yr1973_2004/race_ethnicity/
	Race_recode_W,_B,_AI,_API	breaks down "other" field from Race_recode_White,_Black,_Other	100%; ~95% detailed	
	Origin_recode_NHIA_Hispanic,_N	either hispanic or non-hispanic origin	100%; detailed	

	on-Hisp			
	SEER_historic_stage_A	multiple collapsed stage fields, use caution!	100%; ~95% detailed	
	AJCC_stage_3rd_edition_1988-2003	(1988-2003)	~70%; ~60% detailed	
	SEER_modified_AJCC_Stage_3rd_ed_1988-2003	few details on this field (1988-2003); appears to be an updated version of AJCC_stage_3rd_edition	~70%; ~60% detailed	
	SEER_Summary_Stage_1977	(1995-2000)	~30%; ~1% "unknown"	
	SEER_Summary_Stage_2000	(2001-2003)	~25%; ~1% "unknown"	
•	**Number_of_primaries**	accurate value; based on ALL SEER records, not just this database; all records from one patient reflect same value	100%; detailed	
	First_malignant_primary_indicator	based on all tumors in SEER; outside tumors assumed malignant	100%; detailed	
	State-county_recode	uses FIPS codes for state and county in order: SSCCC	100%; detailed	
	Survival_time_recode	survival for completed years and months: YYMM	100%; ~2% "0000"	"9999" should represent "unknown," but there are none. Consider "0000"="unknown"
	Cause_of_Death_to_SEER_site_recode	indicates death from cancer or non-cancer	100%; ~2% "unknown"	site: seer.cancer.gov/codrecode/1969+_d09172004/index.html
•	**COD_to_site_rec_KM**	indicates death from cancer or non-cancer (replica of Cause_of_Death_to_SEER…)	100%; ~2% "unknown"	site: seer.cancer.gov/codrecode/1969+_d09172004/index.html
	Vital_Status_recode	alive or dead as of 12/31/2004	100%; definitive	
•	**Date_of_Dx_consensus**	added by MN 9-14-07; collapsed Month of Dx and Year of Dx	100%; definitive	

	EOD13_NumPosNodes	added by MN 9-13-07; parsed EOD 13-digit value; key from EOD_1977.pdf	10%; 2% detailed	
	EOD13_NumPosNodesExamined	added by MN 9-13-07; parsed EOD 13-digit value; key from EOD_1977.pdf	10%; 2% detailed	
	EOD4_TumSize	added by MN 9-13-07; parsed EOD 4-digit value; key from EOD_1984.pdf	9%; 6.5% detailed	
	EOD13_TumSize_clin	added by MN 9-13-07; parsed EOD 13-digit value; key from EOD_1977.pdf	11%; 6.5% detailed	
	EOD13_TumSize_oper-path	added by MN 9-13-07; parsed EOD 13-digit value; key from EOD_1977.pdf	11%; 6.5% detailed	
	EOD13_TumSize_collapsed	added by MN 9-17-07; collapsed EOD 13-digit clin and oper-path tumor sizes to oper-path value, unless oper-path = 999 when clin <> 999	11%; 9% detailed	
	EOD2_TumSize_bin1	added by MN 9-14-07; parsed EOD 2-digit for values 2x,3x,5x,6x; key from EOD_1977.pdf; consolidated unknowns 0 and 9 to 999	1%; 0.5% detailed	
	EOD2_TumSize_bin2	added by MN 9-14-07; parsed EOD 2-digit for values 7x,8x; key from EOD_1977.pdf; consolidated unknowns	0.5%; 0.001% detailed	
	EOD2_TumSize_bin3	added by MN 9-14-07; parsed EOD 2-digit for values <>2x,3x,5x,6x,7x,8x; recoded to 999 for unknown	0.8%; 100% "unknown"	
	EOD2_TumSize_collapsed	added by MN 9-18-07; collapsed EOD2_TumSize bins 1-3	2%; 0.5% detailed	
	CS_TumSize_consensus	added by MN 9-18-07; consensus recode of "CS-Tumor_Size" field; see TumorSize_coding_table.xls for details	8%; 6% detailed	
	EOD10_TumSize_consensus	added by MN 9-18-07; consensus recode of "EOD-Tumor_Size" field; see TumorSize_coding_table.xls for details	7%; 5% detailed	
	EOD4_TumSize_consensus	added by MN 9-18-07; consensus recode of "EOD4_TumSize" field; see TumorSize_coding_table.xls for details	9%; 7% detailed	
	EOD13_TumSize_consensus	added by MN 9-18-07; consensus recode of "EOD13_TumSize_collapsed" field; see TumorSize_coding_table.xls for details	11%; 9% detailed	used approximated tumor sizes

	EOD2_TumSize_consensus	added by MN 9-18-07; consensus recode of "EOD2_TumSize_collapsed" field; see TumorSize_coding_table.xls for details	2%; 0.5% detailed	used approximated tumor sizes
•	**TumSize_consensus**	added by MN 9-22-07; MASTER tumor size field; collapsed all TumSize_consensus fields;	100%; definitive	
•	**Survival_time_consensus**	added by MN 9-24-07; recoded Survival_time_recode into a decimal year format	100%; definitive	
•	**COD_BrCa_consensus**	added by MN 9-25-07; all cases of breast cancer deaths according to "COD_to_site_recode_KM" are flagged "1"	100%; definitive	
•	**ER_status_consensus**	added by MN 9-25-07; collapsed ER status data from Tumor_Marker_1 (1990-2003) and CS_Site-Specific_Factor_1 (2004); 0=negative, 1=positive, 2=borderline, 9=unknown/unspecified	73%; 51% definitive	
•	**PR_status_consensus**	added by MN 9-25-07; collapsed PR status data from Tumor_Marker_2 (1990-2003) and CS_Site-Specific_Factor_2 (2004); 0=negative, 1=positive, 2=borderline, 9=unknown/unspecified	73%; 50% definitive	
•	**NumPosNodes_consensus**	added by MN 9-26-07; collapsed data from Regional_Nodes_Positive and EOD13_NumPosNodes; if Node_pos-neg_consensus = 0, NumPosNodes_consensus = 0; upperbound groups were excluded	20%; 30% definitive	
•	**NumNodesExamined_consensus**	added by MN 9-26-07; collapsed data from Regional_Nodes_Examined and EOD13_NumNodesExamined; upperbound groups were excluded	61%; 61% definitive	
•	**Node_pos-neg_consensus**	added by MN 9-26-07; collapsed data from Regional_Nodes_Positive, EOD13_NumPosNodes, EOD-Old_2_Digit, and EOD-Old_4_Digit; includes "node positive" patients who have unknown exact number of positive nodes; 0=negative, 1=positive	74%; 74% definitive	

Source: "PATH:\SEER_field_names.xls"

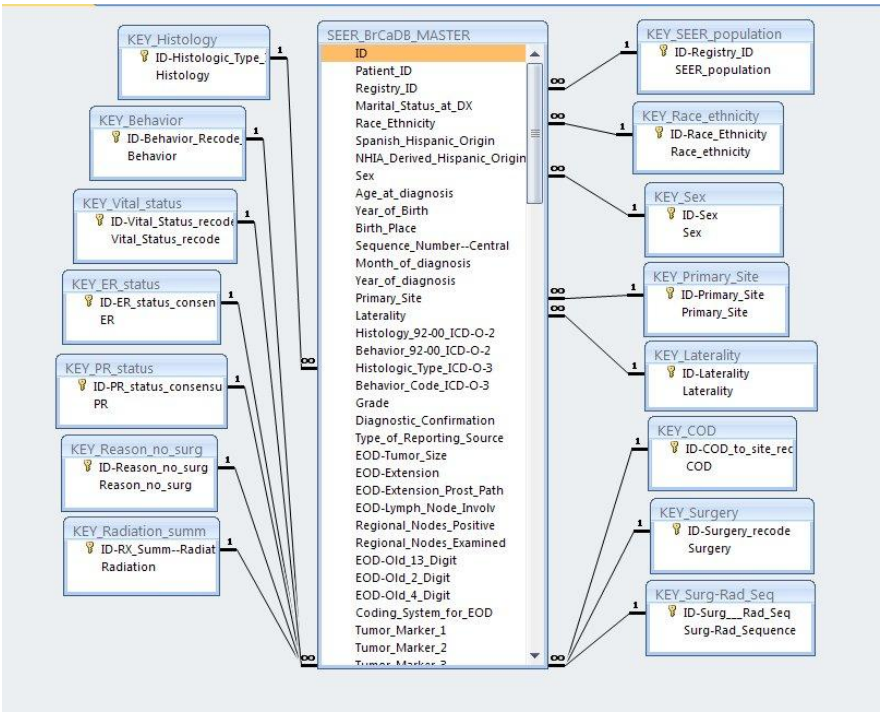
Relationships in the database

One of the most important aspects to our relational database is the use of foreign key lookup tables. To avoid storing the same string many times in the database, many fields use numeric values that correspond to a definition. This method of database management is essential for our database to be located in a single file, rather than having the database reference several linked tables. However, in order to make the data useful, we needed to create foreign key lookup tables, or 'definition' tables containing primary IDs, and then create a relationship between the definition tables and the encoded fields of the master table.

We have created several KEY tables which contain the numeric value IDs, and their corresponding definitions. In some cases, we have defined the numeric value for unknown (e.g. "999") as null. In doing so, we effectively simplify the analysis of the data by excluding the unknown values from the dataset. Note that this only affects select queries, and the original data remains untouched.

Once we had created key tables that correspond to the numeric values of fields from the master table, we needed to define the relationship between the two tables. To do this, we opened the "Relationships" dialog (the  button), right-clicked "show tables," and clicked-and-dragged the field from the master table onto the ID field of the definition table. Because the numeric data are meaningless without the foreign key lookups, it is good database practice to select the "enforce referential integrity" checkbox when creating the relationship. This item prevents modification of the foreign key field from one table without confirmation or equal modification of the definition table.

Once we have created these relationships, we can make use of the relational database structure when performing a select query. This procedure is explained in the discussion of the Core Table select query.



Consensus Tumor Size

We have an entry for tumor size data for every patient in the database, although many patients have the designation “unknown,” or its equivalent, as the value. These data are dispersed among various fields, and exist in various formats. Specifically, the fields we needed to consider were EOD-Tumor_Size, EOD-Old_13_Digit, EOD-Old_2_Digit, EOD-Old_4_Digit, and CS_Tumor_Size. Also note that the tumor size information for a given patient is found in only one of the five fields, which actually made collapsing the fields to create a consensus tumor size much easier and more accurate.

Table of fields containing tumor size information:

Schema	N malignant	years	lowerbound bin value	N lowerbound	N lower/ N malig	intermediate data (in mm)	upperbound bin value	N upperbound	N upper/ N malig
SEER Coding & Staging (CS)	49,364	2004	none (1mm)	n/a		continuous	>= 989mm	3	0.01%
EOD 10 digit	455,538	1988-2003	<= 3mm	5,789	1.27%	continuous	>= 990mm	9	0.00%
EOD 4 digit	66,600	1983-1987	<= 3mm	197	0.30%	continuous, 96-99mm	>= 100mm	730	1.10%
EOD 13 digit	84,413	1973-1982	<= 4mm	702	0.83%	5-9, 10-19, 20-29, 30-39, 40-49, 50-99	>= 100mm	1,326	1.57%
EOD 2 digit (2x,3x,5x,6x)	6,782	1973-1982	<= 10mm	484	7.14%	11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80	>= 81mm	39	0.58%
EOD 2 digit (7x,8x)	1,014	1973-1982	<= 20mm*	121	11.93%	21-40, 41-60	>= 61mm	41	4.04%
Total	663,711			7,293	1.10%			2,148	0.32%

Source: “PATH:\TumorSize_coding_table.xls”

While we have tumor size information for every patient in the database, some patients are unknown, or do not have continuous size data (bins). We want to have continuous, exact tumor size values because we cannot perform calculations or statistical analyses on size bins, or at least we cannot mix size bins values with discrete values when doing such calculations. Ideally, every patient would have an exact tumor size, and there would be a singular value to describe "unknowns." However, the tumor sizes for many patients are part of a size bin, and we needed to assign these patients an exact size based on the mean tumor size of patients within a size bin from a similar population, or to assign these patients the "unknown" classification. There were 3 cases for the binning problem: lowerbound bins (like “<=

3mm"), intermediate bins (like "40mm-49mm"), and upperbound bins (like "10mm+"). What further complicates matters is that the bins do not generally overlap (e.g. one coding scheme bin is 10-19mm, while another is 11-20mm). While it would be possible to assign each bin a discrete value, this would not be an accurate description of central tendency, because lowerbound and upperbound size bins contain much internal variance because they have no fixed limits.

Before assigning exact tumor sizes to the patients in the size bins, we agreed on a nomenclature for tumor size. **Our "consensus" nomenclature for tumor sizes is that 001-988 represents continuous size data in mm, and 999 indicates "Unknown/Unspecified."**

We decided to categorize all lowerbound bins as unknown, except <=10mm and <=20mm bins, because these groups contain large amounts of data for which, we believe, a derived measure of central tendency is actually descriptive.[5]

We decided to categorize all upperbound bins as unknown, because any measure of central tendency would have a very large amount of variance that would render the data useless.

We did derive measures of central tendency (mean) for all intermediate bins, except 96-99mm, because there are no similar populations from which we could derive a mean value. **All mean values were rounded to nearest mm.**

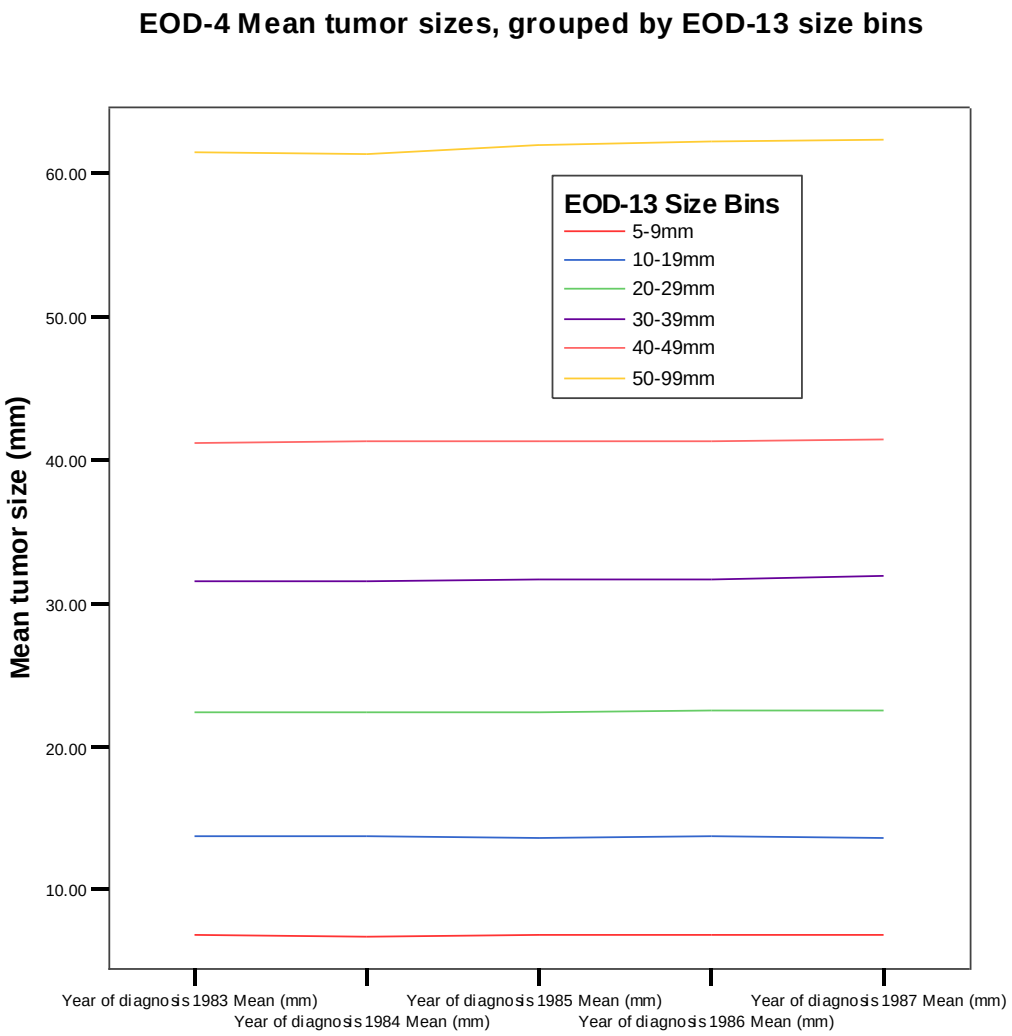
In order to identify which populations with continuous size data were most similar to the populations with binned size data, we considered the chronological proximity, and we then sought similarity in the percentage of patients in each size bin (our assumption was that if two populations contained approximately the same percentage of patients in each size bin, then it would be justifiable to use the mean tumor size from the continuous data to describe the dataset that contained only binned sizes). Our binned size data comes from populations between 1973 and 1982. The EOD 4-digit data begins in 1983 (ends in 1987), and thus it meets our chronological criterion. We then wanted to know which of the 4 years would best approximate the size bins from the earlier populations (contain a similar percentage of patients in each bin). If we see no appreciable change in tumor size over the time period, then we are confident in using the whole 5-year interval from 1983-1987 to estimate the mean tumor size of each size bin.

[5] JSM proposed that we keep these groups instead of ascribing them "unknown" value.

Mean EOD 4-digit tumor sizes by EOD 13-digit size bins:

EOD-4 Tumor Size by EOD-13 Size Bins, by year	Year of diagnosis									
	1978		1979		1980		1981		1982	
	Mean (mm)	std. dev.	Mean (mm)	std. dev.	Mean (mm)	std. dev.	Mean (mm)	std. dev.	Mean (mm)	std. dev.
5-9mm	6.88	1.59	6.79	1.64	6.91	1.55	6.99	1.56	6.87	1.53
10-19mm	13.81	2.65	13.82	2.70	13.65	2.71	13.83	2.66	13.70	2.67
20-29mm	22.43	2.66	22.47	2.64	22.51	2.66	22.53	2.67	22.58	2.64
30-39mm	31.62	2.43	31.65	2.45	31.69	2.43	31.76	2.48	31.92	2.56
40-49mm	41.20	2.17	41.36	2.25	41.40	2.29	41.40	2.30	41.45	2.33
50-99mm	61.52	12.23	61.45	12.06	62.04	12.54	62.32	12.25	62.44	12.53

Source: "PATH:\SPSS_analyses\CS_EOD10_EOD4_TumSize-malig_data.spo"



Source: "PATH:\SPSS_analyses\CS_EOD10_EOD4_TumSize-malig_data.spo"

Note that, for each size group, the mean value by year does not change appreciably.

Percentage of population in each EOD-13 size bin for EOD-4 and EOD-13:

	EOD4 binned to EOD13, comparison					
	5-9mm	10-19mm	20-29mm	30-39mm	40-49mm	50-99mm
	% of Pop.	% of Pop.	% of Pop.	% of Pop.	% of Pop.	% of Pop.
EOD4 1983-1987 (continuous)	7.0%	32.7%	27.9%	14.8%	7.4%	9.7%
EOD13 1972-1982 (bins)	4.0%	26.6%	28.6%	18.3%	9.6%	13.0%

Source: "PATH:\SPSS_analyses\CS_EOD10_EOD4_TumSize-malig_data.spo"

Because these percentages are fairly consistent with each other, we assume that the two EOD-4 populations may be used to describe the EOD-13 population. Additionally, because the EOD-2 population is further sub-divided in multiple, overlapping tumor size groups, we cannot perform a similar analysis of EOD-4 versus EOD-2 to estimate validity in assigning exact tumor sizes. The EOD-2 population is, however, essentially interspersed with the EOD-13 population throughout the period 1973-1982. The patients were coded using either EOD-13 or EOD-2, and there does not seem to be any trend for coding one of the other (i.e., nothing to do with population location, etc.). Therefore, because we cannot analyze the EOD-2 population against the EOD-4, we will assume that the interspersion of EOD-2 and EOD-13 means that the validity of assigning mean EOD-4 sizes to the EOD-13 is also extensible to the EOD-2 population. The overall results are as follows:

Mean sizes for EOD-13 Size Bins:

		Mean (mm)	Std. Dev.
EOD-4 (1983-1987) Mean Tumor Size by EOD-13 Size Bins	5-9mm	7	1.56
	10-19mm	14	2.68
	20-29mm	23	2.65
	30-39mm	32	2.47
	40-49mm	41	2.27
	50-99mm	62	12.32

Mean sizes for EOD-2 (1) Size Bins:

		Mean (mm)	Std. Dev.
EOD-4 (1983-1987) Mean Tumor Size by EOD-2 (1) Size Bins	<= 10mm	8	1.91
	11-20mm	17	2.93
	21-30mm	27	2.92
	31-40mm	37	2.86
	41-50mm	48	2.71
	51-60mm	58	2.48
	61-70mm	69	2.42
	71-80mm	79	2.18

Mean sizes for EOD-2 (2) Size Bins:

		Mean (mm)	Std. Dev.
EOD-4 (1983-1987) Mean Tumor Size by EOD-2 (2) Size Bins	<= 20mm	14	4.54
	21-40mm	30	5.69
	41-60mm	52	5.62

This following table outlines the bin, the mean and median value, and the similar population from which we derived that value.

Summary of Consensus Tumor Size Modifications:

Schema	original code	bin	mean value	std. dev.	source of mean	recoded to:	notes
CS		>= 989mm				999	
EOD 10 digit		<= 3mm				999	
EOD 10 digit		>= 990mm				999	
EOD 4 digit		<= 3mm				999	
EOD 4 digit		>= 100mm				999	
EOD 4 digit		96-99mm				999	
EOD 13 digit	1	<= 4mm				999	
EOD 13 digit	2	5-9mm	7	1.56	EOD 4 (1983-1987)	7	
EOD 13 digit	3	10-19mm	14	2.68	EOD 4 (1983-1987)	14	
EOD 13 digit	4	20-29mm	23	2.65	EOD 4 (1983-1987)	23	
EOD 13 digit	5	30-39mm	32	2.47	EOD 4 (1983-1987)	32	
EOD 13 digit	6	40-49mm	41	2.27	EOD 4 (1983-1987)	41	
EOD 13 digit	7	50-99mm	62	12.32	EOD 4 (1983-1987)	62	
EOD 13 digit	8 -> 999	>= 100mm				999	
EOD 2 digit (2x,3x,5x,6x)	1	<=10mm	8	1.91	EOD 4 (1983-1987)	8	EOD-4 data only continuous above 3mm
EOD 2 digit (2x,3x,5x,6x)	2	11-20mm	17	2.93	EOD 4 (1983-1987)	17	
EOD 2 digit (2x,3x,5x,6x)	3	21-30mm	27	2.92	EOD 4 (1983-1987)	27	
EOD 2 digit (2x,3x,5x,6x)	4	31-40mm	37	2.86	EOD 4 (1983-1987)	37	
EOD 2 digit (2x,3x,5x,6x)	5	41-50mm	48	2.71	EOD 4 (1983-1987)	48	
EOD 2 digit (2x,3x,5x,6x)	6	51-60mm	58	2.48	EOD 4 (1983-1987)	58	

EOD 2 digit (2x,3x,5x,6x)	7	61-70mm	69	2.42	EOD 4 (1983-1987)	69	
EOD 2 digit (2x,3x,5x,6x)	8	71-80mm	79	2.18	EOD 4 (1983-1987)	79	
EOD 2 digit (2x,3x,5x,6x)	9 -> 999	>= 81mm				999	
EOD 2 digit (7x,8x)	1 or 6 --> 20	<= 20mm	14	4.54	EOD 4 (1983-1987)	14	EOD-4 data only continuous above 3mm
EOD 2 digit (7x,8x)	2 or 7 --> 40	21-40mm	30	5.69	EOD 4 (1983-1987)	30	
EOD 2 digit (7x,8x)	3 or 8 --> 60	41-60mm	52	5.62	EOD 4 (1983-1987)	52	
EOD 2 digit (7x,8x)	4 or 9 --> 999	>= 61mm				999	

Source: “PATH:\TumorSize_coding_table.xls”;

Note: all "000" values, or respective equivalents, were recoded to "999" because they offered no useful data

Discussion of the tumor size data

One should note that, in creating a consensus tumor size field, we used mean tumor sizes from one cohort, and assigned those values to another cohort, estimating that the two cohorts actually represent the same population, and that the arithmetic mean of a size bin is a useful measure of centrality. We deemed that if the cohorts were similar in their distribution of tumor sizes, then such reassignment was permissible. The data we used to determine population homogeneity of tumor size was the count/percentage of patients within a size bin over time.

If one wishes to exclude the estimated tumor size data, one must select against the cases with binned tumor size (EOD 2-digit, 13-digit, etc.).

Consensus Number of Positive Lymph Nodes, Number of Nodes Examined, and General Node Positivity:

Table of sources for Lymph Node data:

Nodal Info Source	type of info	years	notes
EOD13_NumPosNodes	Number of Positive Nodes	1973-1982	0-7, 8=8+ PosNodes, 9=NodePos_non-specific
EOD13_NumNodesExamined	Number of Nodes Examined	1973-1982	00-98, 99=unknown/not specified
EOD-Old_2_Digit	Node Positive/Negative	1973-1982	left-most digit = 5,8,-,& means "Lymph Nodes Involved"; 2,3,6,7,9 means no LN involved
EOD-Old_4_Digit	Node Positive/Negative	1983-1987	right-most digit = 1-8 means LN involved, 0 means no LN involved
Regional_Nodes_Positive	Number of Positive Nodes	1988+	00-89, 90=90+ PosNodes, 95=positive aspiration performed, 97=NodePos_non-specific, 98-99=not done/unknown; 2004 values are exact copy of CS_Site-Specific_Factor_3
Regional_Nodes_Examined	Number of Nodes Examined	1988+	00-89, 90=90+ NodesExamined, 95=positive aspiration performed, 96-98=Nodes removed but non-specific, 99=unknown
CS_Site-Specific_Factor_3	Number of Positive Nodes	2004+	00-89, 90=90+ PosNodes, 95=positive aspiration performed, 97=NodePos_non-specific, 98-99=not done/unknown

Source: “PATH:\Nodal_info_table.xls”

Three new fields were made to include consensus data on lymph nodes. They are:

NumPosNodes_consensus
NumNodesExamined_consensus
Node_pos-neg_consensus

Note that NumPosNodes_consensus excludes upperbound groups, such as “8+,” because these data are useless for our calculations. Also note that Node_pos-neg_consensus includes "node positive" patients who had an unknown exact number of positive nodes.

MN considered patients with a reported NumPosNodes but no value for NumNodesExamined. The data appears to be consistent, continuous, and valid, and so we did not exclude these cases.

MN checked that NumPosNodes <= NumNodesExamined for every case.

Characterization of the data

In this section of the report we will characterize the SEER Breast Database, offering frequencies and survival rates for relevant sub-populations within the dataset.

Original data source: SEER_BrCaDB_73-04_MASTER.accdb
SPSS data export for investigations: SEER_BreastDB_2007.sav
Output documents SEER_2007_characterization_[#].spo
Excel file for graphs: SEER_BreastDB07_report_creation-characterization.xlsx.

Core Stats:

Total cases of breast tumors	799,999
Unique patients	731,806
Median follow-up (n=731,806)	4.416 years

Statistics for Unique Patients (n=731,806)

Frequency, by Sex:

Year of Diagnosis	Sex	
	Female	Male
	Count	Count
Total	726861	4945
1973	7410	70
1974	9718	81
1975	9880	64
1976	9631	74
1977	9541	75
1978	9616	66
1979	9967	82
1980	10171	71
1981	10695	78
1982	10885	81
1983	11593	91
1984	12348	88
1985	13651	88
1986	14339	82
1987	15775	88
1988	15676	117

Year of Diagnosis	Sex	
	Female	Male
	Count	Count
1989	15343	114
1990	16269	114
1991	16861	113
1992	24109	143
1993	23798	159
1994	24486	178
1995	25549	152
1996	26109	192
1997	27694	174
1998	29101	169
1999	29637	196
2000	57716	398
2001	58828	412
2002	58312	358
2003	55882	393
2004	56271	384

Frequency, by Cause of Death:

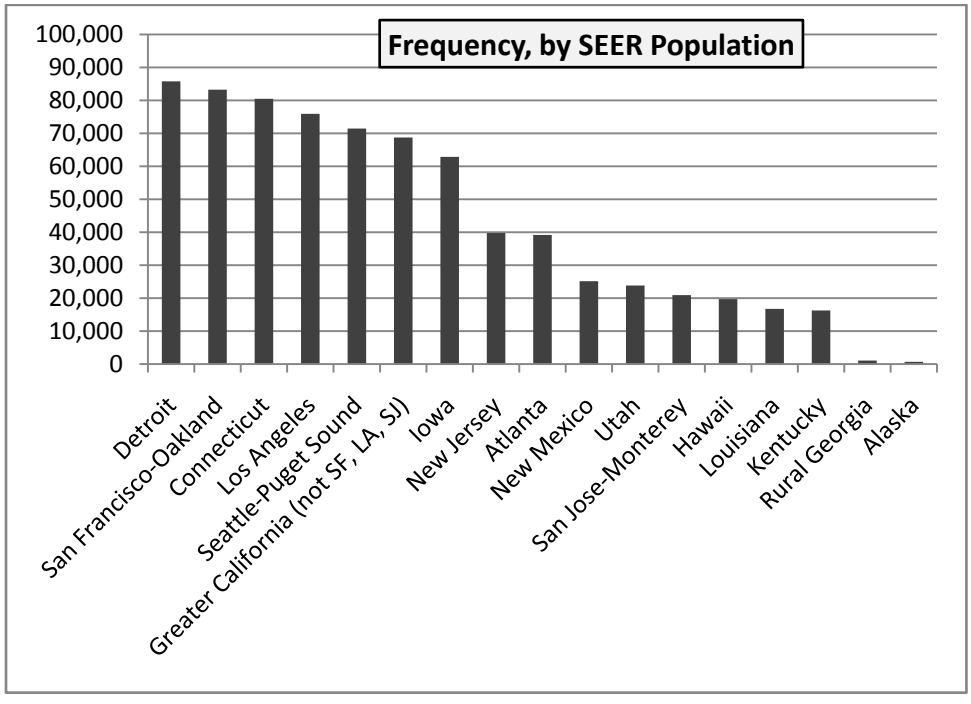
Cause of Death	Count
Alive	478699
Breast	117304
Diseases of Heart	43692
Other Cause of Death	14069
Cerebrovascular Diseases	11680
State DC not available or state DC available but no COD	8866
Lung and Bronchus	6633
Chronic Obstructive Pulmonary Disease and Allied Cond	5765
Pneumonia and Influenza	5062
Miscellaneous Malignant Cancer	4425
Diabetes Mellitus	3426
Colon excluding Rectum	2886
Alzheimers (ICD-9 and 10 only)	2512
Accidents and Adverse Effects	2217
Ovary	1811
Pancreas	1805
Atherosclerosis	1599
Nephritis, Nephrotic Syndrome and Nephrosis	1522
Septicemia	1378

Cause of Death	Count
Non-Hodgkin Lymphoma	1073
Chronic Liver Disease and Cirrhosis	1046
Hypertension without Heart Disease	963
Symptoms, Signs and Ill-Defined Conditions	823
Stomach	744
Aortic Aneurysm and Dissection	716
Other Infectious and Parasitic Diseases	677
In situ, benign or unknown behavior neoplasm	649
Brain and Other Nervous System	574
Acute myeloid	573
Myeloma	564
Other Diseases of Arteries, Arterioles, Capillaries	544
Uterus, NOS	484
Urinary Bladder	445
Suicide and Self-Inflicted Injury	440
Esophagus	439
Stomach and Duodenal Ulcers	424
Kidney and Renal Pelvis	413
Corpus Uteri	413
Liver	399
Rectum and Rectosigmoid Junction	384
Soft Tissue including Heart	347
Melanoma of the Skin	317
Other Acute Leukemia	234
Chronic Lymphocytic Leukemia	195
Cervix Uteri	191
Congenital Anomalies	132
Aleukemic, subleukemic and NOS	131
Gallbladder	129
Intrahepatic Bile Duct	127
Other Biliary	113
Homicide and Legal Intervention	100
Acute Monocytic Leukemia	95
Tongue	89
Prostate	82
Other Non-Epithelial Skin	82
Thyroid	80
Bones and Joints	69
Gum and Other Mouth	68
Larynx	66
Vulva	65
Small Intestine	58
Other Oral Cavity and Pharynx	56
Tuberculosis	55

Cause of Death	Count
Other Endocrine including Thymus$	53
Human Immunodeficiency Virus (HIV) (1987+)	53
Hodgkin Lymphoma	50
Peritoneum, Omentum and Mesentery	49
Acute Lymphocytic Leukemia	49
Other Digestive Organs	48
Other Female Genital Organs	43
Chronic Myeloid Leukemia	41
Vagina	35
Ureter	28
Salivary Gland	27
Nose, Nasal Cavity and Middle Ear	27
Nasopharynx	27
Mesothelioma (ICD-10 only)+	25
Tonsil	22
Retroperitoneum	21
Other Myeloid/Monocytic Leukemia	20
Oropharynx	20
Anus, Anal Canal and Anorectum	20
Pleura	19
Trachea, Mediastinum and Other Respiratory Organs	18
Other Urinary Organs	16
Hypopharynx	16
Floor of Mouth	14
Eye and Orbit	14
Other Lymphocytic Leukemia	10
Complications of Pregnancy, Childbirth, Puerperium	8
Syphilis	4
Testis	3
Certain Conditions Originating in Perinatal Period	3
Kaposi Sarcoma (ICD-10 only)+	2
Other Male Genital Organs	1
Lip	1
Total	731806

Frequency, By SEER Population:

SEER Population	Count
Detroit	85771
San Francisco-Oakland	83270
Connecticut	80466
Los Angeles	75908
Seattle-Puget Sound	71433
Greater California (not SF, LA, SJ)	68749
Iowa	62884
New Jersey	39757
Atlanta	39152
New Mexico	25138
Utah	23846
San Jose-Monterey	20946
Hawaii	19726
Louisiana	16747
Kentucky	16254
Rural Georgia	1075
Alaska	684
Total	731806

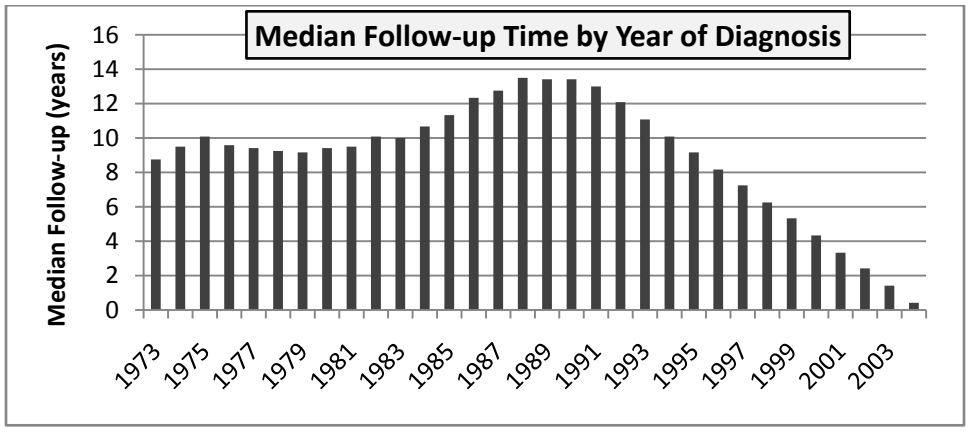


Frequency, by Race/Ethnicity:

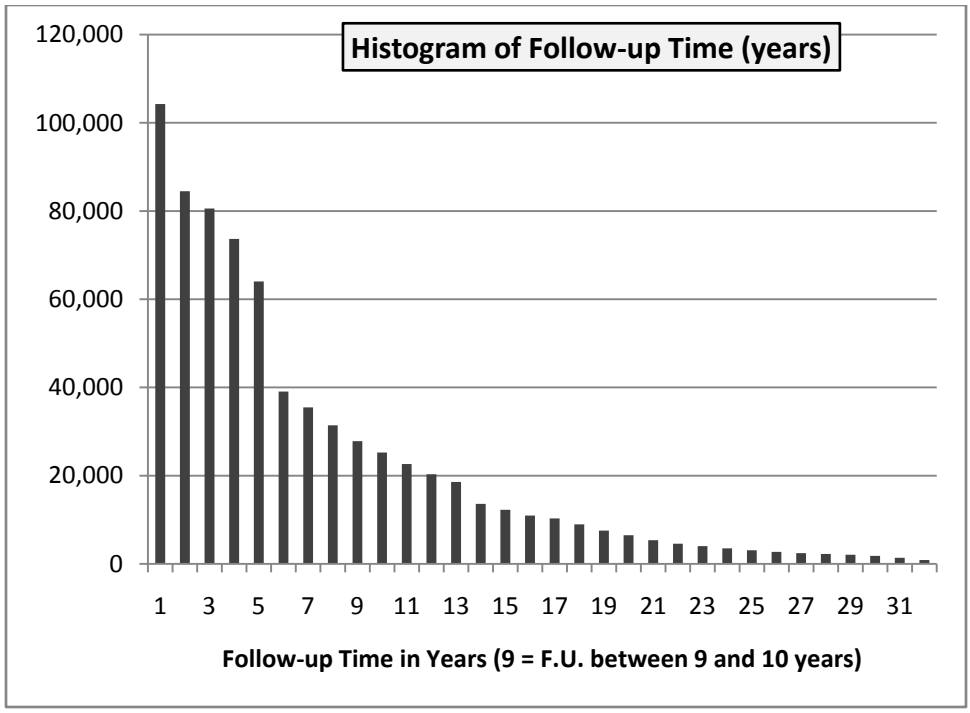
Race/Ethnicity	Count
White	623934
Black	59388
Japanese	10589
Filipino	9624
Chinese	8381
Unknown	4618
Hawaiian	3366
Asian	3000
American Indian (+Aleutian, Alaskan)	2650
Korean	1862
Asian Indian	1736
Vietnamese	1509
Thai	273
Samoan	237
Kampuchean	137
Pacific Islander	118
Laotian	90
Tongan	83
Fiji Islander	56
Guamanian	42
Micronesian	35
Polynesian	24
Hmong	16
Melanesian	15
Tahitian	8
New Guinean	8
Chamorran	7
Total	731806

Median Follow-up Time:

Follow-up Time		Median	Count
Year of Diagnosis	1973	8.750	7480
	1974	9.500	9799
	1975	10.083	9944
	1976	9.583	9705
	1977	9.416	9616
	1978	9.250	9682
	1979	9.166	10049
	1980	9.416	10242
	1981	9.500	10773
	1982	10.083	10966
	1983	10.000	11684
	1984	10.666	12436
	1985	11.333	13739
	1986	12.333	14421
	1987	12.750	15863
	1988	13.500	15793
	1989	13.416	15457
	1990	13.416	16383
	1991	13.000	16974
	1992	12.083	24252
	1993	11.083	23957
	1994	10.083	24664
	1995	9.166	25701
	1996	8.166	26301
	1997	7.250	27868
	1998	6.250	29270
	1999	5.333	29833
	2000	4.333	58114
	2001	3.333	59240
	2002	2.416	58670
	2003	1.416	56275
	2004	.416	56655



Follow-up Time in Years (Binned by 1 year)	Count
1	104246
2	84482
3	80569
4	73659
5	64044
6	39077
7	35490
8	31421
9	27830
10	25262
11	22638
12	20330
13	18559
14	13588
15	12275
16	10941
17	10299
18	8967
19	7552
20	6502
21	5365
22	4568
23	4010
24	3531
25	3093
26	2721
27	2433
28	2244
29	2066
30	1817
31	1363
32	864

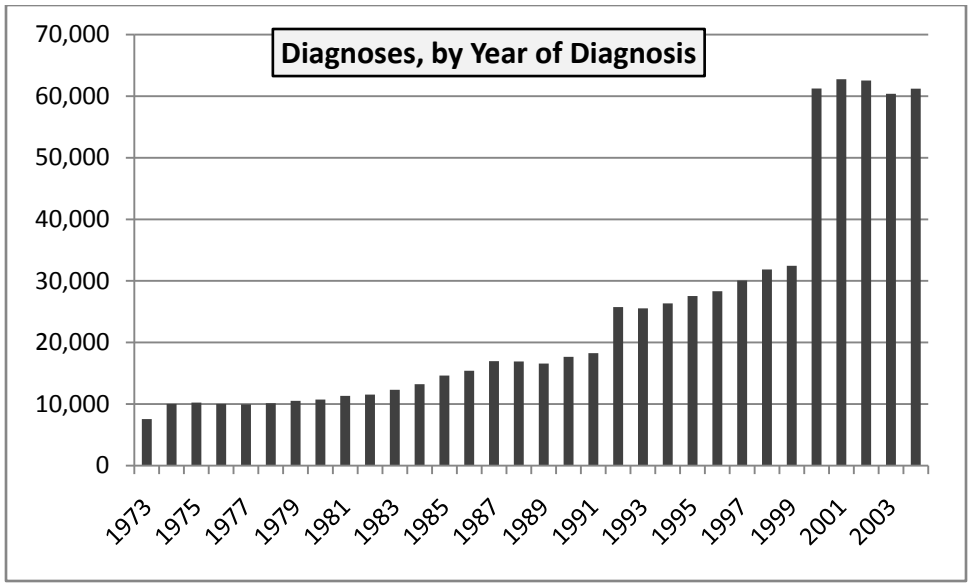

Histogram of Follow-up Time (years)
120,000
100,000
80,000
60,000
40,000
20,000
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Follow-up Time in Years (9 = F.U. between 9 and 10 years)

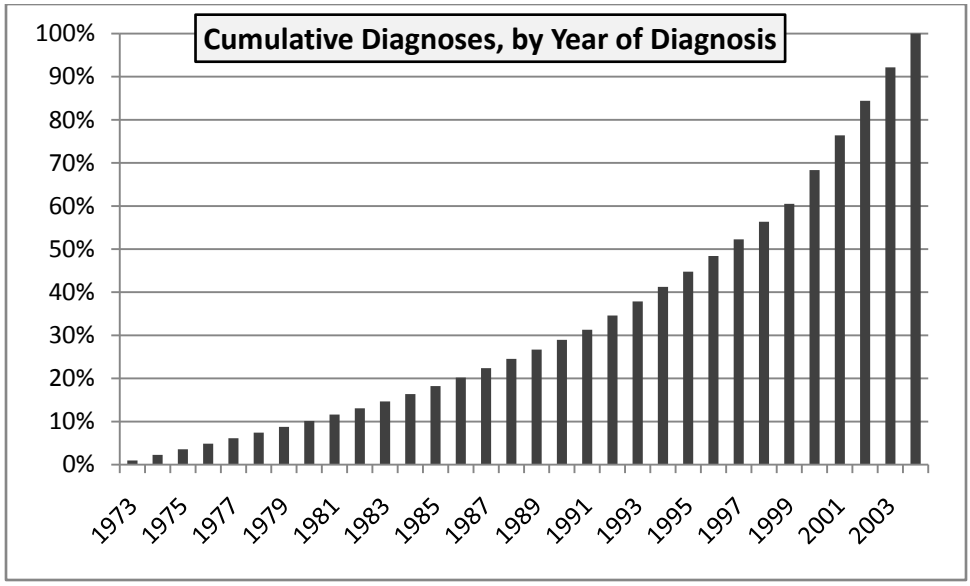
Statistics for Tumor Diagnoses (n=799,999)

Frequency, by Year of Diagnosis:

Year of Diagnosis	Count
1973	7561
1977	9964
1974	10018
1976	10025
1978	10121
1975	10238
1979	10518
1980	10732
1981	11326
1982	11530
1983	12318
1984	13234
1985	14623
1986	15406
1989	16577
1988	16908
1987	16955
1990	17661
1991	18262
1993	25534
1992	25750
1994	26341
1995	27546
1996	28333
1997	30060
1998	31862
1999	32462
2003	60379
2004	61213
2000	61245
2002	62535
2001	62762
Total	779999

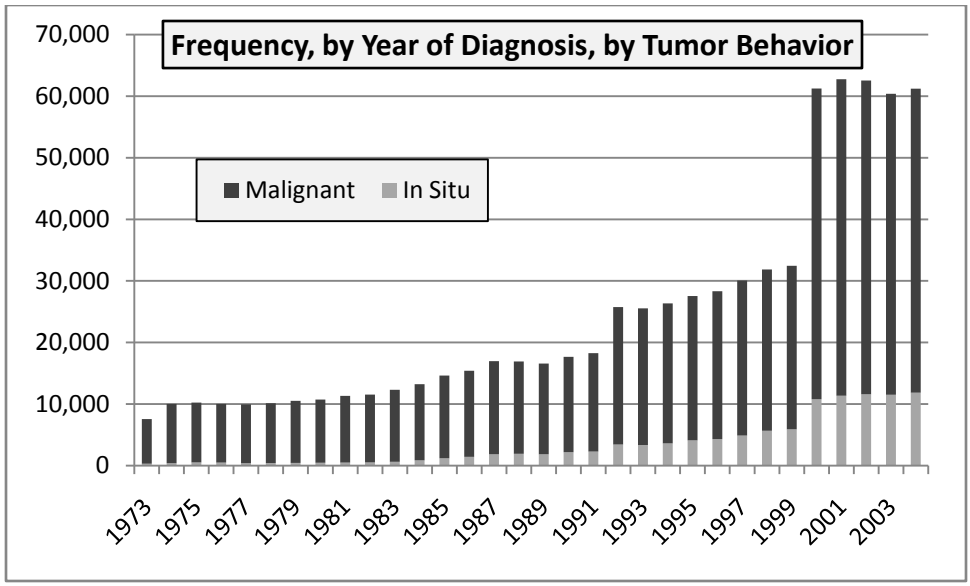
(frequencies describe biopsied tumors, not individual patients)


Diagnoses, by Year of Diagnosis
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
1973
1975
1977
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003


Cumulative Diagnoses, by Year of Diagnosis
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1973
1975
1977
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003

Frequency, by Tumor Behavior, by Year of Diagnosis:

		Behavior	
		In Situ	Malignant
		Count	Count
Year of Diagnosis	1973	310	7251
	1977	393	9571
	1974	380	9638
	1976	488	9537
	1978	405	9716
	1975	521	9717
	1979	432	10086
	1980	459	10273
	1981	486	10840
	1982	517	11013
	1983	613	11705
	1984	870	12364
	1985	1184	13439
	1986	1420	13986
	1989	1848	14729
	1988	1920	14988
	1987	1849	15106
	1990	2171	15490
	1991	2277	15985
	1993	3347	22187
	1992	3432	22318
	1994	3604	22737
	1995	4100	23446
	1996	4302	24031
	1997	4886	25174
	1998	5660	26202
	1999	5906	26556
	2003	11502	48877
	2004	11849	49364
	2000	10777	50468
	2002	11581	50954
	2001	11366	51396
	Total	110855	669144

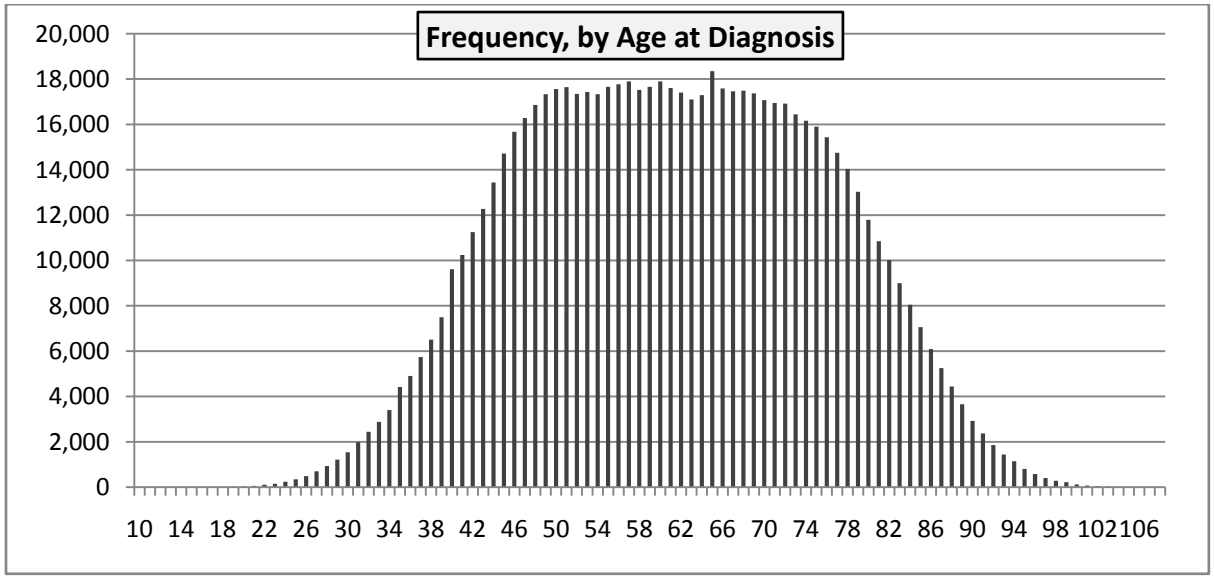

Frequency, by Year of Diagnosis, by Tumor Behavior
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Malignant
In Situ
1973
1975
1977
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003

Frequency, by Age at Diagnosis:

Age at Diagnosis	Count
10	1
11	1
12	2
13	2
14	7
15	8
16	6
17	10
18	12
19	28
20	35
21	54
22	107
23	149
24	239
25	348
26	486
27	701
28	929
29	1215
30	1540
31	1982
32	2443
33	2881
34	3404
35	4417
36	4908
37	5735
38	6505
39	7498
40	9615
41	10238
42	11253
43	12270
44	13439
45	14713
46	15671
47	16280
48	16859
49	17336
50	17560
51	17645
52	17350
53	17430

Age at Diagnosis	Count
54	17330
55	17655
56	17770
57	17896
58	17525
59	17655
60	17900
61	17607
62	17406
63	17103
64	17292
65	18349
66	17587
67	17458
68	17488
69	17366
70	17068
71	16946
72	16917
73	16442
74	16163
75	15902
76	15437
77	14753
78	14034
79	13031
80	11793
81	10847
82	10016
83	8996
84	8045
85	7057
86	6092
87	5252
88	4438
89	3655
90	2921
91	2373
92	1857
93	1443
94	1146
95	802
96	581
97	403
98	281
99	219

Age at Diagnosis	Count
100	118
101	70
102	48
103	32
104	17
105	11
106	13
107	9
108	2
Total	779929



Frequency, by Laterality:

		Count
Laterality	Left	393520
	Right	377166
	Paired	7080
	Bilateral	1148
	Unilateral, but unspecified	1085
	Total	779999

Frequency of Cases Describing First Malignant Primary Diagnosis:

		Count
First Malignant Primary	Yes	584637
	No	195362
	Total	779999

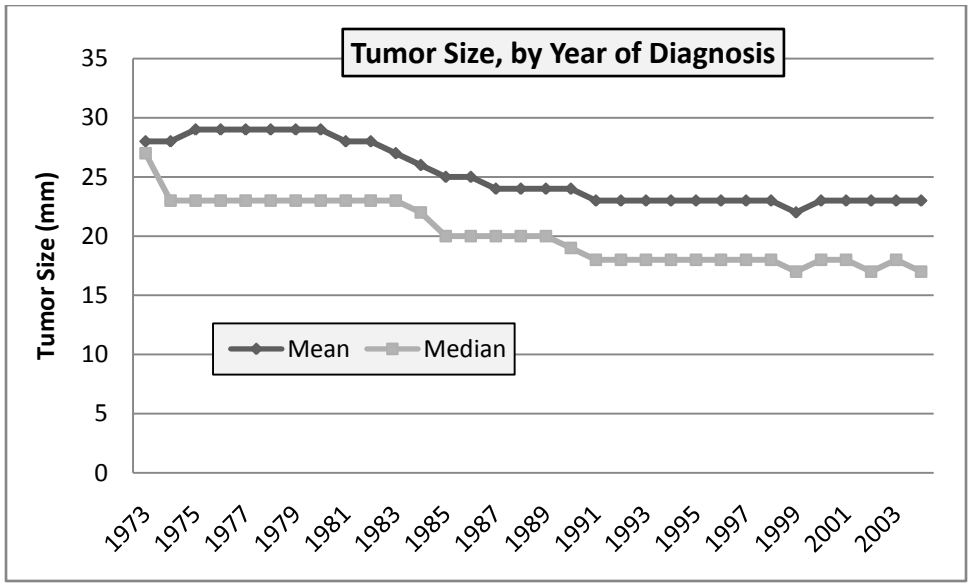
Statistics for First Malignant Primary Tumors (n=584,637)

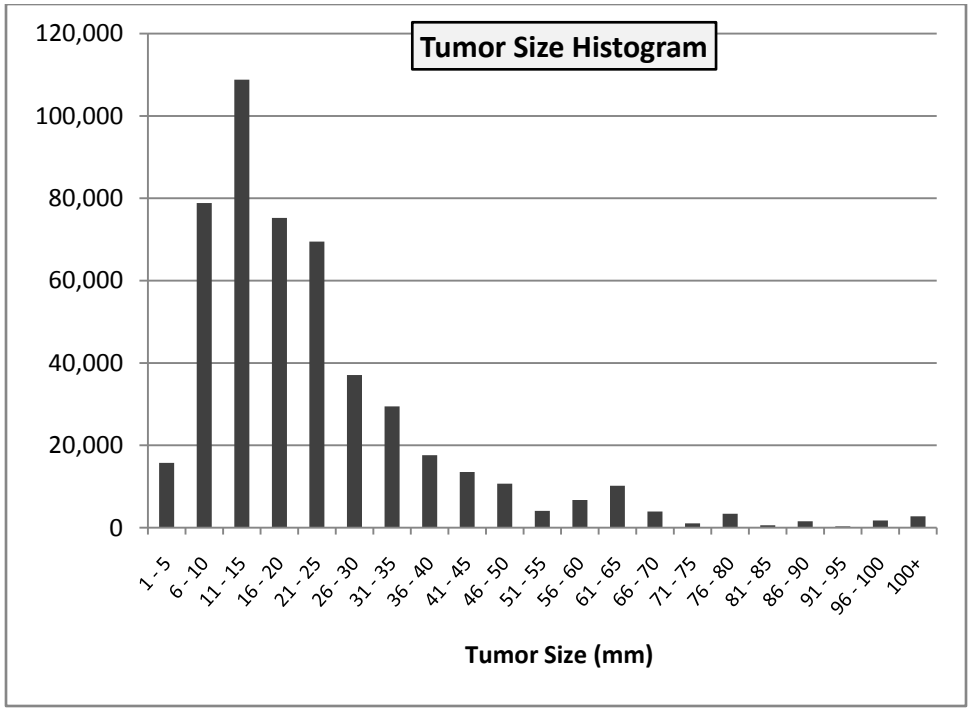
The following statistics describe the SEER population of first malignant primary tumor diagnoses.

Tumor Size:

Tumor Size (mm)		Count	Mean	Standard Deviation	Median	Minimum*	Maximum	Percentile 95
Total		584637	24	20	20	1	985	62
Year of Diagnosis	1973	6673	28	15	27	7	79	62
	1974	8815	28	15	23	7	79	62
	1975	8798	29	16	23	7	62	62
	1976	8651	29	16	23	7	62	62
	1977	8612	29	16	23	7	62	62
	1978	8681	29	16	23	7	62	62
	1979	8993	29	16	23	7	62	62
	1980	9165	29	16	23	7	62	62
	1981	9674	28	15	23	7	62	62
	1982	9815	28	16	23	7	62	62
	1983	10423	27	16	23	4	95	60
	1984	10884	26	16	22	4	95	60
	1985	11832	25	16	20	4	95	60
	1986	12244	25	16	20	4	95	60
	1987	13223	24	15	20	4	95	55
	1988	13056	24	23	20	4	985	60
	1989	12801	24	19	20	4	300	60
	1990	13483	24	21	19	4	913	60
	1991	13954	23	20	18	4	800	55
	1992	19401	23	22	18	4	931	60
	1993	19339	23	20	18	4	900	60
	1994	19806	23	21	18	4	700	60
	1995	20398	23	21	18	4	939	60
	1996	20849	23	19	18	4	520	60
	1997	21899	23	20	18	4	700	60
	1998	22631	23	19	18	4	712	60
	1999	22997	22	19	17	4	778	60
	2000	44038	23	21	18	4	880	60
	2001	44718	23	21	18	4	980	60
	2002	44170	23	21	17	4	925	60
	2003	42255	23	21	18	4	900	60
	2004	42359	23	22	17	1	790	60

*Smallest recorded tumor size varied according to the year/coding scheme used

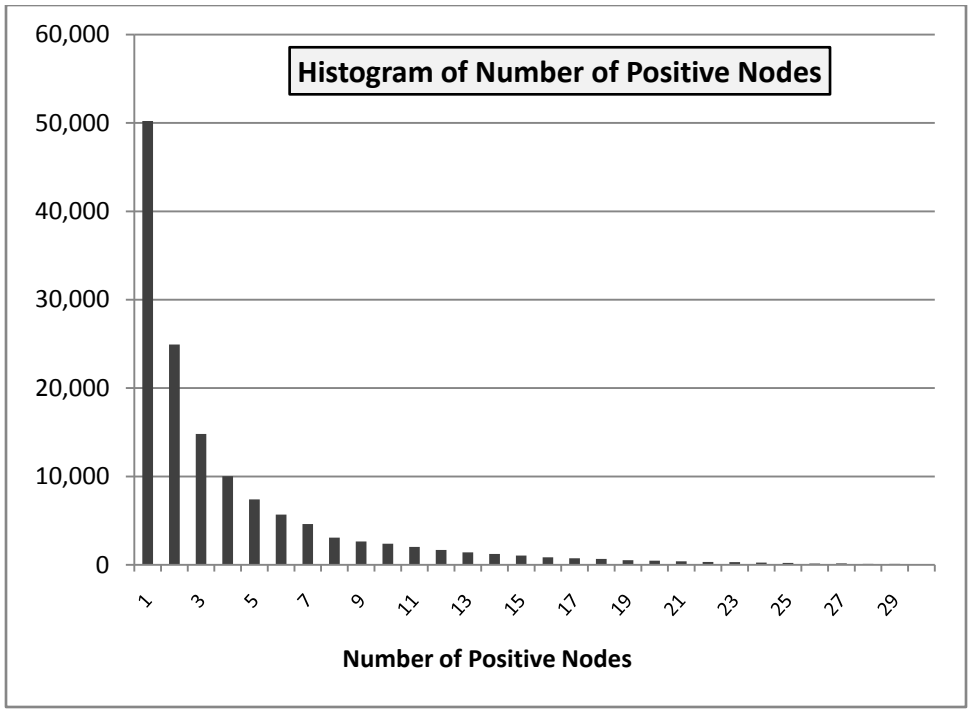

Tumor Size, by Year of Diagnosis
Tumor Size (mm)
35
30
25
20
15
10
5
0
Mean
Median
1973
1975
1977
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003


Tumor Size Histogram
120,000
100,000
80,000
60,000
40,000
20,000
0
1 - 5
6 - 10
11 - 15
16 - 20
21 - 25
26 - 30
31 - 35
36 - 40
41 - 45
46 - 50
51 - 55
56 - 60
61 - 65
66 - 70
71 - 75
76 - 80
81 - 85
86 - 90
91 - 95
96 - 100
100+
Tumor Size (mm)

By Number of Positive Nodes:

Number of Positive Nodes	Count
Total Node Neg.	303193
Total Node Pos.	139062
1	50215
2	24931
3	14818
4	10027
5	7399
6	5682
7	4616
8	3075
9	2642
10	2383
11	2035
12	1676
13	1414
14	1225
15	1045
16	847
17	735
18	678
19	525
20	477
21	403
22	319
23	303
24	247
25	215
26	149
27	154
28	105
29	100
30	99
31	64
32	54
33	52
34	51
35	48
36	35
37	33
38	16

Number of Positive Nodes	Count
39	21
40	26
41	18
42	15
43	13
44	10
45	12
46	7
47	4
48	7
49	1
50	2
51	7
52	2
53	4
54	1
55	1
56	2
58	3
60	2
61	4
69	1
73	1
74	1
75	1
78	1
79	1
84	1
88	1


Histogram of Number of Positive Nodes
60,000
50,000
40,000
30,000
20,000
10,000
0
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
Number of Positive Nodes

By Grade:

Grade	Count
1	67779
2	156652
3	146905
4	13340
Total	384676

By ER Status:

ER Status	Count
Borderline	1884
Negative	75040
Positive	254128

By PR Status:

PR Status	Count
Borderline	3103
Negative	107603
Positive	212130

By ER/PR Status:

ER/PR Status	Count
ER-/PR-	65472
ER-/PR+	7928
ER+/PR-	41364
ER+/PR+	203243

By Histological Type:

Histology	Count
Ductal	398506
Lobular	41623
Intraductal+Lobular in situ	32659
Adenocarcinoma	19579
Carcinoma, NOS	16654
Mucinous adenocarcinoma	13245
Medullary carcinoma	8027
Comedocarcinoma	7830
Tubular adenocarcinoma	7360
Infiltr. duct mixed with other types of carcinoma	5295
Inflammatory carcinoma	5241
Neoplasm, malig	5240
Paget's disease w/ inf. duct. carc.	2703
Scirrhous adenocarcinoma	2692
Papillary adenocarcinoma	2159
Papillary carcinoma	1668
Paget disease and intraductal ca.	1452
Phyllodes tumor, malig	1356
Cribriform carcinoma	1044
Infiltrating ductular carcinoma	985
Apocrine adenocarcinoma	952
Infiltrating lobular mixed with other types of carc.	879
Mucin-producing adenocarcinoma	758
Intracystic carcinoma	660
Paget disease, mammary	622
Adenoid cystic carcinoma	494
Metaplastic carcinoma	465
Medullary carcinoma with lymphoid stroma	342
Signet ring cell carcinoma	324
Squamous cell carcinoma	302
Carcinoma, undifferentiated	270
Carcinoma, anaplastic	245
Atypical medullary carcinoma	241
Small cell carcinoma	175
Hemangiosarcoma	170
Solid carcinoma	162
Intraductal micropapillary carcinoma	161
Carcinosarcoma	144
Adenocarcinoma with squamous metaplasia	131
Adenosquamous carcinoma	129
Spindle cell carcinoma	116
Tumor cells, malig	105

Histology	Count
Secretory carcinoma of breast	92
Sarcoma	92
Large cell carcinoma	92
Neuroendocrine carcinoma	86
Pleomorphic carcinoma	81
Stromal sarcoma	75
Adenocarcinoma with spindle cell mataplasia	63
Clear cell adenocarcinoma	61
Adenocarcinoma with apocrine metaplasia	59
Fibrous histocytoma	53
Fibrosarcoma	47
Spindle cell sarcoma	46
Adenocarcinoma w cartilag. & oss. metaplas.	42
Squamous cell carcinoma, keratinizing	34
Mixed cell adenocarcinoma	33
Unspecified histological type	31
Pseudosarcomatous carcinoma	31
Acinar cell carcinoma	28
Leiomyosarcoma	24
Adenocarcinoma with mixed subtypes	22
Epithelial-myoepithelial carcinoma	21
Squamous cell carcinoma, spindle cell	20
Malignant myoepithelioma	19
Carcinoid tumor	18
Osteosarcoma	17
Duct carcinoma, desmoplastic type	17
Glycogen-rich carcinoma	16
Carcinoma simplex	15
Liposarcoma	14
Polymorphous low grade adenocarcinoma	13
Alveolar adenocarcinoma	13
Adenocarcinoma with neuroendocrine differen.	12
Cystadenocarcinoma	10
Myxoid liposarcoma	9
Giant cell sarcoma	9
Mixed tumor, malig	7
Malig tumor, spindle cell	7
Liposarcoma, well differentiated	7
Sweat gland adenocarcinoma	6
Dermatofibroma	6
Chondroma	6

Histology	Count
Papillary cystadenoma	5
Lipid-rich carcinoma	5
Giant cell carcinoma	5
Trabecular adenocarcinoma	4
Pleomorphic liposarcoma	4
Mucinous cystadenocarcinoma	4
Hemangioendothelioma, malig	4
Fibromyxosarcoma	4
Adenomyoepithelioma	4
Verrucous carcinoma	3
Papillary squamous cell carcinoma	3
Mucoepidermoid carcinoma	3
Mesenchymoma, malig	3
Malig tumor, giant cell	3
Granular cell carcinoma	3
Adenosarcoma	3
Squamous cell carcinoma, adenoid	2
Papillary transitional cell neoplasm (low malig. Potential)	2
Paget disease, extramammary	2
Myosarcoma	2
Lymphoepithelioma	2
Large cell neuroendocrine carcinoma	2
Infantile fibrosarcoma	2
Granular cell tumor, malig	2
Epithelioma	2
Epithelioid sarcoma	2
Carcinoma w/ osteoclast-like giant cells	2
Alveolar rhabdomyosarcoma	2
Undifferentiated sarcoma	1
Telangiectatic osteosarcoma	1
Superficial spreading melanoma	1
Superficial spreading adenocarcinoma	1
Squamous cell carcinoma, lg. cell, non-ker.	1
Small cell sarcoma	1
Small cell - large cell carcinoma	1
Rhabdomyoma	1
Renal cell carcinoma	1
Pleomorphic rhabdomyosarcoma	1
Pilomatrixoma	1
Papillary cystic tumor	1
Nonencapsulating sclerosing carcinoma	1
Myxoid leiomyosarcoma	1
Myxoid chondrosarcoma	1

Histology	Count
Mixed type liposarcoma	1
Melanoma	1
Malig tumor, small cell	1
Malig tumor, clear cell	1
Hermangiopericytoma, malig	1
Ewing sarcoma	1
Epithelioid leiomyosarcoma	1
Epithelioid hemangioendothelioma, malig	1
Endometrial stromal nodule	1
Embryonal rhabdomyosarcoma	1
Cystic hypersecretory carcinoma	1
Clear cell sarcoma	1
Chrondroblastic osteosarcoma	1
Bronchiolo-alveolar adenocarcinoma	1
Basaloid carcinoma	1
Basal cell carcinoma	1
Astrocytoma	1
Angiomyosarcoma	1
Alveolar soft part sarcoma	1
Total	584,637