PharmaSUG 2014 - Paper AD27

Expediting Access to Critical Pathology Data

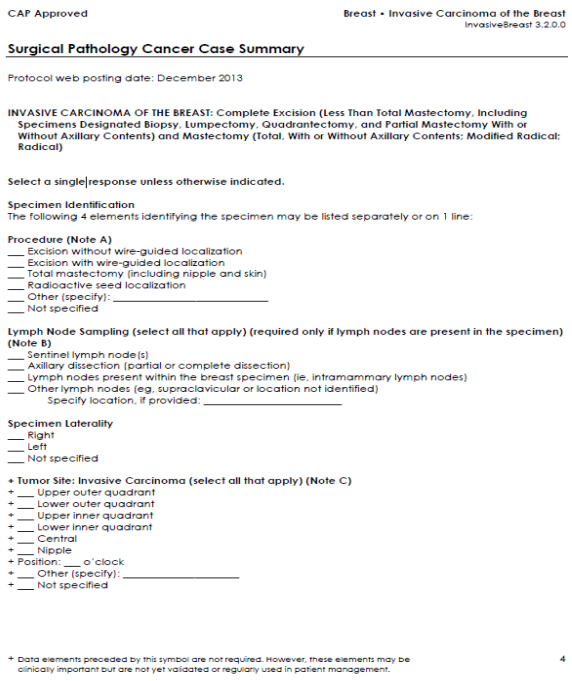
Rebecca Ottesen, City of Hope, Duarte, CA
Leanne Goldstein, City of Hope, Duarte, CA
Julie Kilburn, City of Hope, Duarte, CA
Joyce Niland, City of Hope, Duarte, CA

ABSTRACT

Abstracting information from pathology notes is often quite cumbersome. Clinical Research Associates and Tumor Registrars typically have to read through all of the diagnosis information and manually enter the data into a database for fields such as tumor size, lymph nodes, and staging. This manual process can lead to data that is subject to interpretation and data entry errors, in addition to the workload burden. In this presentation, we will demonstrate approaches to simplifying data abstraction from pathology reports using various SAS® programming techniques. First, we consider the use of semi-coded College of American Pathologists (CAP) synoptic worksheet data based on a check list filled out by pathologists when reviewing a surgical specimen. We also evaluate the approach of performing string searches on the unstructured pathology dictation using functions such as INDEX() and SUBSTR(). Finally, we demonstrate an approach of identifying data from the pathology diagnosis using SAS Text Miner in SAS Enterprise Miner Workstation 12.1. With all of these tools at hand, it is much easier to meet research needs and to analyze pathology data efficiently.

INTRODUCTION

The first approach to accessing pathology diagnostic information is the use of synoptic worksheets produced by the College of American Pathologists, an organization consisting of board-certified pathologists who produce peer-reviewed cancer protocols in an effort to streamline the data capture of surgical pathology findings and to improve quality cancer diagnosis and care. The CAP protocol checklists (Figure 1) are designed to capture key fields from a cancer diagnosis as structured data, to supplement the traditional unstructured pathology dictation. This coded worksheet data represents information used for diagnosis on a consistently structured synoptic report to assist physicians in communicating more effectively about patient care. Even though the data are coded, it typically requires human abstraction of data into research databases and cancer registries. The abstraction process is facilitated by these organized checklists but the data still are entered manually.



CAP Approved

Breast • Invasive Carcinoma of the Breast
InvasiveBreast 3.2.0.0

Surgical Pathology Cancer Case Summary

Protocol web posting date: December 2013

INVASIVE CARCINOMA OF THE BREAST: Complete Excision (Less Than Total Mastectomy, Including Specimens Designated Biopsy, Lumpectomy, Quadrantectomy, and Partial Mastectomy With or Without Axillary Contents) and Mastectomy (Total, With or Without Axillary Contents: Modified Radical; Radical)

Select a single response unless otherwise indicated.

Specimen Identification
The following 4 elements identifying the specimen may be listed separately or on 1 line:

Procedure (Note A)
___ Excision without wire-guided localization
___ Excision with wire-guided localization
___ Total mastectomy (including nipple and skin)
___ Radioactive seed localization
___ Other (specify): _____
___ Not specified

Lymph Node Sampling (select all that apply) (required only if lymph nodes are present in the specimen) (Note B)
___ Sentinel lymph node(s)
___ Axillary dissection (partial or complete dissection)
___ Lymph nodes present within the breast specimen (ie, intramammary lymph nodes)
___ Other lymph nodes (eg, supraclavicular or location not identified)
Specify location, if provided: _____

Specimen Laterality
___ Right
___ Left
___ Not specified

+ Tumor Site: Invasive Carcinoma (select all that apply) (Note C)
+ ___ Upper outer quadrant
+ ___ Lower outer quadrant
+ ___ Upper inner quadrant
+ ___ Lower inner quadrant
+ ___ Central
+ ___ Nipple
+ Position: ___ o'clock
+ ___ Other (specify): _____
+ ___ Not specified

+ Data elements preceded by this symbol are not required. However, these elements may be clinically important but are not yet validated or regularly used in patient management.

4

Figure 1. Example of a CAP worksheet for invasive breast cancer patients

Electronic access to coded synoptic data is helpful in that the data is entered in near real-time, it is high quality diagnostic data entered by pathologists as a requirement of the cancer center, and contains critical core diagnostic information such as cancer staging. However, this data source also has limitations as it is narrow in scope with little treatment information, and only focuses on certain specimens under evaluation. An additional limitation is that at City of Hope we only have access to synoptic data for surgical resections after 2006. Quite often we would like to know more detail about a diagnosis which could involve interpretation of outside biopsies or a diagnosis of metastasis (non-local spread of cancer) that relies on information across other specimens. To incorporate more information or evaluate cases pre-dating 2006, we need to pull data from the unstructured text of the pathology report and final diagnosis.

SYNOPTIC WORKSHEETS

At the time of the clinical diagnostic dictation by the pathologist (Figure 2) the data for the synoptic worksheet is entered into a database (Figure 3). (Note that we will use a fictitious diagnosis as an example throughout this paper.) This consistency in timeframe leads to real-time data entry and less redundancy. At City of Hope the synoptic worksheet data are imported to the data warehouse from a vendor based system: Cerner ® CoPATH v3.2.2.214. This relational database consists of many of hierarchical tables, not only for the purpose of synoptic reporting, but also related to many areas of patient care such as billing, encounters, diagnosis, and physician identifiers. This database utilizes dictionary tables to house the common coding of data elements and the relevant patient data that is linked to the synoptic reporting of data specimens.

Final Diagnosis

BREAST AND AXILLARY LYMPH NODES, RIGHT, MASTECTOMY WITH AXILLARY NODE DISSECTION (A):
- INFILTRATING DUCTAL CARCINOMA (5.2 CM), PLEOMORPHIC TYPE WITH A MINOR COMPONENT SHOWING TUBULAR FEATURES (SEE NOTE AND SYNOPTIC REPORT)
- MINOR COMPONENT OF DUCTAL CARCINOMA-IN-SITU WITH PAGETOID INVOLVEMENT OF THE DUCTS
- ATYPICAL DUCTAL HYPERPLASIA (BLOCK A3), SCLEROSING ADENOSIS AND COLUMNAR CELL HYPERPLASIA
- BENIGN NIPPLE
- RESECTION MARGINS FREE OF CARCINOMA (2.5 FROM THE DEEP MARGIN)
- THE INVASIVE CARCINOMA IS NEGATIVE FOR ER AND PR EXPRESSION AND EQUIVOCAL FOR HER-2/neu OVEREXPRESSION (SEE NOTE)

- METASTATIC ADENOCARCINOMA (UP TO 6.5 CM) IN TWENTY-EIGHT OF THIRTY AXILLARY LYMPH NODES (28/30), WITH EXTRANODAL EXTENSION
- THE METASTATIC ADENOCARCINOMA IS NEGATIVE FOR HER-2/neu GENE AMPLIFICATION

Figure 2. Fictitious example of unstructured text in a pathology report

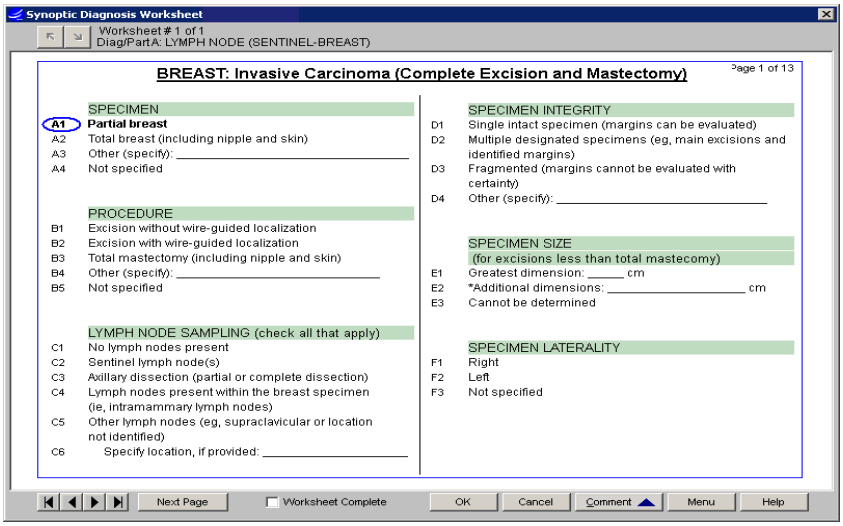


Figure 3. User interface of the synoptic worksheet database entry screen at City of Hope

Accessing synoptic data to utilize it for analytics is challenging due to its structure, which consists of multiple relational database tables that need to be joined. The default synoptic report view provided by the database is also difficult to use. The structure of the view is vertical; it contains one column with one row per data element corresponding to values from the CAP checklist (Figure 4). In addition, the key identifiers used in the backend database convey no metadata, such as "cop1625," a key that corresponds to a text based label (Figure 5).

Medical Record Number	Surgical Path Number	Category	Filled In Information
9999999	X10-12345	Breast-Size of Invasive Component	Greatest dimension: 5.2 cm
9999999	X10-12345	Breast-Histologic Type	Ductal carcinoma in situ
9999999	X10-12345	Breast-Histologic Type	Invasive ductal
9999999	X10-12345	Breast-Histologic Grade	Tubule-Minimal less than 10% (score = 3)
9999999	X10-12345	Breast-Histologic Grade	Nuclear-Marked variation in size, nucleoli, chromatin clumping, etc (score = 3)
9999999	X10-12345	Breast-Histologic Grade	40x-6 to 10 mitoses per 10 HPF (score = 2)
9999999	X10-12345	Breast-Histologic Grade	Nottingham Score-Grade III: 8-9 points
9999999	X10-12345	Breast-Pathologic Staging (pTNM)	pT3: Tumor more than 5.0 cm in greatest dimension
9999999	X10-12345	Breast-Pathologic Staging (pTNM)	pN3a: Metastasis in 10 or more axillary lymph nodes (at least 1 tumor deposit greater than 2.0 mm), or metastasis to the intraclavicular lymph nodes
9999999	X10-12345	Breast-Pathologic Staging (pTNM)	Specify: Number examined: 30
9999999	X10-12345	Breast-Pathologic Staging (pTNM)	Number involved: 28
9999999	X10-12345	Breast-Pathologic Staging (pTNM)	pMX: Cannot be assessed
9999999	X10-12345	Breast-Margins	Margins uninvolved by invasive carcinoma

Figure 4. Default synoptic worksheet print out provided by the database

Medical Record Number	Surgical Path Number	category_id	category_name	synoptic_value_id	fillin_char	SVname	seq
9999999	X10-12345	cop7205	Breast-Size of Invasive Component	cop1625	5.2	Greatest dimension: ___ cm	6
9999999	X10-12345	cop8205	Breast-Histologic Type	cop9625		Ductal carcinoma in situ	7
9999999	X10-12345	cop8205	Breast-Histologic Type	cop5725		Invasive ductal	8
9999999	X10-12345	cox08	Breast-Tubule Formation	cop3925		Tubule-Minimal less than 10% (score = 3)	9
9999999	X10-12345	cox18	Breast-Nuclear pleomorphism	cop6925		Nuclear-Marked variation in size, nucleoli, chromatin clumping, etc (score = 3)	10
9999999	X10-12345	cop9205	Breast-Histologic Grade	cop1035		40x-6 to 10 mitoses per 10 HPF (score = 2)	11
9999999	X10-12345	cox48	Breast-Total Nottingham Score	cop5035		Nottingham Score-Grade III: 8-9 points	12
9999999	X10-12345	cop7335	Breast-Pathologic Staging (pTNM)	cop4235		pT3: Tumor more than 5.0 cm in greatest dimension	13
9999999	X10-12345	cox98	Breast-Regional Lymph Nodes (pN)	cop2435		pN3a: Metastasis in 10 or more axillary lymph nodes (at least 1 tumor deposit greater than 2.0 mm), or metastasis to the intraclavicular lymph nodes	14
9999999	X10-12345	cop7335	Breast-Pathologic Staging (pTNM)	cop6435	30	Specify: Number examined: ____	15
9999999	X10-12345	cop7335	Breast-Pathologic Staging (pTNM)	cop7435	28	Number involved: ____	16
9999999	X10-12345	cop7335	Breast-Pathologic Staging (pTNM)	cop8435		pMX: Cannot be assessed	17
9999999	X10-12345	cop3305	Breast-Margins	cop3535		Margins uninvolved by invasive carcinoma	18

Figure 5. Structure of the synoptic worksheet data from the database

To create a more analytic-friendly data set SAS was used to convert the somewhat cryptic backend data into a more usable view of the synoptic worksheet data. The result of our “data massaging” creates a data set with one row per specimen, meaningful column names and coded values rather than the original text based data. This result was achieved through a series of steps involving multiple joins and data set restructuring to fit our needs.

Step 1) Combine all relevant data dictionary tables and join them with the specimen level data by key ids.

Step 2) Rather than use the text-based category_id and synoptic_value_id keys, we leverage the use of the numerically coded seq field. This field comes directly from the backend database and corresponds to the order of display on the screen, however, it is a unique numeric code within each CAP worksheet. In

addition, a code of 6 is much easier to reference in analysis than a text based code of 'cop1625'.

Step 3) The category_id and category_name variables are not helpful as they often are too broad across actual variables where we would want to store information individually. For example 'cop7335' corresponds to 'Breast-Pathologic Staging (pTNM)' when in reality this category holds data for total number of lymph nodes, number of lymph nodes positive and M stage, which are distinct fields. To alleviate this situation, a one-time effort was taken to review the entire data dictionary of possible categories and code values, and then identify meaningful custom variable names for corresponding collections of codes.

Step 4) The user defined variable names set created in Step 3), key identifiers per the backend database, and seq codes are joined with the specimen level data created in Step 1) into a master data set.

Step 5) Finally the master data set is flipped with PROC TRANSPOSE according to the user defined variable names created in Step 3). The result is a horizontal analytic data set that contains one record per specimen.

The final analytic data set consists of: the seq code (in order to create a unique numeric code for each category); the fill-in fields for any non-coded data; and a text based category name (based on the SVname variable, to provide the corresponding character label). The PREFIX= option for PROC TRANSPOSE in step 5) allows us to utilize a consistent naming convention for the resulting data: the coded value is assigned a variable name, and has a corresponding _Char variable for the text description, and an optional _Fill-in field for further detail (e.g. Tstage, Tstage_Char, Tstage_Fillin). An example of a portion of the resulting data set**,** shown in Figure 6**,** uses the coded seq data which is formatted for readability.

The SAS System

Medical record number	Surgical path number	Greatest dimension	In situ histology	Invasive histology	Tubule formation	Nuclear grade	Nottingham grade
9999999	X10-1234	5.2	Ductal carcinoma in situ	Invasive ductal	Minimal less than 10% (score = 3)	Marked variation in size, nucleoli, chromatin clumping, etc (score = 3)	40x-6 to 10 mitoses per 10 HPF (score = 2)

Figure 6. Formatted SAS data in a horizontal structure

In our resulting analytic data set we now have information that is easier to handle in SAS with statistical procedures. We can obtain the results we need whether it be on the coded, character or fill-in text synoptic data. An example of frequencies of hormone receptor and Her2neu status is shown in Figure 7. This transformation of the synoptic worksheet data has now been scheduled as a batch job that runs once a week when the relevant synoptic data in the warehouse is refreshed. The result is a ready to use analytic data set every Monday morning.

Estrogen Receptor

ER_Char	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Null	6	0.8	6	0.8
Estrogen Receptor, Immunoreactive tumor cells present (>= 1%)	577	76.93	583	77.73
Estrogen Receptor, Less than 1% immunoreactive cells present	34	4.53	617	82.27
Estrogen Receptor, No immunoreactive tumor cells present	86	11.47	703	93.73
Estrogen Receptor, Results unknown	4	0.53	707	94.27
Estrogen Receptor, Results, Other (specify): _____	43	5.73	750	100

Progesterone Receptor

PR_Char	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Null	8	1.07	8	1.07
Progesterone Receptor, Immunoreactive tumor cells present (>= 1%)	491	65.47	499	66.53
Progesterone Receptor, Less than 1% immunoreactive cells present	73	9.73	572	76.27
Progesterone Receptor, No immunoreactive tumor cells present	128	17.07	700	93.33
Progesterone Receptor, Results unknown	5	0.67	705	94
Progesterone Receptor, Results, Other (specify): _____	45	6	750	100

Her2 by IHC

Her2IHC_Char	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Null	43	5.73	43	5.73
HER2/neu-Results-Equivocal (Score 2+)	233	31.07	276	36.8
HER2/neu-Results-Negative (Score 0)	90	12	366	48.8
HER2/neu-Results-Negative (Score 1+)	234	31.2	600	80
HER2/neu-Results-Other (specify): _____	64	8.53	664	88.53
HER2/neu-Results-Positive (Score 3+)	75	10	739	98.53
HER2/neu-Results-Results unknown	11	1.47	750	100

Figure 7. Frequency output of selected character fields

STRING SEARCHES

Our work with the synoptic data is limited for several reasons. The main restriction is that the synoptic reports performed at our cancer center are only on the surgical resection of the primary cancer site, and not on all patient specimens. However, for research purposes investigators may be interested in obtaining not only the primary site specimen but also the specimen containing metastasis. For example, if a breast cancer patient has metastasis to the brain, only the breast specimen would have a synoptic report but not necessarily the brain specimen. This brings to light a problem of how to identify the metastasis specimen for a breast cancer patient or even how to identify patients with metastasis in general without human review of the medical record.

A simple way of looking for metastatic cancer patients and their specimens is by conducting a text string search against the unstructured text in the final diagnosis (as shown in Figure 2) for the word *METASTATIC* or its variants, for example *METASTASES* and *METASTASIS.*

Suppose the text from Figure 2 was in the variable *finaldiagnosis* in the SAS data set called *figure2*. The following DATA step could be used to identify final diagnoses with any mention of metastasis, metastatic or metastases. Note that the final diagnosis data is consistently entered in caps therefore the use of the UPCASE() function is not necessary.

```
DATA tmp;
      SET figure2;
      loc=INDEX(finaldiagnosis,"METAS");
RUN;
```

The INDEX() function searches *finaldiagnosis* for the string METAS and gives the position number of the start of the first instance of that string. Using the root METAS covers all variants of the words indicating METASTATIC and will capture metastasis cases even when there are misspellings, such as METASTAIC or METASTATISTIS. Once the position of the METAS string is found, the SUBSTR() function can be used to return a 'blurb' of the surrounding text and understand metastasis in context. For example,

```
DATA tmp;
      SET figure2;
      loc=INDEX(finaldiagnosis,"METAS");
      if loc>0 then blurb = SUBSTR(finaldiagnosis, MAX(loc-10,1),
               MIN(100, length(finaldiagnosis)-MAX(loc-10,10)+1));
RUN;
```

The SUBSTR() function is only used here if loc>0, in other words only if the string “METAS” is found in the *finaldiagnosis* variable. The first argument in the SUBSTR() function is the variable name with the text to search. The second argument is the starting position of substring. Here, we wish to have 10 characters before the first instance of the string “METAS” therefore MAX(loc-10,1) is used. IF “METAS” is found within first 10 characters of *finaldiagnosis*, the substring begins at the start of the final diagnosis, which is position 1, otherwise it starts at 10 characters before the first instance of the string “METAS”. The third argument gives the length of the desired substring or how many characters should be captured past the position in argument 2. For this third argument, MIN(100, length(finaldiagnosis)-MAX(loc-10,10)+1))is used. This means that 100 characters should be captured. However, if 100 characters goes beyond the length of the string, then just give the remainder of the string = length(finaldiagnosis)-MAX(loc-10,10)+1). Figure 8 shows the result of running this set of commands on the unstructured final diagnosis text from Figure 2.

Obs	finaldiagnosis	loc	blurb
1	BREAST AND AXILLARY LYMPH NODES, RIGHT, MASTECTOMY WITH AXILLARY NODE DISSECTION (A): -INFILTRATING DUCTAL CARCINOMA (5.2 CM), PLEOMORPHIC TYPE WITH A MINOR COMPONENT SHOWING TUBULAR FEATURES (SEE NOTE AND SYNOPTIC REPORT) -MINOR COMPONENT OF DUCTAL CARCINOMA-IN-SITU WITH PAGETOID INVOLVEMENT OF THE DUCTS -ATYPICAL DUCTAL HYPERPLASIA (BLOCK A3), SCLEROSING ADENOSIS AND COLUMNAR CELL HYPERPLASIA -BENIGN NIPPLE -RESECTION MARGINS FREE OF CARCINOMA (2.5 FROM THE DEEP MARGIN) -THE INVASIVE CARCINOMA IS NEGATIVE FOR ER AND PR EXPRESSION AND EQUIVOCAL FOR HER-2/neu OVEREXPRESSION (SEE NOTE) -METASTATIC ADENOMARCINOMA (UP TO 6.5 CM) IN TWENTY-EIGHT OF THIRTY AXILLARY LYMPH NODES (28/30) , WITH EXTRANODAL EXTENSION -THE METASTATIC ADENOCARCINOMA IS NEGATIVE FOR HER-2/neu GENE AMPLIFICATION	597	E NOTE) -METASTATIC ADENOMARCINOMA (UP TO 6.5 CM) IN TWENTY-EIGHT OF THIRTY AXILLARY LYMPH NODES (2

Figure 8. Example of using INDEX() and SUBSTRING() function against final diagnoses text

The output in Figure 8 shows that the first instance of the “METAS” is in position, loc=597, and the blurb provides the surrounding text to that first instance of METAS. From the blurb, we can see that Metastatic Adenocarcinoma to the lymph nodes is indicated for this specimen.

TEXT MINER

Using INDEX() and SUBSTR() functions are very helpful in parsing the final diagnosis text but their results can be flawed. For example, there may be other words used to indicate metastasis or the substring may capture too little or too much information. SAS Enterprise Miner & Text Miner can be very useful for parsing the final diagnosis text.

The first step is to load the final diagnosis text data set as a Data Source node in SAS Enterprise Miner. We then create a diagram and attach a Text Parsing node to the data set node in order to analyze and pull apart the unstructured text. Next, a Text Filter node is added to categorize all of the words in the document (Figure 9). To review results, while on the Text Filter node we can click on the Filter Viewer ellipsis under Results in the lower left properties box.

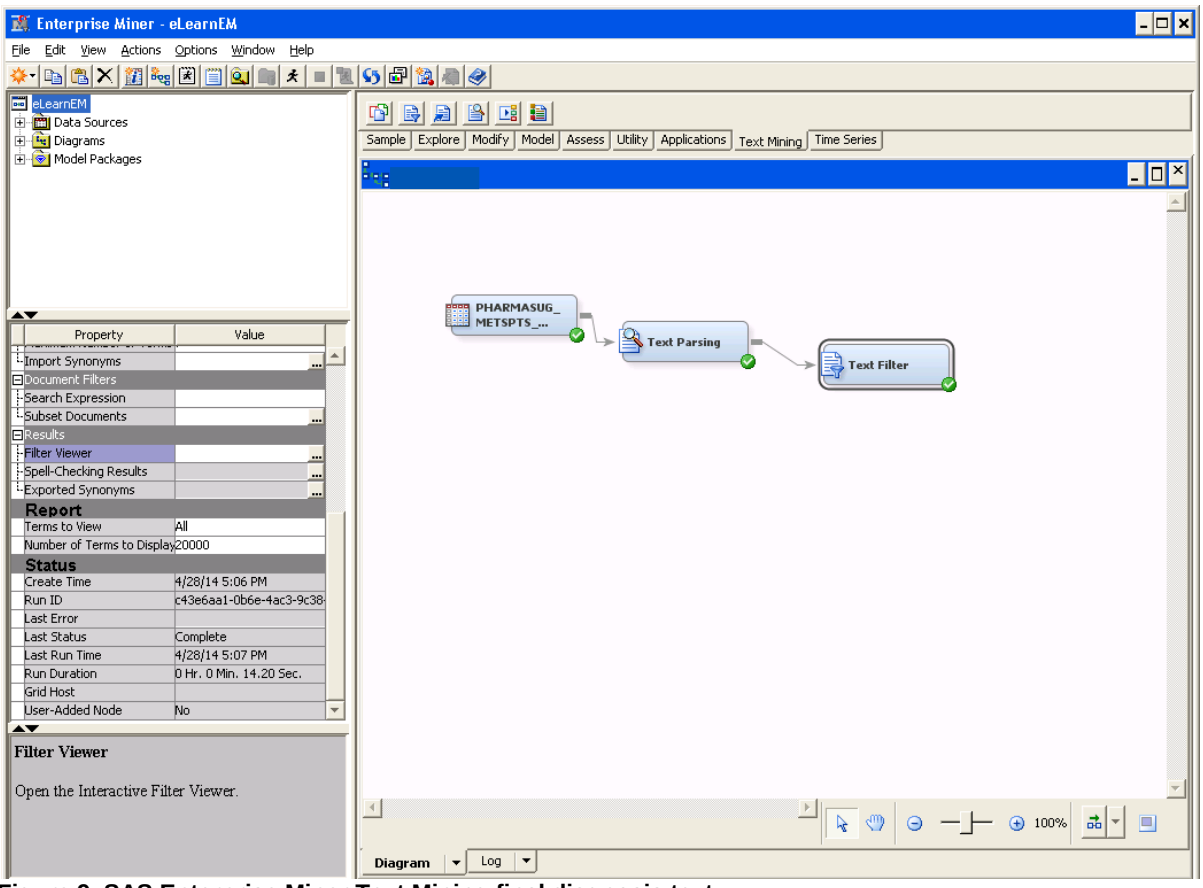


Figure 9. SAS Enterprise Miner Text Mining final diagnosis text

The results provide documents and terms as shown in Figure 10. All documents have been parsed into terms, and the frequency and number of documents where these terms occur are given. This text parsing and filtering is very useful in identifying other metastasis related words such as “macrometastasis”. It also allows for identification of more misspellings such as “metastasisleft breast”. In the search box, we can add all terms related to metastasis and click Apply. This runs a filter on the documents for only the terms selected. The TEXTFILTER_SNIPPET shows the text surrounding the phrase and bolds the phrase indicated such as “metastasis” or “metastatic”. Finally, if we highlight any one of the terms, for example “metastasis”, right click and select “View Concept Links” SAS gives a term map such as the one in the lower right corner of Figure 10. This process shows what words are closely related to metastasis. Some of these terms are useful and make sense such as “distant metastasis” and “regional lymph node metastasis” and others may not be as useful such as “40x” and “40x objective”.

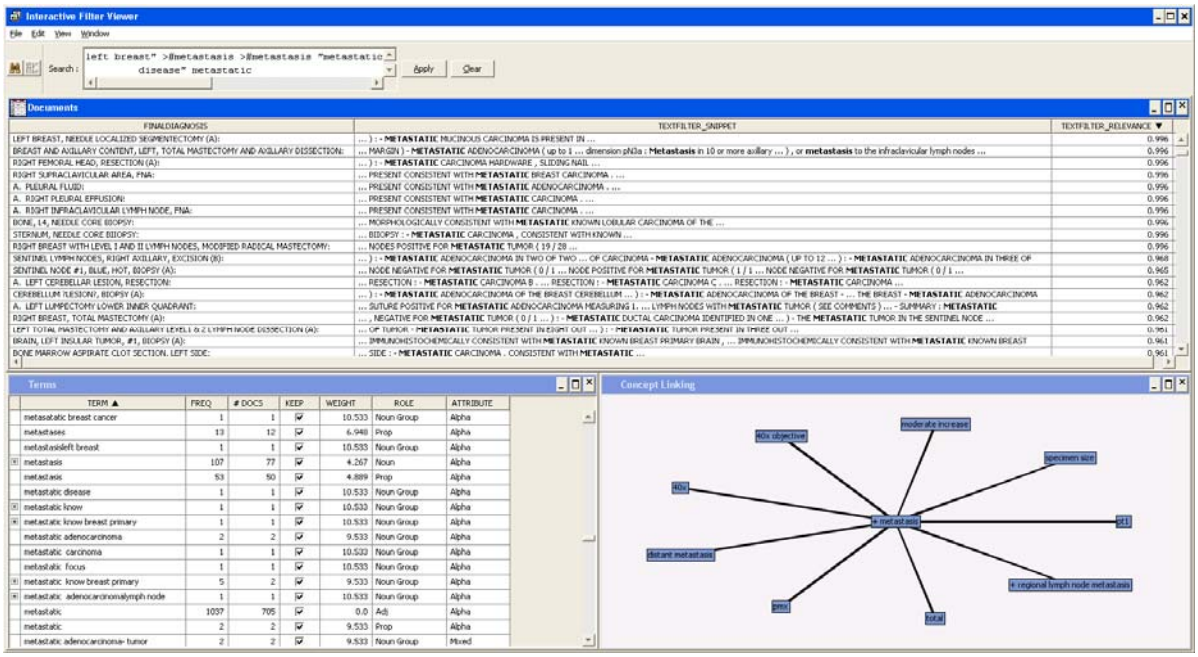


Figure 10. Results of text parsing node for metastasis in SAS Enterprise Miner

With a few extra steps, the filtered data set containing the text filter snippet can be saved to a permanent data set for later use with Base SAS. In the Enterprise Miner diagram, a Score node is attached to the Text Filter node and then a SAS Code node is attached to the Score node. In the SAS Code node, click on the Code Editor ellipsis (…) under Train (Figure 11) to add code.

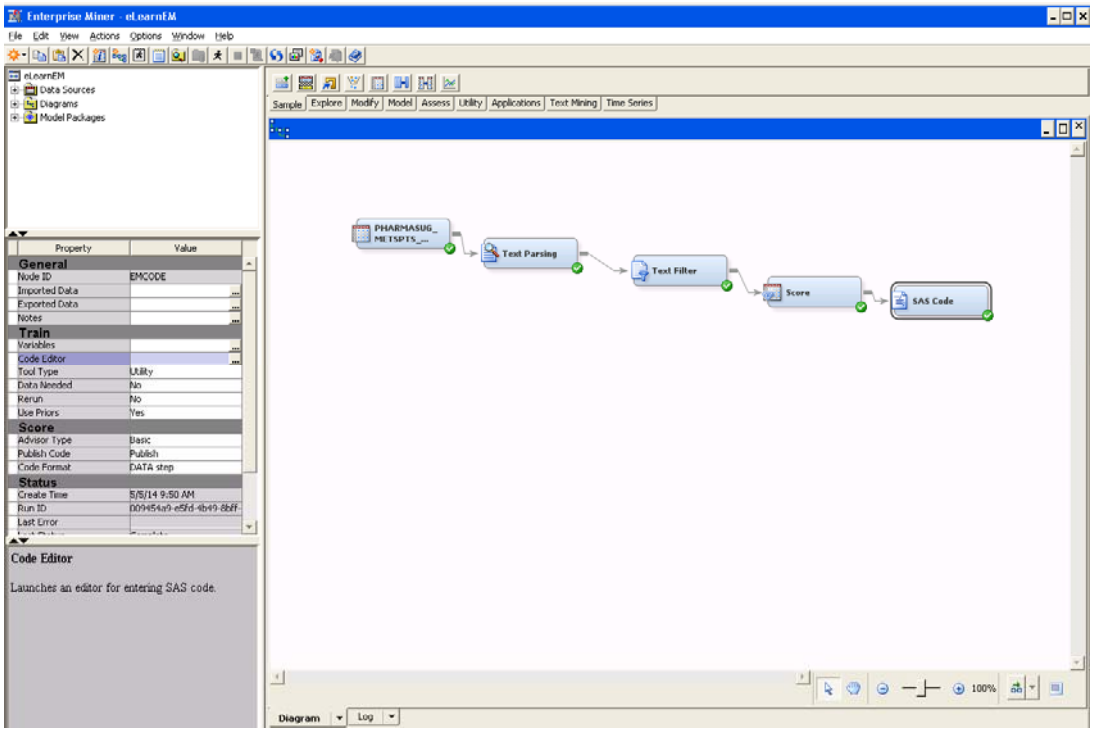


Figure 11. Attaching a score node and SAS code to create a permanent SAS data set

In the Training Code screen, the library name for the location storage location and data set names are specified. The simple code below saves the filtered data set to “C:\” as the permanent data set called PERM_METAS_DATA.

```
LIBNAME MYLIB "C:\";
DATA MYLIB.PERM_METAS_DATA;
      SET &EM_IMPORT_DATA;
RUN;
```

The &EM_IMPORT_DATA data set corresponds to the macro name SAS Enterprise Miner assigns to the filtered data to be exported. Figure 12 shows the SAS Code within the Training node. Once this code is entered, and the Score and SAS Code Nodes are run, the data will be permanently saved for later use in Base SAS.

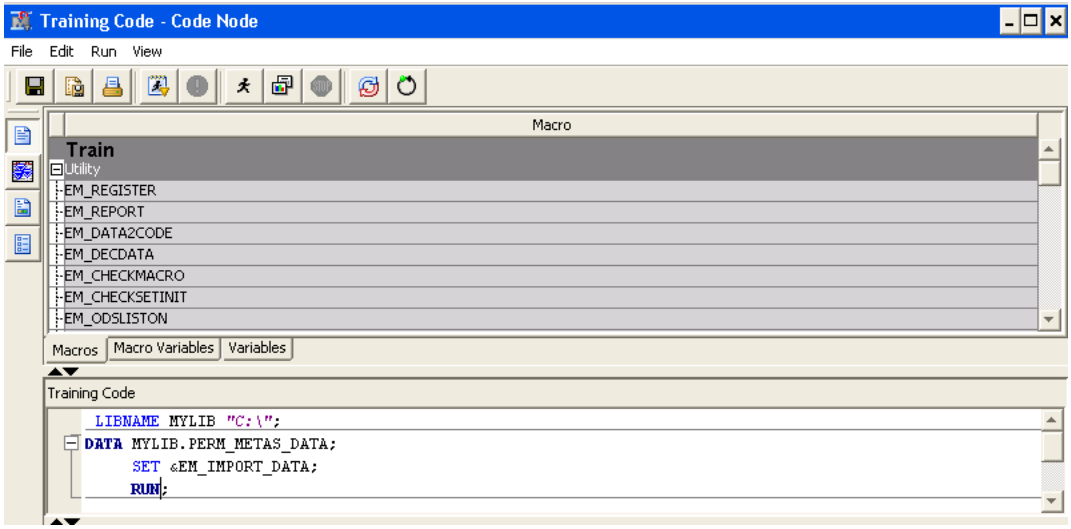


Figure 12. Training code screen for creating permanent SAS data set

Overall the SAS Enterprise Miner and Text Miner are useful in parsing text and providing guidance with what terms and misspelling are present in the final diagnoses of pathology reports. Using these tools gets us one step closer to the answer of whether or not the patient has metastasis. While the resulting data set requires a final review to screen out any remnant cases without metastasis that were missed due to context, it does have an initial filter applied so that there are fewer cases that the researcher would need to review.

CONCLUSION

We present several useful methods of how to handle unstructured pathological data. Using synoptic report data, string searches, or text mining tools can all be useful in simplifying the abstraction of the cumbersome pathology report. While these methods are simple they provide some of the most critical information that is needed to define response or predictor variables for analysis without the work of manually entering the data into a database. We can use synoptic worksheet data to define variables of interest at the time of surgical resection and combine this information with the results of our string searches to see which patients developed a distant recurrence, or even other sources of information that have been entered into our Cancer Registry. For future directions, we plan to examine the use of SAS Context Analytics in providing contextual meaning to the cancer diagnoses and other features in the pathology report that would not be easily found otherwise.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the first author at:

Name: Rebecca Ottesen
Enterprise: City of Hope
Address: 1500 Duarte Road
City, State ZIP: Duarte, CA 91010
Work Phone: (626) 256-4673
E-mail: rottesen@coh.org