Paper SAS165-2014

Extracting Key Concepts from Unstructured Medical Reports Using SAS® Text Analytics and SAS® Visual Analytics

J. Gregory Massey, Radhikha Myneni, M. Adrian Mattocks, Eric C. Brinsfield, SAS Institute Inc.

ABSTRACT

The growing adoption of electronic systems for keeping medical records provides an opportunity for health care practitioners and biomedical researchers to access traditionally unstructured data in a new and exciting way. Pathology reports, progress notes, and many other sections of the patient record that are typically written in a narrative format can now be analyzed by employing natural language processing contextual extraction techniques to identify specific concepts contained within the text. Linking these concepts to a standardized nomenclature (for example, SNOMED CT, ICD-9, ICD-10, and so on) frees analysts to explore and test hypotheses using these observational data. Using SAS® software, we have developed a solution in order to extract data from the unstructured text found in medical pathology reports, link the extracted terms to biomedical ontologies, join the output with more structured patient data, and view the results in reports and graphical visualizations. At its foundation, this solution employs SAS® Enterprise Content Categorization to perform entity extraction using both manually and automatically generated concept definition rules. Concept definition rules are automatically created using technology developed by SAS, and the unstructured reports are scored using the DS2/SAS® Content Categorization API. Results are post-processed and added to tables compatible with SAS® Visual Analytics, thus enabling users to visualize and explore data as required. We illustrate the interrelated components of this solution with examples of appropriate use cases and describe manual validation of performance and reliability with metrics such as precision and recall. We also provide examples of reports and visualizations created with SAS Visual Analytics.

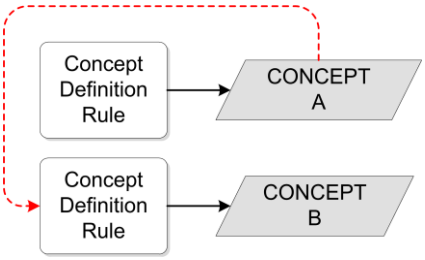
INTRODUCTION

Medical records contain tremendous amounts of valuable unstructured information often recorded in physicians' clinical notes or pathology reports. Historically, specially trained personnel have manually abstracted these data, but this is a slow, laborious process leading to reporting delays and sometimes erroneous conclusions (Midthune, Fay, Clegg, & Feuer 2005; Horm & Kessler 1986; Ries et al. 2004). Automated text mining techniques not only speed the abstraction process but also produce data that might be used for cohort discovery, outcomes research, and clinical decision support.

The SAS® Health and Life Sciences Research and Development team has created the SAS® Health Outcomes Analysis solution. This product extracts, standardizes, stores, and visualizes many types of electronic health record data by leveraging various elements of SAS proprietary technologies. A key component of the product is SAS® Enterprise Content Categorization. Enterprise Content Categorization is a rule-building engine that can be used for document categorization and named entity extraction (Albright, Punuru, & Surratt 2013). Its functionality includes tokenization, part-of-speech tagging, and coreference definition. Health Outcomes Analysis strictly uses the named entity extraction capability to abstract text strings and associate them with standardized terms, thereby linking multiple synonymous expressions and facilitating statistical analysis.

KEY CONCEPTS

Named entity extraction is a form of natural language processing that identifies words and phrases based on carefully designed rules. Enterprise Content Categorization groups these rules according to the concepts they are created to extract. These rules are developed using the SAS® Enterprise Content Categorization Studio user interface where concepts are organized as a taxonomy. The concepts within a particular taxonomy constitute an Enterprise Content Categorization project. The many different types of concept definition rules available in Enterprise Content Categorization provide the flexibility to train the model for more effective performance. They also allow a concept defined by one group of rules to be inserted within rules used to define another concept (Fig. 1). This recursive nesting of concepts within one another permits greater flexibility and renders the model less brittle when applied across different groups of documents.



`CLASSIFIER:scalp` ➔ CONCEPT_A

`CONCEPT:CONCEPT_A laceration` ➔ CONCEPT_B

Sample Text: There is a 2cm scalp laceration superior to the AS. The remaining scalp = NSF.

Figure 1. Enterprise Content Categorization allows predefined concepts to be recursively incorporated within rules that define other concepts. This example demonstrates how the word "scalp" (extracted by CONCEPT_A) is inserted into the concept definition rule for CONCEPT_B to extract "scalp laceration."

The disambiguation process clarifies confusion arising from words or phrases with more than one possible meaning. An example is the word "present," which might refer to the act of physically occupying a particular space and time or denote a gift on some special occasion. Skillful execution of the disambiguation process is essential to successful entity extraction. Enterprise Content Categorization provides the REMOVE_ITEM and NO_BREAK rules as two ways to disambiguate potentially confounding terms.

Precision and recall are classic metrics used to evaluate relevance in information retrieval systems (Carterette 2009). Precision (positive predictive value) measures the amount of retrieved information that is relevant, and recall (sensitivity) measures the amount of relevant information that is retrieved. The F Score is a harmonic mean, or weighted average, of the two values. To evaluate the success of the Health Outcomes Analysis Text Analytics Project, these statistics were calculated after analyzing various types of pathology reports using the following formulae.

$$Precision = \frac{Correctly\ Extracted\ Concepts}{(Correctly\ Extracted\ Concepts + Incorrectly\ Extracted\ Concepts)}$$

$$Recall\ = \frac{Correctly\ Extracted\ Concepts}{(Correctly\ Extracted\ Concepts + Concepts\ Not\ Extracted)}$$

$$F\ Score = \frac{2(Precision\ \times\ Recall)}{(Precision + Recall)}$$

BUILDING THE CONCEPT DEFINITION RULES

The Health Outcomes Analysis Text Analytics Project initially focused on extracting data from pathology reports for cancer patients. We began by identifying the tumor diagnosis, tissue site, and pathologic stage and grade. Since then the project has expanded to include 83 concepts containing over 320,000 concept definition rules. These rules gather data describing tumor size and behavior, the presence of genetic markers, and much more. Lexicons containing specific words mapped to a standardized term form the basic building blocks of the project. These standardized terms allow the project to group words or phrases that are synonyms, permitting scientists to conduct more effective analyses. These standardized terms also provide a link to various medical ontologies (e.g., SNOMED CT). In information science ontologies are used to associate concepts in a formalized structure based on logical interrelationships (Yu 2006). This association provides the means for analysts to roll up more sparsely populated data to a less granular level and fill a cohort with an appropriate sample size (Fig. 2). Due to the complexity of the biomedical domain, ontologies like SNOMED CT link terms along multiple paths. Consequently, a group of synonyms might not all map to the same standardized term. In the Health Outcomes Analysis data model we join these different standards to form additional linkages and create associations across multiple branches (Fig. 3).

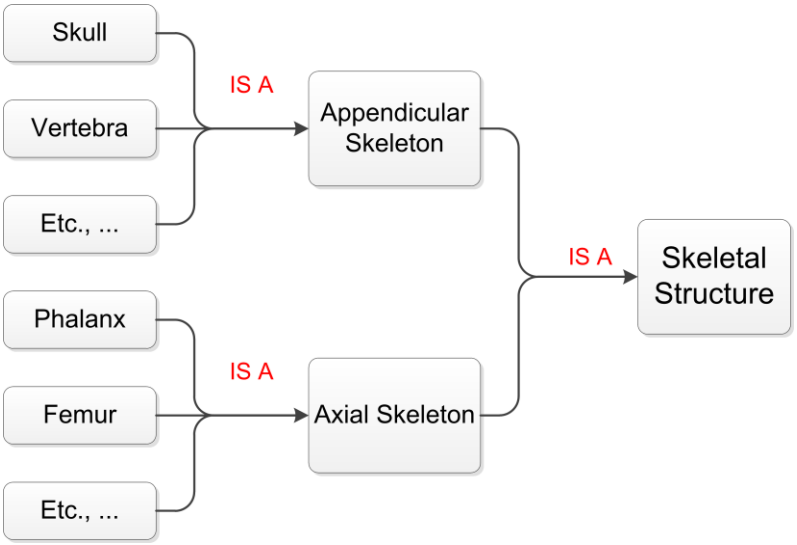


Figure 2. Mapping extracted text to a standard term allows analysts to aggregate sparse granular data (e.g., tumors affecting the femur and phalanx) at a higher level (e.g., all tumors affecting the axial skeleton).

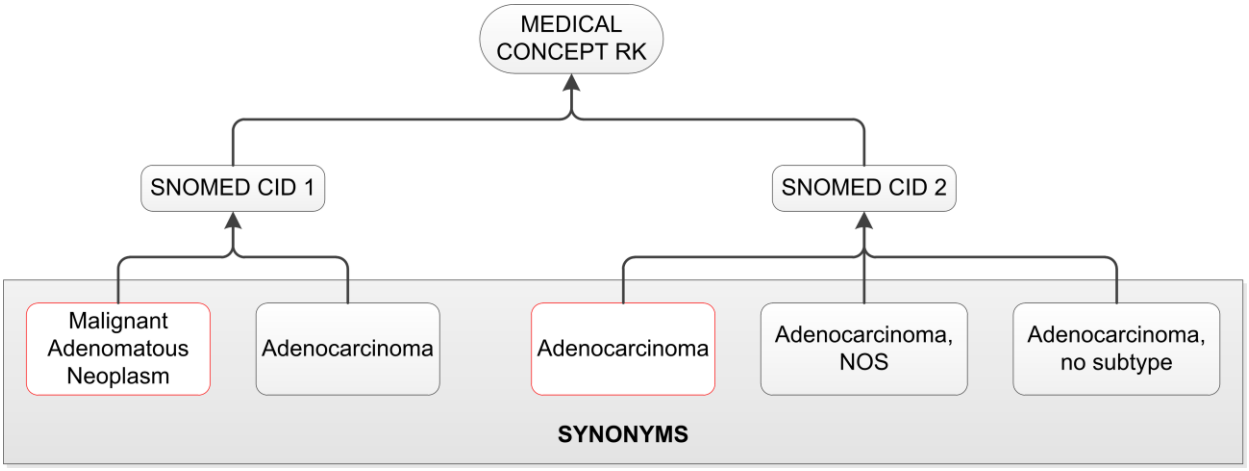


Figure 3. Multiple branches of complex biomedical ontologies might be populated by multiple synonyms. Each branch of synonyms maps to a different standard (in this case a SNOMED Concept ID) that then must be linked in the data model. For clarity, at the synonym level a preferred term can be selected (red box).

RETURNED INFORMATION

Enterprise Content Categorization's CLASSIFIER and REGEX rules provide the option to define returned information (Fig. 4). This information is not present in the text document but instead is added to the Enterprise Content Categorization output as a separate data element when an entity defined by the rule is extracted. In the Health Outcomes Analysis Text Analytics Project, we populate this function with a standardized term used in every rule targeting a synonym for the standard. By linking to this standard we are able to aggregate synonymous terms no matter how they are written in the text. As demonstrated above, biomedical ontologies provide the means to further aggregate these standard terms based on their semantic relationships to other entities.

In Enterprise Content Categorization the CLASSIFIER rule uses a comma as the delimiter between the rule and the returned information. If the string targeted by a CLASSIFIER rule contains a comma (e.g., Cary, NC), then the comma must be replaced with \c (e.g., Cary\c NC) to prevent truncation of the results. Instead of a simple comma, the REGEX rule requires the word INFO bounded by commas as the delimiter between the rule text and the returned information.

SAMPLE TEXT: “L great toe, proximal phalanx: closed complete non-displaced transverse fx. Remaining phalanges are unremarkable.”

CONCEPT DEFINITON RULES

```
CLASSIFIER:phalanx,phalanx
REGEX:phalanx\w+,INFO,phalanx
```

RESULTS

Concept Definition Rule	*Text Abstracted*	*Returned Information*
CLASSIFIER	phalanx	phalanx
REGEX	phalanx, phalanges	phalanx

Figure 4. In this example, “phalanx” is the standard term used for both rules. Phalanx and phalanges are synonyms for the same skeletal structure, so both concept definition rules return “phalanx” as the returned information.

RULES BUILT ON CONTEXT

Enterprise Content Categorization provides the CONCEPT_RULE and C_CONCEPT concept definition rules to extract entities based on the surrounding text (Fig. 5). The CONCEPT_RULE may use Boolean operators to define its parameters, but it uses these parameters to target a specific entity—not the entire string defined by those parameters. The C_CONCEPT is more restrictive, using the words most proximate to the target to define the context in which it should be extracted. This approach narrows the context that triggers extraction and improves the project’s precision.

SAMPLE TEXT: “The liver parenchyma is compressed by a cream-colored 3.5cm tumor. This liver tumor is solitary with well-defined margins.”

CONCEPT DEFINITION RULES

```
CONCEPT_RULE:(SENT,”_c{liver}”,”tumor”)
C_CONCEPT:_c{liver} tumor
```

RESULTS

Concept Definition Rule	*Text Abstracted*	*Context*
CONCEPT_RULE	liver	“Liver” and “tumor” occupy the same sentence
C_CONCEPT	liver	“Liver” immediately precedes the word “tumor”

Figure 5. These examples show contextual-based concept definition rules. The blue words in the sample text highlight the extracted terms. The green words are used by the concept definition rules to identify the context in which the target term should be extracted.

RULES FOR EXTRACTING FACTS

For the purposes of contextual extraction, a fact might be described as a text string defining a relationship between two concepts. The required concepts are comprised of words or specific text strings. Enterprise Content Categorization names these concepts *arguments*. The Enterprise Content Categorization predicate rules (PREDICATE_RULE, SEQUENCE) utilize arguments to identify and extract facts contained within a body of text (Fig. 6). Similar to the CONCEPT_RULE, the PREDICATE_RULE might use Boolean operators to define the relationship between the arguments. The SEQUENCE rule depends on the specific order of the supporting arguments, words, or tokens that identify a fact.

SAMPLE TEXT: "Sarcoma, liver, left lobe, removal – 7 cm in greatest dimension, involving canaliculi. Consistent with bone primary. Patient has a previous history of sarcoma in proximal right ulna."

CONCEPT DEFINITION RULES

```
PREDICATE_RULE:(tmr,size):(SENT,"_tmr{sarcoma}","_size{cm}")
SEQUENCE:(hx,tmr):_hx{history} of _tmr{sarcoma}
```

RESULTS

Concept Definition Rule	*Text Abstracted*	*Arguments*
PREDICATE_RULE	Sarcoma, liver, left lobe, removal – 7 cm	Sarcoma, cm
SEQUENCE	history of sarcoma	history, sarcoma

Figure 6. These examples show fact extraction using predicate rules.

THE ART OF DISAMBIGUATION

Enterprise Content Categorization provides several methods for disambiguating similar entities. Available concept definition rules include NO_BREAK and REMOVE_ITEM. The NO_BREAK rule prevents partial matches of text strings designed to be extracted as a single token (Fig. 7). The REMOVE_ITEM rule searches the document and prevents matches to its associated concept if there is also a match to a second concept (Fig. 8).

SAMPLE TEXT

Sentence 1: "The biopsy showed histologic findings consistent with basal cell carcinoma."
Sentence 2: "This type of carcinoma can typically be cured with surgical excision."

CONCEPT	**CONCEPT DEFINITION RULE**
TUMORLONG	`CLASSIFIER:basal cell carcinoma`
TUMORSHORT	`CLASSIFIER:carcinoma`
TUMORSHORT	`NO_BREAK:_c(TUMORLONG)`

RESULTS

Concept	*Text Abstracted*	*Sentence*
TUMORLONG	basal cell carcinoma	Sentence 1
TUMORSHORT	carcinoma	Sentence 2

Figure 7. The NO_BREAK rule prevents the TUMORSHORT concept from extracting "carcinoma" (red text) when applied to Sentence 1.

SAMPLE TEXT: "The biopsy showed histologic findings consistent with basal cell carcinoma. This type of carcinoma can typically be cured with surgical excision."

CONCEPT	**CONCEPT DEFINITION RULE**
TUMORLONG	`CLASSIFIER:basal cell carcinoma`
TUMORSHORT	`CLASSIFIER:carcinoma`
TUMORSHORT	`REMOVE_ITEM:(ALIGNED,"_c{TUMORSHORT}","TUMORLONG")`

RESULTS

Concept	*Text Abstracted*
TUMORLONG	basal cell carcinoma

Figure 8. The REMOVE_ITEM rule prevents the TUMORSHORT concept from extracting "carcinoma" (red text) when there is a match for "basal cell carcinoma" (blue text) in the same document.

Conflicts can also be resolved using available priority settings. Priority might be given to a concept as well as individual rules. Assigning a higher priority affects results when Overlapping Concept Matches is set to Best on the LITI tab of Project Settings in Enterprise Content Categorization Studio (Fig. 9). Increase the priority for an individual concept by entering a numeric value on the concept's Data tab (Fig. 10). The default priority is 10.

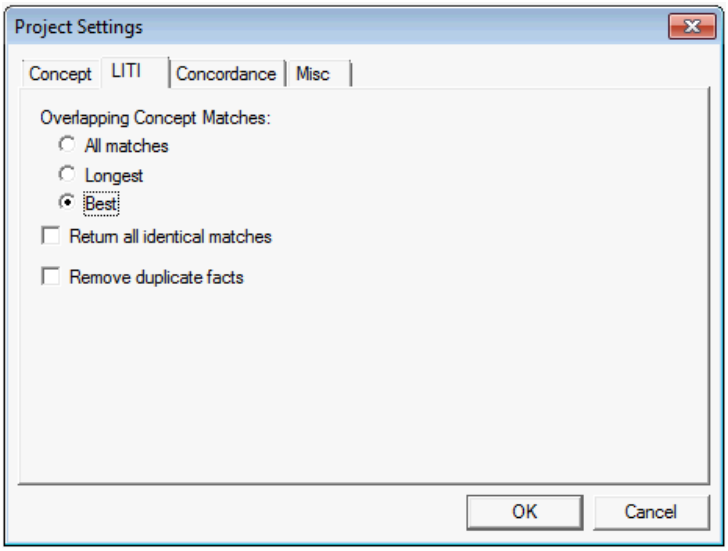


Figure 9. Selecting Best returns the longest match or the concept with the highest priority.

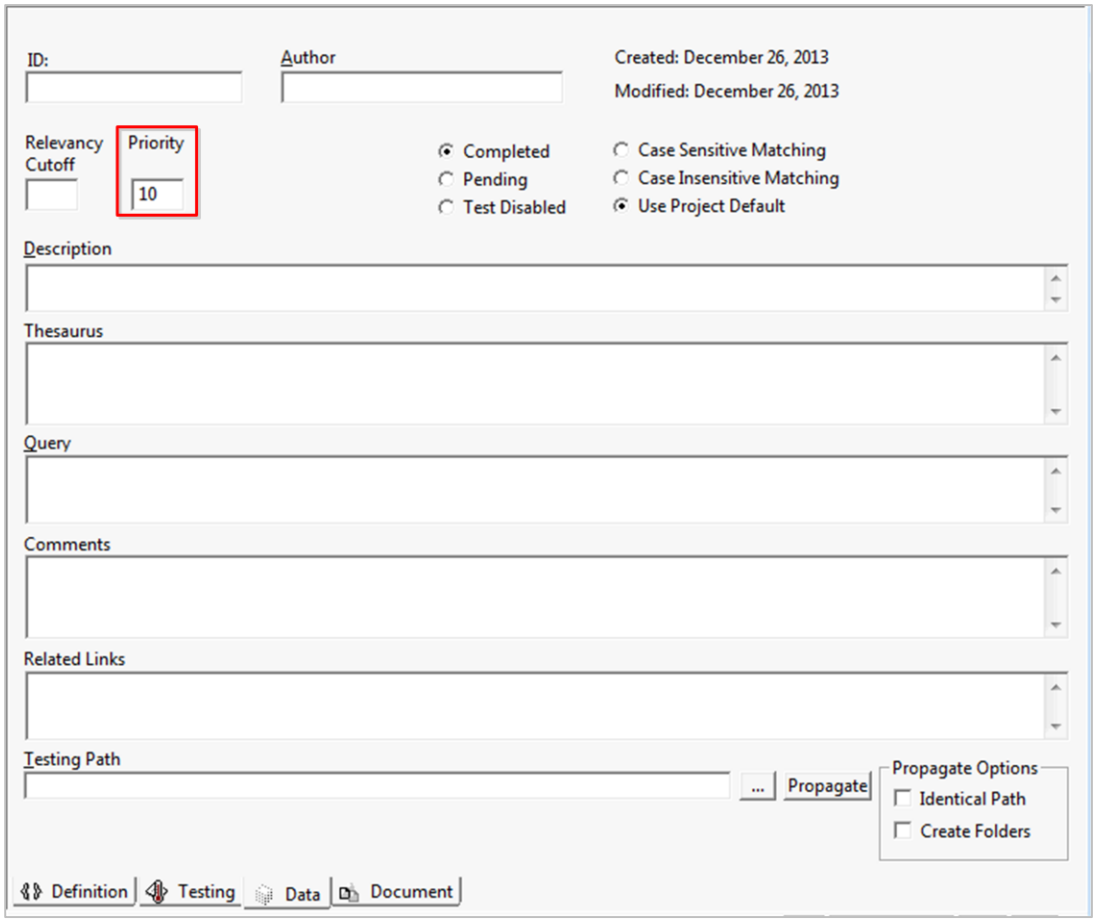


Figure 10. Set Priority on the concept Data tab.

Under some circumstances a single concept definition rule must be prioritized over the entire concept. For example, consider the concept TUMORLONG that contains the following CLASSIFIER rules:

```
CLASSIFIER:basal cell carcinoma
CLASSIFIER:chronic myelogenous leukemia
CLASSIFIER:non-small cell carcinoma
```

To ensure "basal cell carcinoma" is not captured as part of a longer text string extracted by another concept using a predicate rule, the CLASSIFIER rule could be replaced with the following:

```
CONCEPT:PRIORITY=15:basal cell carcinoma
```

In the Health Outcomes Analysis Text Analytics Project we also use the type of concept definition rule as a method to extract the correct entity from pathology reports. When Overlapping Concept Matches is set to Best or Longest, predicate rules consume the text bounded by their arguments rendering the text unavailable to other types of concept definition rules. Conversely, the context-dependent rules (C_CONCEPT, CONCEPT_RULE) do not. We use this difference to strategically differentiate concepts employing similar lexicons but in different contexts (e.g., separating a true diagnosis from a differential diagnosis).

AUTOMATING THE DOCUMENT SCORING PROCESS

Enterprise Content Categorization Studio provides the means for users to create concepts and their associated definition rules. Once the rules are compiled, the project is uploaded to a server that scores the corpus of text documents via a Java, Python, or SAS DS2 API. Java and Python results are output as a tab-delimited flat file. We used the SAS DS2 API to embed document scoring into a SAS® Data Integration ETL process that automated scoring and improved overall performance.

PRECISION AND RECALL STATISTICS

As part of the development process, we evaluated the Health Outcomes Analysis Text Analytics Project model using multiple types of pathology reports. We calculated precision, recall, and the F Score after manually annotating randomly selected reports and reviewing the output from the model (Table 1).

Report Type	Annotated Entities (N)	Precision	Recall	F Score
Cytopathology (Fine Needle Aspiration)	307	0.86	0.94	0.90
Cytopathology (Gynecology)	372	0.97	0.81	0.88
Dermatopathology	189	0.94	0.77	0.85
Surgical Pathology	558	0.94	0.85	0.89

Table 1. Precision, recall, and F Score statistics are shown from various pathology report types.

VISUALIZATIONS

SAS® Visual Analytics provides an excellent reporting platform for exploring and visualizing data extracted from medical reports. The capabilities of the software are beyond the scope of this paper, but below we have included two examples of reports that might be used to understand data derived using entity extraction. Because we lacked a sufficient sample of anonymized human pathology reports, we analyzed veterinary pathology reports for the purpose of creating these visualizations. Figure 11 is a report driven by the bar graph, demonstrating the frequency of tumor diagnoses segmented by behavior. Selecting one of the bars filters the remaining graphics to reveal data about tumors with either benign, malignant, or uncertain behavior. In this context, uncertain behavior refers to tumors with the potential to transform from benign to malignant. Data associated with the malignant category are shown in Figure 11. Figure 12 is a network diagram that explores relationships between tumor diagnosis, tumor site, and breed. Tumor size is represented by the thickness of the arrows linking the various elements.

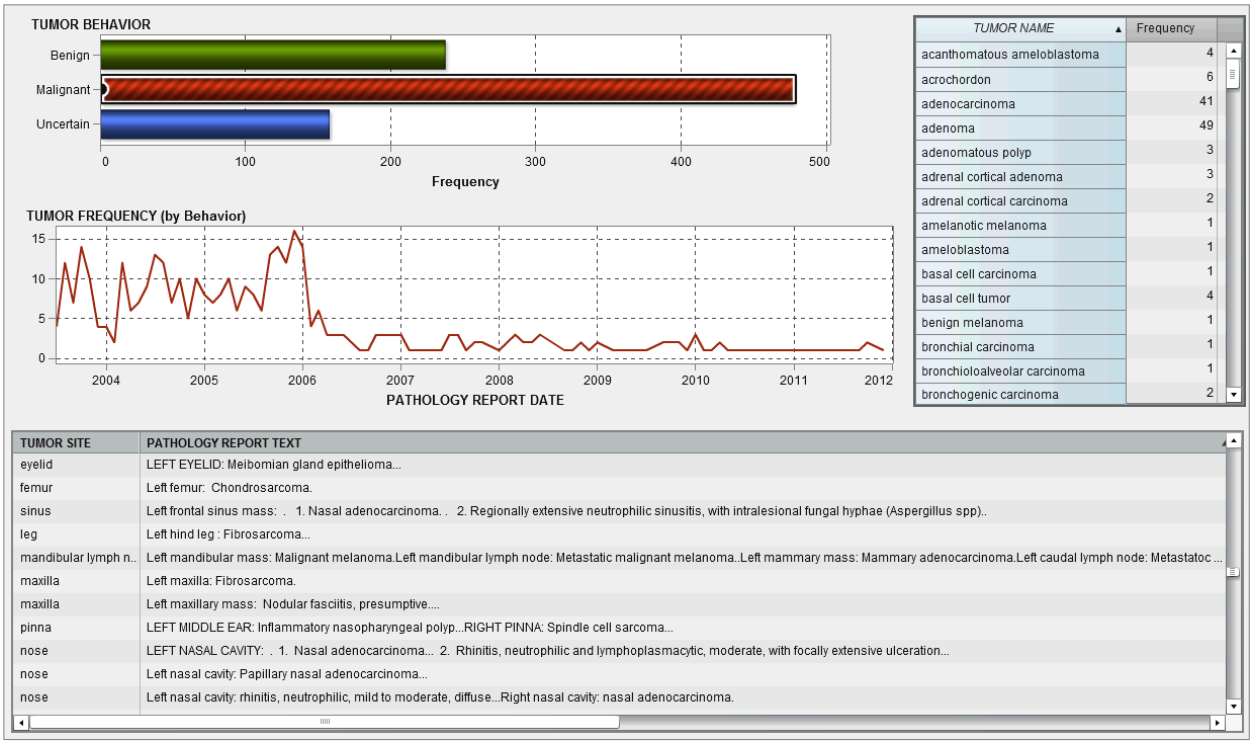


Figure 11. This SAS Visual Analytics report shows data derived from entity extraction of surgical pathology reports. Data are filtered by selecting one of the bars in the horizontal bar graph. This example illustrates data from malignant tumors.

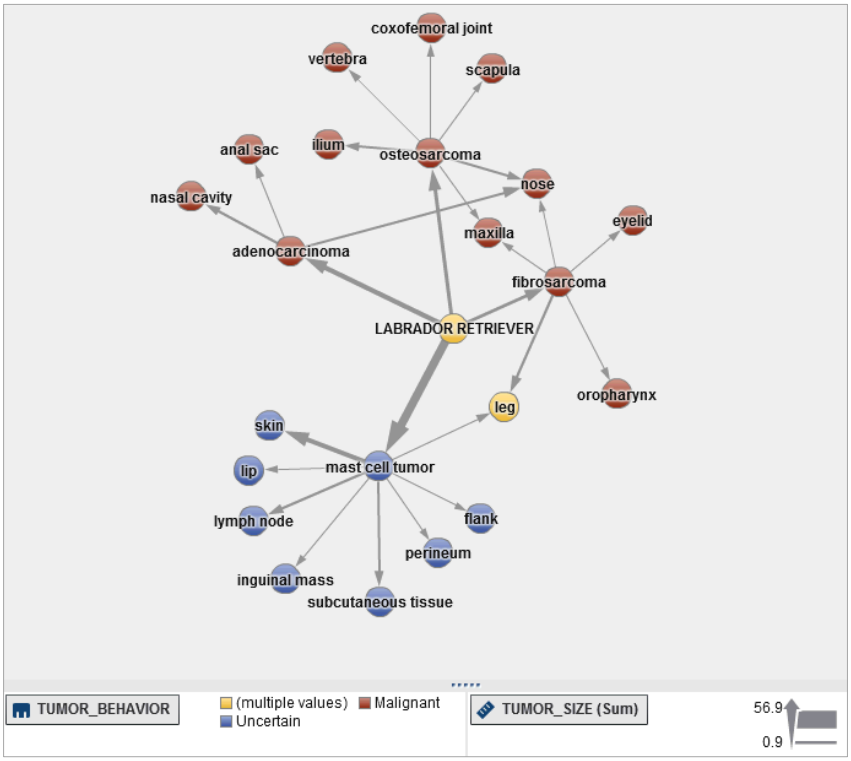


Figure 12. This example shows how SAS Visual Analytics can be used to explore relationships between tumor diagnoses and affected tissues. In this figure tumor behavior is color coded. Red nodes are associated with a diagnosis of adenocarcinoma, osteosarcoma, or fibrosarcoma (malignant behavior). Blue nodes are associated with a diagnosis of mast cell tumor (uncertain behavior). Yellow nodes are associated with both types of behavior. Arrow thickness represents relative tumor size.

CONCLUSION

The purpose of the Health Outcomes Analysis Text Analytics Project was to recover key elements of unstructured medical records and convert them to a structured format for data analysis. We accomplished this by linking extracted text to standardized terms that are aggregated or associated with similar or synonymous findings. We used SAS Enterprise Content Categorization to perform named entity extraction after creating custom concept definition rules. With the aid of subject matter experts and SAS Enterprise Content Categorization, the Health Outcomes Analysis Text Analytics Project achieved a high level of precision and recall when tested on multiple types of pathology reports.

REFERENCES

Albright R., J. Punuru, and L. Surratt. 2013. “Relate, Retain, and Remodel: Creating and Using Context-Sensitive Linguistic Features in Text Mining Models.” *Proceedings of the SAS Global 2013 Conference*. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings13/100-2013.pdf.

Carterette, B. 2009. “Precision and Recall.” In *Encyclopedia of Database Systems*, ed. L. Lui and M. T. Özsu, 2126–2127. New York, NY: Springer Science+Business Media.

Horm, J. W., and L.G. Kessler. 1986. “Falling Rates of Lung Cancer in Men in the United States.” *Lancet* 327 (8478): 425–426.

Midthune, D. N., M. P. Fay, L. X. Clegg, and E. J. Feuer. 2005. “Modeling Reporting Delays and Reporting Corrections in Cancer Registry Data.” *Journal of the American Statistical Association* 100 (469): 61–70.

Ries, L. A. G., M. P. Eisner, C. L. Kosary, B. F. Hankey, B. A. Miller, L. Clegg, A. Mariotto, E. J. Feuer, and B. K. Edwards, ed. 2004. *SEER Cancer Statistics Review, 1975–2001.* Bethesda, MD: National Cancer Institute.

Yu, A. C. 2006. “Methods in Biomedical Ontology.” *Journal of Biomedical Informatics* 39 (3): 252–266.

ACKNOWLEDGEMENTS

The authors are grateful to Saratendu Sethi, Janardhana Punuru, James Zawisza, Dan Zaratsian, and Lane Surratt for enduring numerous questions while sharing their time and technical knowledge of the Enterprise Content Categorization software. We would also like to thank Jeannette Bensen, M.S., Ph.D., and Leigh Thorne, M.D., for providing invaluable subject matter expertise. Finally, we must thank Keven Flammer, D.V.M., Jim Holland, Nicole Barker-Scoggins, and the North Carolina State University College of Veterinary Medicine for providing pathology reports used to create data for our visualizations.

CONTACT INFORMATION

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

SAS Institute Inc.
100 SAS Campus Drive
Cary, NC 27513
+1 919 677 4444 (fax)

Greg Massey	Radhikha Myneni	Adrian Mattocks	Eric Brinsfield
+1 919 531 3774 (phone)	+1 919 531 3736 (phone)	+1 919 531 3203 (phone)	+1 919 531 0213 (phone)
greg.massey@sas.com	radhikha.myneni@sas.com	adrian.mattocks@sas.com	eric.brinsfield@sas.com