Proceedings

First SUGCDSC Conference

August 31, 2018

Silver Spring Civic Building at Veterans Plaza
8525 Fenton Street, Silver Spring, MD 20910
Washington DC Area, USA

SAS User Group for Clinical Data Standards and Codes (SUGCDSC)
PO Box 1032, Rockville, MD 20849, USA
https://sugcdsc.org

Text mining protocols with SAS Enterprise Guide to create therapeutic library

Vidya Muthukumar, Advanced Clinical, IL

In a CRO or Pharmaceutical setting, one may have received emails from management asking if anyone had experience working in different indications such as Asthma and ABPA (Allergic bronchopulmonary aspergillosis) or Aesthetic or Cosmetic Dermatology products or worked with biomarkers to make a pitch for potential capabilities bidding projects.

While it may not be easy for programmers or biostatisticians to quickly recall protocol indications, it is likely that team members may overlook key experience, if they worked on different studies.

This paper demonstrates a method to text mine keywords from protocols in PDF format to search for key indications and output the data to a library as an index of therapeutic areas. Since SAS cannot read PDF files directly, PDF files need to be converted to .txt or .xls files using Adobe Export PDF. Text mining a PDF protocol requires the first few pages to be converted to excel spreadsheet which will be used to text mine and create a therapeutic library.

Using SAS EG, one can build a text-mined SAS hash object table to match key indications from protocols and output it to create a therapeutic library of studies. The object of this paper is to provide an efficient solution using text mining and SAS EG to create a therapeutic library to not only help streamline a company's capabilities pipeline but also potentially help with resource allocation to key therapeutic areas.

The Protocol is the key document that describes a complete plan of research activity in the framework of a clinical study[1]; specifically:

- the study objective(s),
- design,
- methodology,
- eligibility requests for prospective subjects and controls;
- intervention regimen(s),
- proposed methods of analysis of data;
- statistical considerations, and
- organization of the study.

The protocol usually provides the background and rationale for the trial, but these could be represented in other protocol referenced documents. When trying to create a library of protocols, the essential element to keep in mind is to text mine key words (such as therapeutic area, formulary, drug names, treatment, etc) from the protocols and create an index of key text-mined words as a base for setting up a therapeutic library.

However, many pharmaceutical companies or CROs lack an effective or efficient method to text mine protocols that are often saved in different formats. While the format of a protocol may vary from study to study, whether a final .docx version of the protocol is available or the study has a final /pdf version of the protocol, the absence of non-standardized formats for protocols can make it difficult to text-mine protocols to create a therapeutic library.

For purposes of this paper, the approach was to take final PDF versions of a sample of five study protocols and save them in a sandbox directory folder while testing setup of a therapeutic library.

When text-mining the protocol PDF documents, it is essential to categorize text into the following:

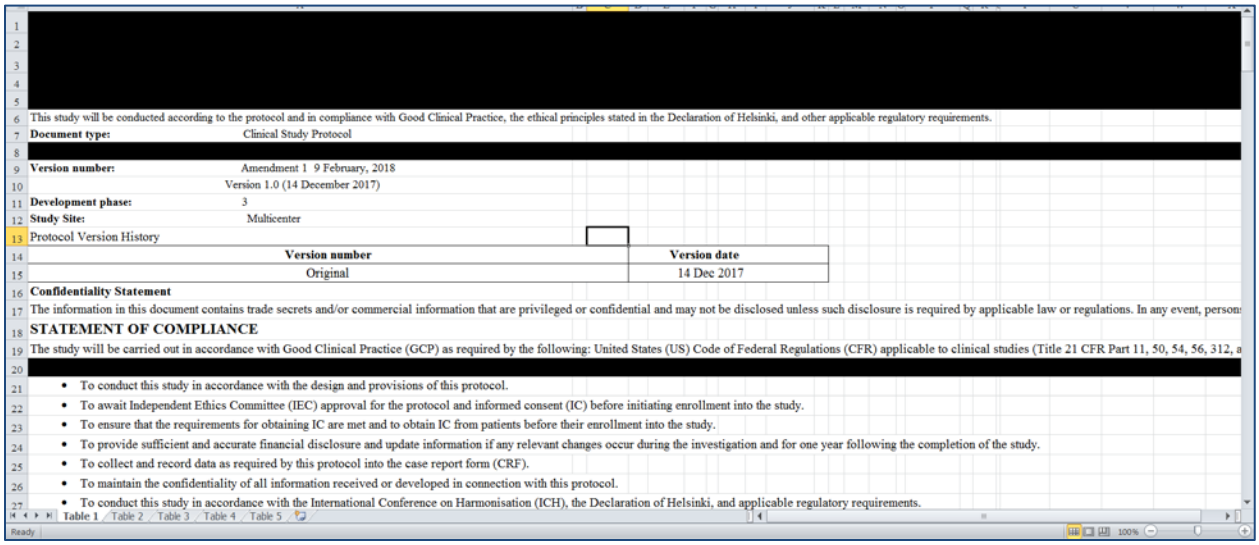
Table 1
Protocol Library Metadata Categories
Trial name
Project ID
Study phase
Sponsor
Therapeutic area
Intervention type
Condition under study
Description of trial
Reference link to primary manuscript

The Metadata categories will comprise a compendium of all keywords from key therapeutic areas and stored in an excel file which will be converted to a SAS dataset as Metadata library of keywords.

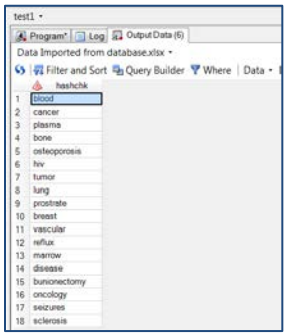
Some of the sources used to identify key therapeutic area from where data was collected included: https://www.centerwatch.com/clinical-trials/therapeutic-description.aspx[2]and www.clinicaltrials.gov[3].

Once keywords from therapeutic areas are identified, the webpage or downloadable data is exported as an excel file and saved as a dataset. This will be the underlying text-mined therapeutic metadata generated.

The next step is to text-mine the PDF protocols. Since PDF files cannot be easily text-mined, one of the options used to convert PDF files to an excel file was to use freeware trial version open-source application called https://smallpdf.com/. Once the PDF protocols are converted to excel files, the excel files are converted to SAS datasets. Screenshot below of a PDF protocol file converted to an excel file, which will be converted to a SAS dataset.



The therapeutic metadata table stored earlier will then be compared against the text-mined and converted PDF protocols→Excel→SAS dataset and a hash objects table will be created.



Using simple data step, proc sql, merge and SAS Hash Objects in SAS Enterprise Guide, once can identify matches between the text-mined PDF dataset and the Metadata library dataset to create a therapeutic library of the protocols.

Some of the challenges faced during text mining the clinical trials PDF protocols was the scope of creating the index by the therapeutic area using a list specific to organizational divisions (e.g., “cardiovascular device” or “cardiovascular megatrials”). The challenge was to convert the therapeutic area terminology to a list of standard medical specialties.

Another foreseeable challenge will be the ongoing maintenance for the protocol library within the organization. Maintenance tasks include tracking clinical trials as they near completion, communicating with the trial team to acquire the final protocol and associated documents, indexing documents and uploading the PDF protocols into the library. One can estimate that 10% of a full-time equivalent (FTE) employee’s efforts will be required on a continuing basis (actual time spent on maintenance will likely average 12.7 hours/month, or about 7.5% of an FTE).

Also, there has been some research to show that CROs and pharmaceutical companies spend significant operational resources on protocol distribution to all point-people where this information is needed. For multicenter trials, protocols and their updates are often first distributed to main members, and redistributed to a sub-network of affiliate centers. Research has shown that companies still using the paper format for protocols have to face safe storage or recycling/disposal costs. Also, paper-based protocol distribution methods cause many problems, such as delays in protocol information reaching regulatory bodies and front-line personnel; errors in document copying and distribution; and significant labor and materials costs.

In conclusion, if CROs and pharmaceutical companies use an electronic or web-based protocol maintenance framework, the conceptualization and operationalization of a therapeutic library can be a formidable feature to consider with a view to helping companies with economies of scale and also help them explore the option to use the therapeutic library as a tool for better resource allocation management.

References:

1.https://www.ncbi.nlm.nih.gov/pubmedhealth/
2.https://www.centerwatch.com/clinical-trials/therapeutic-description.aspx
3. www.clinicaltrials.gov