PhUSE EU Connect 2018

Paper TT07

The MetaDataHub - Integrating, searching and presenting study descriptive information

Elisabeth Nyman, Advanced Analytics Center, AstraZeneca, Gothenburg, Sweden
Kerstin Forsberg, Science and Enabling Units IT, AstraZeneca, Gothenburg, Sweden
Michail Doulis, Advanced Analytics Center, AstraZeneca, Gothenburg, Sweden

ABSTRACT

Identifying clinical studies fulfilling specific criteria and getting an overview can be difficult and time-consuming. Information sits in different documents and systems, and it may be challenging to find, integrate and display. We have developed the MetaDataHub concept for integrating study descriptive information. Additionally, it provides search functionality and presents the result in an interactive interface combining various visualizations to give a comprehensive overview. Basic study details originate from the company clinical trial management systems. Variable information comes by indexing study SAS data sets. Study design and endpoint definitions are extracted in a structured format from study documents. By integrating the information, you can combine searches on different types of information and quickly identify studies of interest. To further improve the search functionality, we explored how machine learning can be used to cluster variables and thereby improve identification of similar variables across studies.

INTRODUCTION

Clinical studies generate an enormous amount of data that we need to understand and navigate. The legacy data is a significant asset for the pharma companies and is important in many aspects, from regulatory interactions to reuse of data in exploring new medical opportunities. Critical in all aspects is to understand the studies and their set-up to interpret the data correctly and generate valid analyses. This work, to understand and prepare legacy data, is often very challenging and time-consuming. A recent example was presented at the PhUSE EU 2017 conference: "Converting Legacy Data to "ADaM-like" Data for Making Efficient Regulatory Response"[1]. Information about our clinical studies is spread across different platforms and to get an overview of studies where you want to combine information from multiple sources often requires extensive manual work. This includes identifying the sources, accessing the information and integrating it in a meaningful way before initiating the evaluation.

The purpose of the MetaDataHub concept is to simplify this work by gathering study descriptive information and provide comprehensive overviews where you can perform searches, compare studies, dig into details and be linked back to source documents and systems. By accessing the information directly from the sources, the MetaDataHub will reflect the content in the sources and doesn't need to be updated separately. Applying the concept will generate a reusable assembly of information, and not a one-off solution created for a single case. It will be valuable in any situation where you need to understand, compare or identify studies fulfilling certain criteria. One example of question where the MetaDataHub can be beneficial is to answer a question like "Do we have any Turbohaler studies that compare budesonide/formoterol to formoterol with pre-dose and post-dose FEV1?"

The MetaDataHub concept was developed during a data pooling project. To make an appropriate pooling you need to know your studies, understand the similarities and differences, and know what variables that correspond to each other in the different studies. The concept was developed as an internal crowdfunded innovation project and has also initiated additional innovation projects on the same theme.

In this paper, we describe the MetaDataHub concept, the four main types of information currently covered in the MetaDataHub concept and how they are displayed for the users. It also outlines the ongoing development and future directions. The examples are taken from a MetaDataHub created for respiratory studies.

METADATAHUB CONCEPT

The MetaDataHub concept covers four main types of information; basic study details, study design, variables and endpoint definitions (see Figure 1). The idea is to access the information directly from the source, not to create an additional storing area that needs to be maintained. However, in some cases the information is not available in a structured format useful for this purpose, hence we have explored how this information can be structured to be used here. The MetaDataHub presents an overview of the studies and enables the user to dig further into details and be linked to additional information. A fundamental piece is the search functionality. By integrating information of different kind, you can combine searches across the information types and quickly identify studies of interest. The information

is presented in an interactive interface that allows you to update filtering and directly see the effect as well as switch between visualizations to inspect the studies from different perspectives.

A challenge when integrating information from multiple sources is that different names (codes, numbers, acronyms) may be used for the same study. For example, D5899C00002, 2004-001183-4, NCT00206154 and SHINE are all synonyms referring to the same study and all systems may not use the same study identifier. To address this an AstraZeneca Trial ID (AZT_ID) has been introduced, in this case D5899C00002, that makes the connection point for the different codes, numbers and names in the source systems. The AZT_ID has been implemented company-wide[2].

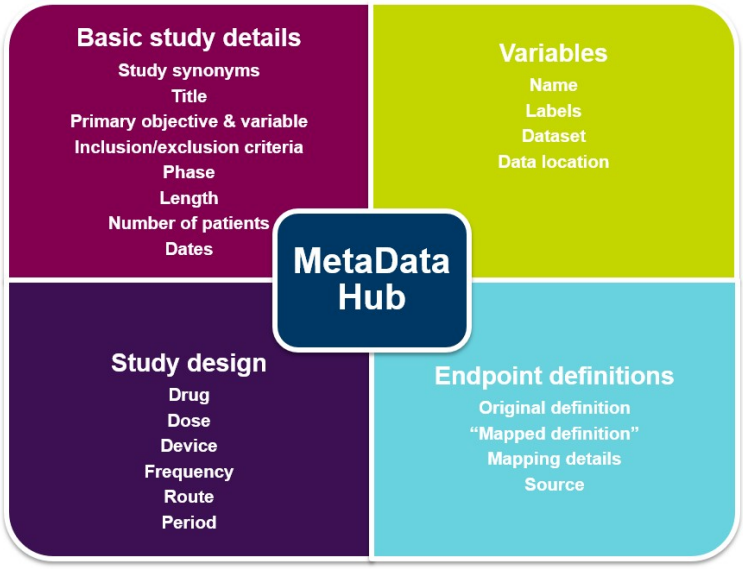


Figure 1. The four different types of information currently covered in the MetaDataHub.

BASIC STUDY DETAILS

Central in the MetaDataHub is the basic study details. These are basic study descriptive information like study synonyms, title, phase, primary objective, primary variable, inclusion/exclusion criteria, length, number of patients and dates. It also contains links to key documents. The sources for this information are internal clinical trial management systems and ClinicalTrials.gov accessed via the Aggregate Analysis of ClinicalTrials.gov (AACT)[3]. Figure 2 shows key dates and basic study details for selected studies.

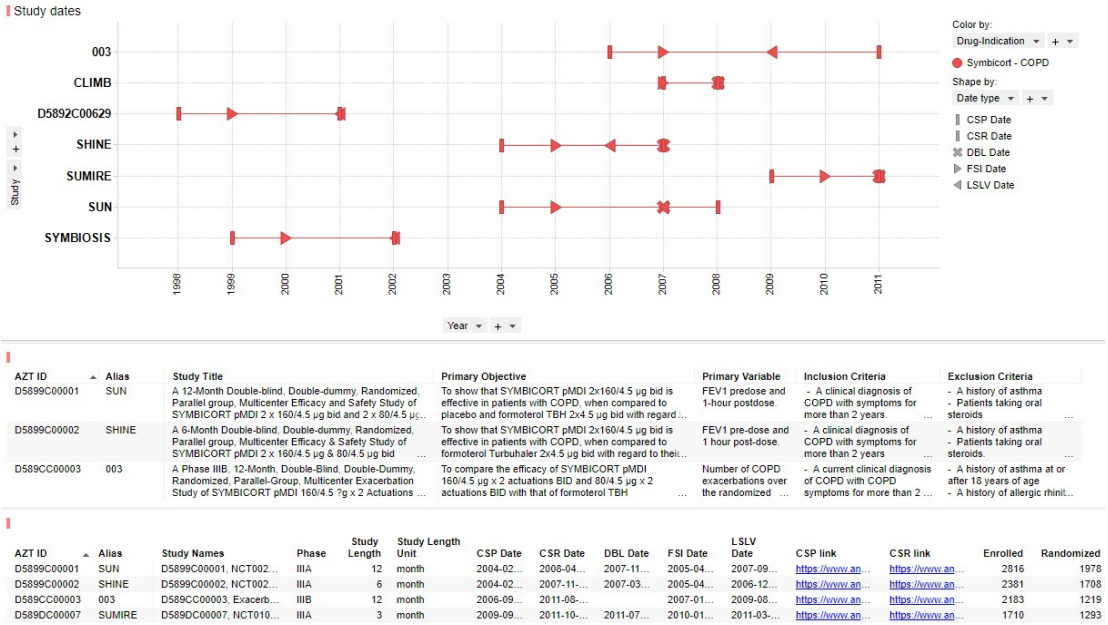


Figure 2. Visualization of key dates and basic study details.

STUDY DESIGN

To understand a study and its design the treatment arms and treatment periods are fundamental. This information is given in different documents and for more recent studies also in the Trial Design Domains in CDISC SDTM[4].

To be able to make controlled searches on the single elements and to create standard design visualizations the information needs to be divided on the individual components. Currently there is no directly suitable source containing this information. Available sources with structurally stored information usually captures the treatment arms as text strings, and the syntax for this text string varies between studies. Ideally you want this to be in the same format, e.g. “drug-dose-frequency”. By this, identical treatment arms across studies will be presented in the same way. We have extracted treatment information from study documents and organized the information in a structured and searchable format. Treatment arm information is divided on drug, substance, dose, frequency, route and device, and the different treatment periods (e.g. screening, treatment and follow-up) are captured together with their lengths. The treatment arms are then summarized in text strings to be used to label the arms. Figure 3 shows visualizations of the study length, treatment periods and treatment arms for selected studies.

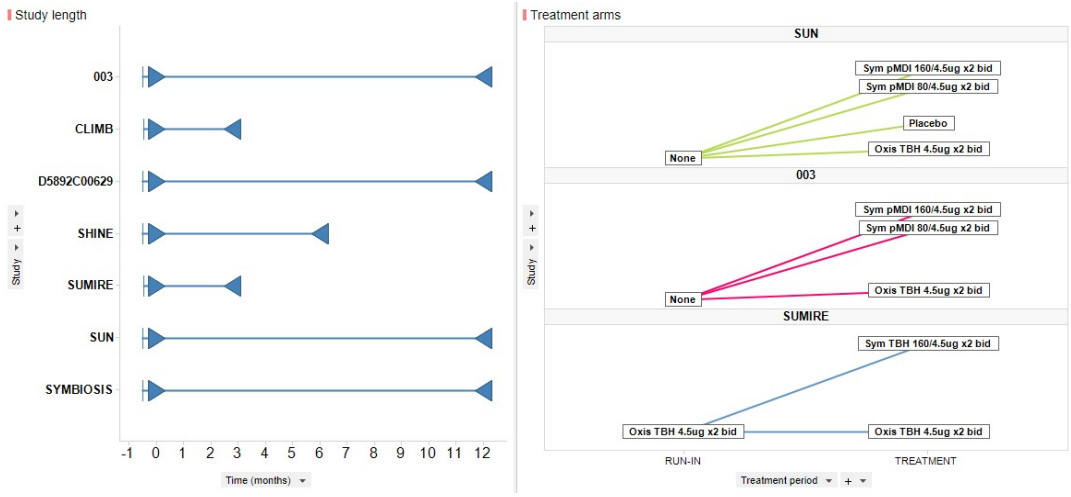


Figure 3. Visualizations on study lengths, treatment periods and treatment arms for selected studies.

VARIABLES

A vital component in the MetaDataHub is information about what data has been collected in the studies. To access this information the study SAS data sets have been scanned and indexed. The index contains information about the variable names and labels and covers both the variables given directly in the data sets and those given as coded values. The latter is the so called “tall and skinny” structure being used by CDISC standards e.g. lab data coded in the variables LBTESTCD and LBTEST. In addition to information about the variables the index also contains information about what data set the variables are available in and the data set location. Figure 4 shows reversibility variables and the data set they are given in for selected studies.

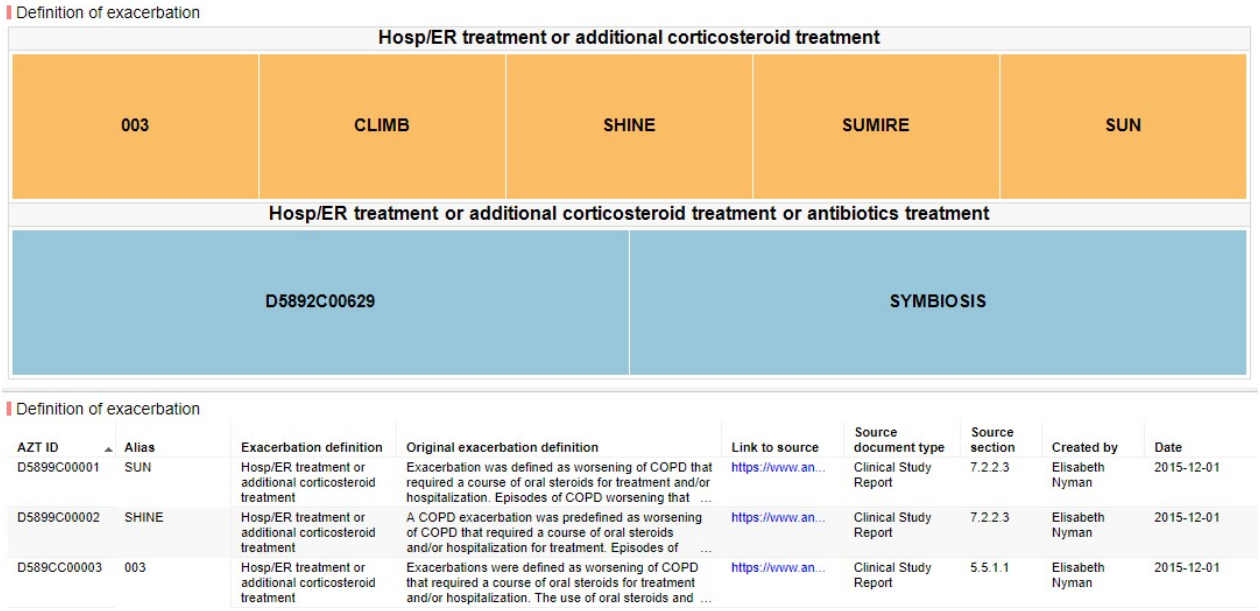
Variables & Datasets

AZT ID	Alias	Dataset Name	Variabel Name	Variable Label	File Path
D5892C00670	SYMBIOSIS	ADSL	FEV1REV	FEV1 reversibility at screening	//SEMLDSW2APP20...
D5899C00001	SUN	_PFT00	_LESS12	% of Reversibility at V1 Category	//SEMLDSW2APP20...
D5899C00001	SUN	_PFT00	_LESS12X	% of Reversibility at V1 Category	//SEMLDSW2APP20...
D5899C00001	SUN	_PFT00	_REVPCT	% of Reversibility at Visit 1	//SEMLDSW2APP20...
D5899C00002	SHINE	_PFT00	_LESS12	% of Reversibility at V1 Category	//SEMLDSW2APP20...
D5899C00002	SHINE	_PFT00	_LESS12X	% of Reversibility at V1 Category	//SEMLDSW2APP20...
D5899C00002	SHINE	_PFT00	_REVPCT	% of Reversibility at Visit 1	//SEMLDSW2APP20...
D589CC00003	003	ADPULM	BLREVRS	Baseline Reversibility (%)	//SEMLDSW2APP20...

Figure 4. Information about reversibility variables, what dataset they are available in and links to the data sets.

ENDPOINT DEFINITIONS

The definition of an endpoint may vary across studies. This is the case in e.g. respiratory studies where the definition of exacerbation has varied over time. To decide whether endpoints are comparable between studies it is crucial to understand how they are defined. In the case of exacerbation, the definition is relevant in many measurements, e.g. time to first exacerbation and number of exacerbations. As the definition is relevant for several endpoints it is more rational to relate it to the whole study rather than to each variable. To address this, we prototyped a solution applying a modified version of Nanopublication[5]. The definitions are usually given as text in a study document and the wording for the same definition may slightly differ across studies. To simplify comparison of the definitions and make it searchable we developed a vocabulary to map the definitions to. To be completely transparent of how this was done the mapped definition is stored together with the exact original definition, the source document, information about who did the mapping and when it was done. Figure 5 shows the definition of exacerbation used in selected studies together with details about the mapping.



AZT ID	Alias	Exacerbation definition	Original exacerbation definition	Link to source	Source document type	Source section	Created by	Date
D5899C00001	SUN	Hosp/ER treatment or additional corticosteroid treatment	Exacerbation was defined as worsening of COPD that required a course of oral steroids for treatment and/or hospitalization. Episodes of COPD worsening that …	https://www.an…	Clinical Study Report	7.2.2.3	Elisabeth Nyman	2015-12-01
D5899C00002	SHINE	Hosp/ER treatment or additional corticosteroid treatment	A COPD exacerbation was predefined as worsening of COPD that required a course of oral steroids and/or hospitalization for treatment. Episodes of …	https://www.an…	Clinical Study Report	7.2.2.3	Elisabeth Nyman	2015-12-01
D589CC00003	003	Hosp/ER treatment or additional corticosteroid treatment	Exacerbations were defined as worsening of COPD that required a course of oral steroids for treatment and/or hospitalization. The use of oral steroids and …	https://www.an…	Clinical Study Report	5.5.1.1	Elisabeth Nyman	2015-12-01

Figure 5. The definition of exacerbation used in selected studies. Together with the mapped definition the original definition given in the source document and details about the mapping is seen.

SEARCH FUNCTIONALITY

A key feature in the MetaDataHub is the search functionality. To be able to search across the different information types and identify studies fulfilling complex criteria, information needs to be integrated. This was done using TIBCO Spotfire[6]. By this, searches to identify e.g. studies with a certain length, treatment arm, endpoint definition and with data on a specific variable could be made. Figure 6 shows an example of how studies can be identified by applying different filters.

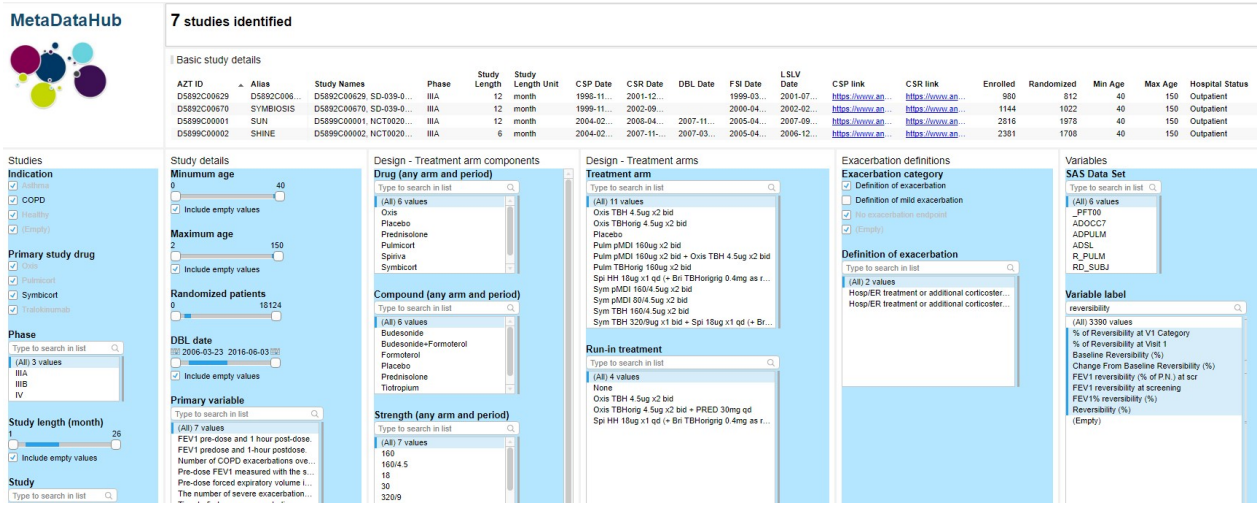


Figure 6. Studies fulfilling specific criteria can be identified by searching and filtering on different types of information.

PRESENTATION OF INFORMATION

To quickly get a basic understanding of the studies the information needs to be presented simple and clear. TIBCO Spotfire was used also for this purpose and we use a combination of various visualizations with traditional lists (Figure 2-5). These visualizations give an overview of the studies and you can also dig further into details. The solution is interactive which allows you to explore individual studies or sets of studies from different perspectives as the selections made in one visualization is coming through in the others. Additionally, the result of changed search and filtering is directly displayed.

ONGOING DEVELOPMENT & FUTURE DIRECTIONS

BASIC STUDY DETAILS & VARIABLES

During the last years AstraZeneca has initiated broad initiatives to simplify access to internal study descriptive information. One initiative covers large scale systematic integration of company clinical trials management systems and legacy registry. Today this information has been systematically integrated and additional information from various sources are continuously being added. This is done through integration with Enterprise Search and an API for accessing clinical study data in human and machine-readable formats. The integrated information can be accessed through a Google look-alike user interface in addition to be accessed by solutions like the MetaDataHub. Another large initiative that has evolved through several Proof of Concept projects covers extensive scanning and indexing of clinical study data to generate a searchable index for data locations and variable and data set information. These company-wide initiatives give a comprehensive foundation with study descriptive information and on top of this we can add additional information to generate a more holistic “multi-angled” view of our studies.

STUDY DESIGN

For the study design to be searchable and to create standard visualizations you want the information to be divided in its different elements. For the MetaDataHub we extracted treatment information from study documents and organized the information in a structured and searchable format. In a follow-up innovation project, we continued this work and developed a prototype to structurally capture this information to fit the MetaDataHub’s need as well as fit with the current standards. The prototype is built in several levels; substance information, formulation information, regime information, and finally this is summarized to treatment arms. Additional development is needed to bring this prototype to a full-scale proposal. Important in continued work is to align with CDISC, the ISO standard IDMP (Identification of Medicinal Products) and the emerging CTR2 project[7]. Additionally, the work to populate this needs to be naturally integrated in the daily work not to be burdensome.

ENDPOINT DEFINITIONS

The initial purpose with the endpoint definition component was to capture definitions key for the studies. However, the need for labelling a study is not limited to endpoint definitions. This principal, to extract core information from a large text, map it to a vocabulary and structurally capture information about the process, can be used for any kind of information. One example could be to annotate studies with sharing restrictions, i.e. a short annotation based on the informed consent. As the component can be broaden beyond endpoint definition a more suitable name for this component could be “study annotation”.

IMPROVED VARIABLE SEARCH BY MACHINE LEARNING

The ability to search across the different types of information is a key feature in the MetaDataHub concept. Searching on the variables names and labels does not always generate the result you are expecting, as the same or similar variable may exist in several studies, but the naming may differ as reversed order of words, abbreviations or synonyms may have been used. To improve the search of variables we have explored to what extent machine learning algorithms can be applied to categorize and cluster variables and thereby further simplify the search of similar variables across studies. As a Proof of Concept, we identified RAW datasets for a specific product from a broad index of clinical study data by applying machine learning and domain knowledge. Pre-processing step was performed focusing on emphasizing medical terms i.e. removing of non-medical content and enriching content according to known abbreviations. The labels of the variables were used as a text corpus and a document clustering method was applied. The distance (dissimilarity) among the variable labels was computed and the variables were clustered with Hierarchical clustering. An overview of the process is seen in Figure 7.

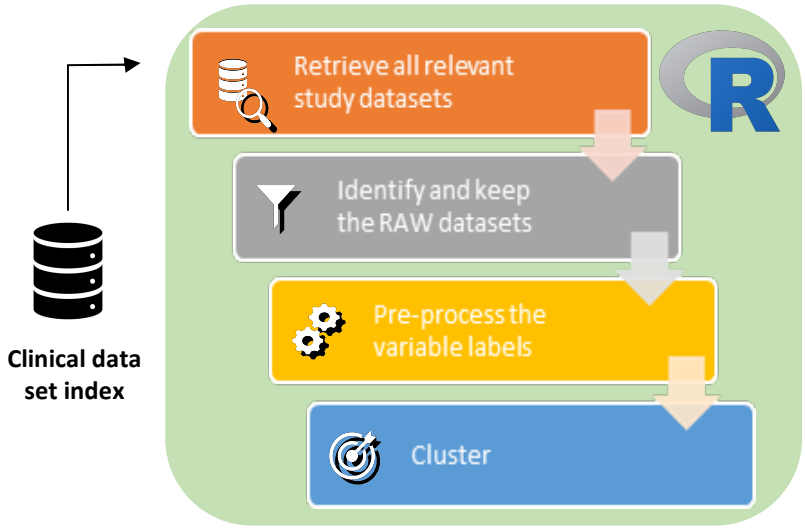


Figure 7. Schematic overview of the steps from data extraction to cluster output in the work to explore to what extend machine learning algorithms can be applied to categorize and cluster variables.

The resulting dendrogram, based on metadata on eight RAW data sets across 36 studies, shows that variables with similar medical content are clustered together or in adjacent clusters (see Figure 8). This demonstrates that machine learning can be used to generate overviews of related variables and can be used to improve searching for similar variables across studies. The method needs to be further developed and adjusted to fit the specific purpose of the clustering.

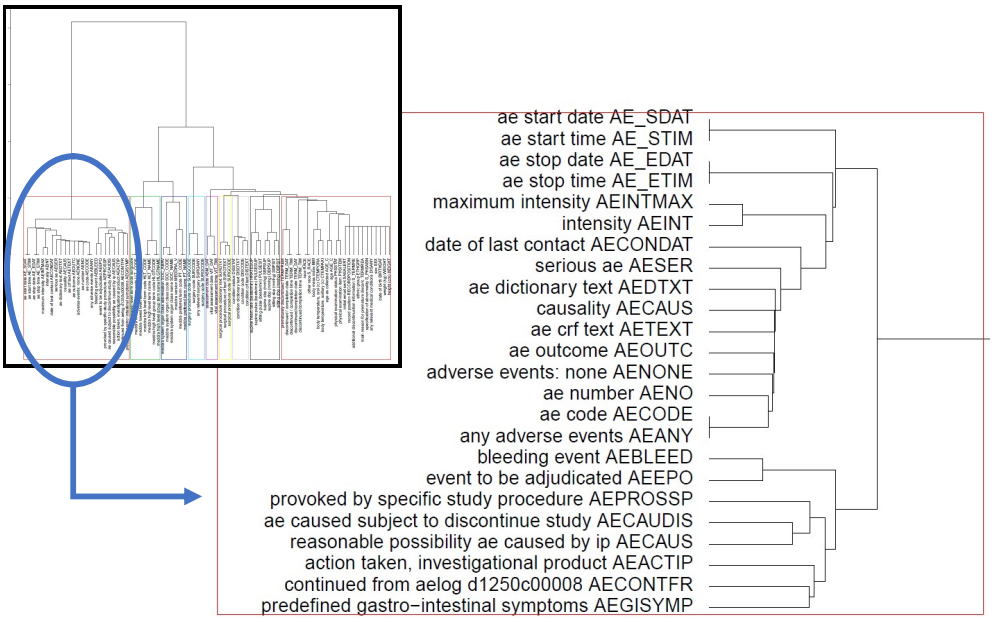


Figure 8. Resulting dendrogram from the variable clustering where a well-formed cluster for the adverse events related variables is seen.

TECHNICAL SOLUTION

The MetaDataHub concept covers integrating, searching and presenting information. In the development we have used TIBCO Spotfire for this. It has integrating capability, provides search functionality, has flexibility in creating visualization in addition to being interactive. However, the concept can be applied using other technical solutions. With growing MetaDataHubs it becomes more critical to choose appropriate technical solutions as it needs to handle

large scale data integration. Additional requirements also come when implementing improved search functionality, and to gain full flexibility in visualization design.

CONCLUSION

Pharmaceutical companies possess an enormous treasury with clinical data. As data may be re-used in different ways it can continue to generate value decades after the studies were conducted. Fundamental when working with clinical studies, legacy studies as well as ongoing and planned, is to understand the study design and study characteristics. For legacy studies it can be challenging to gain this understanding.

The idea with the MetaDataHub is to simplify access to study descriptive information to easily gain a basic understanding of the studies and their set-up. The MetaDataHub concept includes integration of information from multiple source. Searches can be done across the different information types and the information is presented in interactive visualizations, both overviews and detailed information. Currently four main types of information are covered; basic study details, study design, variables and endpoint definitions. However, the concept can be expanded to include additional information types e.g. data sharing restrictions (informed consent review) and biobank sample availability.

To further develop and to facilitate a broader implementation of MetaDataHubs we have run additional explorative activities. An improved information model to capture design information is needed and we have developed a prototype to structurally capture this. We have also investigated, and concluded, that machine learning can be utilized to improve search for similar variables across studies. Additionally, the model to capture endpoint definitions can be modified to capture any kind of annotation.

The MetaDataHub concept is aligned with the larger initiatives AstraZeneca recently has initiated to simplify access to study descriptive information. Learnings from MetaDataHub development is valuable in this work and the outcome of these initiatives will clearly simplify the understanding of our studies, through MetaDataHubs or similar systems.

We also see the value of the MetaDataHub concept in the changing pharmaceutical landscape. Beyond understanding internally conducted studies, we need, because of mergers and acquisitions, build an understanding of externally conducted studies. Similarly, upon a divestment, we need to make studies understandable for an external audience.

REFERENCES

1) Fältström, Bengt G. *Converting Legacy Data to "ADaM-like" Data for Making Efficient Regulatory Response*. PhUSE Annual Conference Edinburgh 2017. https://www.phusewiki.org/docs/Conference%202017%20DH%20Papers/DH03_paper%20NEW.pdf
2) Forsberg, Kerstin & Goude, Daniel. *Study URI*. PhUSE EU Connect Frankfurt 2018.
3) Clinical Trials Transformation Initiative, AACT Database. https://www.ctti-clinicaltrials.org/aact-database
4) Clinical Data Interchange Standards Consortium. https://www.cdisc.org
5) Nanopub.org. http://nanopub.org
6) TIBCO Spotfire. https://spotfire.tibco.com
7) Clinical Data Interchange Standards Consortium, CTR2 (Clinical Trial Registration) Project. https://www.cdisc.org/ctr2-project

ACKNOWLEDGMENTS

The authors would like to thank Nina Mian, Jim Weatherall, Lena Börjesson, Maria Benjegård and Bengt Fältström for reviewing the paper.

CONTACT INFORMATION

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

Elisabeth Nyman
AstraZeneca
Pepparedsleden 1SE-431 83, Mölndal, Sweden
Work Phone: +46-31-776 23 61
Email: Elisabeth.Nyman@astrazeneca.com