PhUSE EU Connect 2018

Paper TT09

Study URI

Kerstin Forsberg, AstraZeneca, Science and Enabling Units IT, Sweden
Daniel Goude, AstraZeneca, Science and Enabling Units IT, Sweden

ABSTRACT

A fundamental problem when integrating information on clinical studies is that we use synonyms (codes, numbers, acronyms) for the same study. For example: D3562C00096, 4522IL/0096, NCT00240331, 2004-001741-15 and AURORA all refer to the same study. To understand that we are referring to the same study, there is a need to cross match the same study across different systems. A Uniform Resource Identifier (URI) identifies an entity/resource, at a global scale simplifying integration of information. By assigning URIs for clinical studies, we have integrated study metadata across various systems. We have also integrated study related information, e.g. links to study datasets, summaries of biological samples, reviews of informed consents and data sharing requests. This paper describes the opportunities, enablers and challenges of implementing Study URIs. We also describe efforts to implement open source code for the identifier logic used in data science tools.

INTRODUCTION

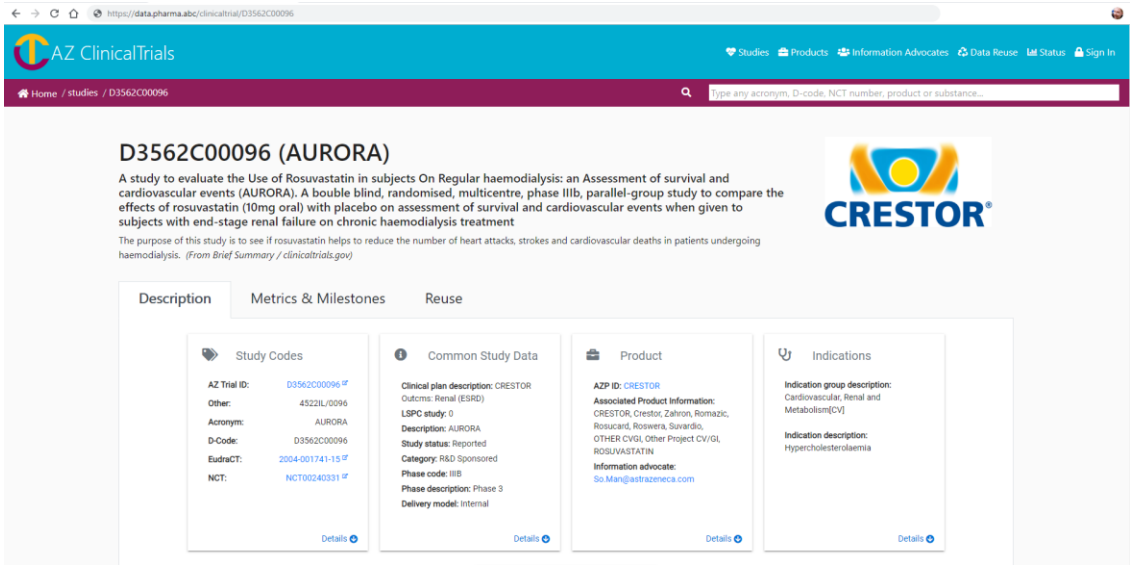
Clinical studies are key entities in clinical research sponsored by pharmaceutical companies, academia and other research institutions. Clinical research studies on human participants are designed to answer specific questions about biomedical or behavioral interventions, including new treatments and known interventions for further comparison and new indications. They often involve multiple centers in multiple countries, and can take years, from planning, conducting and analysis to publication. Furthermore, learnings from clinical study designs, data, and analysis results go beyond the publication of the primary analysis results. Clinical studies will remain key entities over time, as study data can be of renewed interest years after the study was conducted. In the changing pharmaceutical landscape, companies merge, and medical products are divested or acquired. Meanwhile the operating models within companies change, systems are upgraded, information is migrated, and code standards change. As a result, studies can potentially have multiple identifiers associated with them. Some identifiers are public, such as registry IDs and the identifiers assigned by regulatory authorities. Others are internal, such as the codes and acronyms used by sponsors. Some of the study codes are a consequence of different code standards. Different code standards can be seen as a representation of the history of the investigated medical product and of the sponsor company.

An example is that the codes D3562C00096, 4522IL/0096 NCT00240331, 2004-001741-15 and the acronym AURORA all refer to the same study. Sponsored by AstraZeneca, the AURORA study was a randomized, double-blind, placebo-controlled study investigating the use of rosuvastatin in the prevention of cardiovascular disease among patients undergoing chronic hemodialysis. AURORA, together with CORONA and JUPITER, were outcome trials conducted between 2003 and 2008 as part of the so-called GALAXY program.

Rosuvastatin (CRESTOR®) was developed by Shionogi under the code name S-4522 at the end of the 1980s. It was acquired by the British company Zeneca in June 1998 under the code name of ZD4522. Zeneca was formed in June 1993 by the merger of the pharmaceuticals and agrochemicals businesses of Imperial Chemical Industries (ICI). In 1999 Zeneca and the Sweden-based pharmaceutical company

Astra AB merged to form AstraZeneca. The code name for rosuvastatin in AstraZeneca became `AZD4522` and the study codes—the so-called D-codes—all start with `D356`. In 2015, seven years after it was reported, AURORA was one of the studies of interest in a collaboration project with the Montreal Heart Initiative (MHI) in Quebec, Canada, to search the genomes of up to 80,000 patients for genes associated with cardiovascular diseases and diabetes, their complications and treatment outcomes.

Lists of studies of interest, as in the MHI case, often include a mixture of new and old study codes, NCT-numbers, acronyms, and references such as “study 96”. To sort out these lists, we needed an easy to use service to get all codes, numbers, and acronyms. Hence, we needed a common way to identify a study across different internal and public databases.



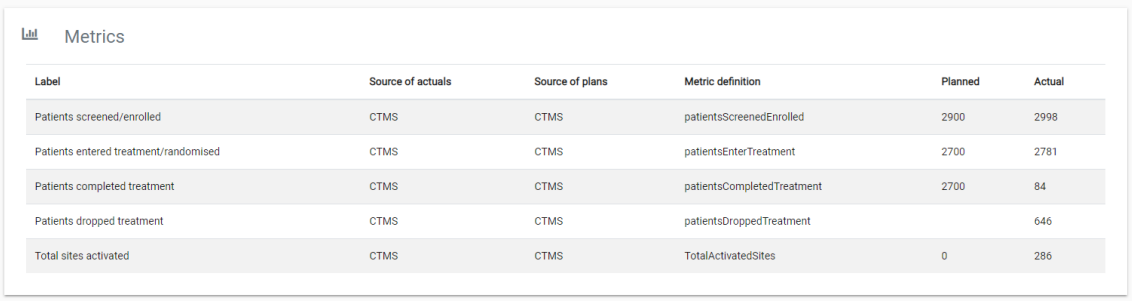
Study Codes

Code value	Code type	Source	Source column	Source ID column	Source ID	Code URL
D3562C00096	AZ Trial ID	MLCS-NG	AZT_ID	AZT_ID	D3562C00096	https://data.pharma.abc/clinicaltrial/D3562C00096
4522IL/0096	Other	MLCS	MS_STUDY_NAME	MS_STUDY_ID	24488	
AURORA	Acronym	MLCS	MS_STUDY_ACRONYM	MS_STUDY_ID	24488	
D3562C00096	DCode	MLCS	MS_STUDY_SYNONYMS	MS_STUDY_ID	24488	
4522IL/0096	Other	IMPACT	TRIAL_ALIAS_CODE	TRIAL_NO	10017	
D3562C00096	DCode	IMPACT	SPECIAL_NAME	TRIAL_NO	10017	
2004-001741-15	EudraCT	IMPACT	REGULATORY_REF	TRIAL_NO	10017	https://www.clinicaltrialsregister.eu/ctr-search/search?query=2004-001741-15
NCT00240331	NCT	IMPACT	REGULATORY_REF	TRIAL_NO	10017	https://clinicaltrials.gov/ct2/show/NCT00240331

Figure 1. Internal Study Page, with study code details,
for the AURORA study resolved via the Study URI

A Uniform Resource Identifier (URI) identifies an entity/resource such as a clinical study at a global scale and simplifies integration of information. By assigning AZ Trial ID / Study URIs to clinical studies we have integrated study information, not only study codes but also other metadata such as study title, indication, investigated medical product, metrics and milestones per study and per country. The AZ Trial ID acts as the key to reference important study information using API (Application Program Interface) endpoints. The Study URI for AURORA is: `https://data.pharma.abc/clinicaltrial/D3562C00096`. Such a Study URI consists of: 1) Global namespace: `https://data.pharma.abc` 2) Resource type: `clinicaltrial`, 3) AZ Trial ID: `D3562C00096`. Together these components form a shared, global identifier for the study. The

global Study URI is resolved inside the firewall to a Study Page URL (Uniform Resource Locator) to present the integrated information about the study, including an overview of all codes, acronyms and links (fig 1) as well as an overview of metrics on study level and on country level (fig 2).


Metrics

Label	Source of actuals	Source of plans	Metric definition	Planned	Actual
Patients screened/enrolled	CTMS	CTMS	patientsScreenedEnrolled	2900	2998
Patients entered treatment/randomised	CTMS	CTMS	patientsEnterTreatment	2700	2781
Patients completed treatment	CTMS	CTMS	patientsCompletedTreatment	2700	84
Patients dropped treatment	CTMS	CTMS	patientsDroppedTreatment		646
Total sites activated	CTMS	CTMS	TotalActivatedSites	0	286

Figure 2. Study Metrics for the AURORA study.

Using the URIs, we have integrated study related information from other sources, e.g. links to study datasets and summaries of biological samples. The URIs have also simplified the creation of new tables by moving information and extending the information from various Microsoft Excel files and SharePoint sites. This tabulated and organized information facilitates reuse of study data such as review of informed consents and data sharing requests (fig 3).

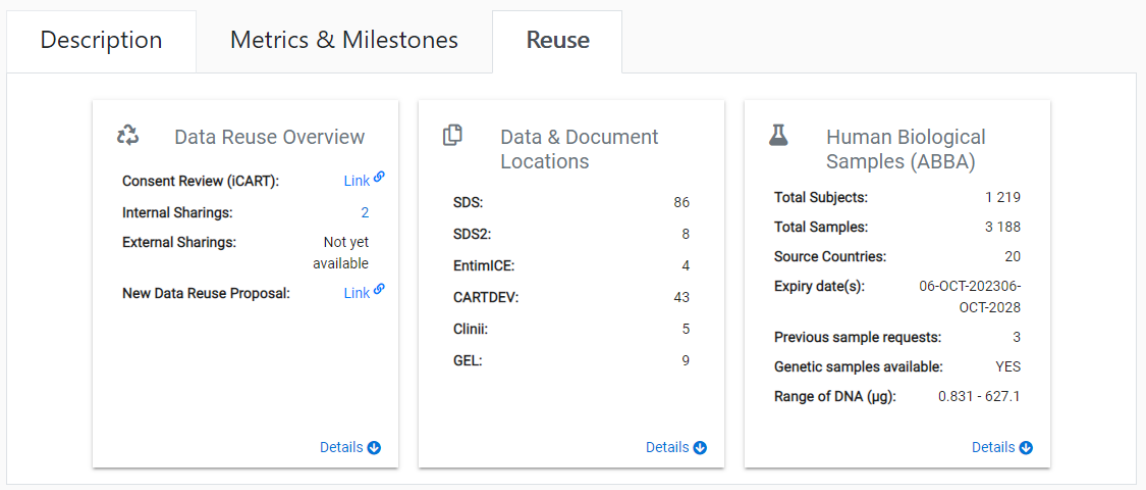


Figure 3. Reuse information for the AURORA study.

Assigning AZ Trial IDs / Study URIs to over 20.000 studies have exposed fine-granular, and quite tedious, challenges in how information about studies is represented and maintained in internal and public databases. In this paper, we describe some of these challenges, together with enablers and opportunities in implementing Study URIs.

CHALLENGES

With AURORA as an example, we illustrate below some of the detailed issues we have encountered with the broad scope of “all studies” to support reuse of data from already reported studies, incorporating new acquired studies, and preparing old studies for divestment. With a narrower scope, for example focusing on ongoing studies in support of clinical operations with overviews of study metrics and milestones, many of these issues will not be surfaced.

MULTIPLE CODES IN DIFFERENT COLUMNS

Our two main sources of study metadata are 1) IMPACT, AstraZeneca's Clinical Trial Management System (CTMS), and 2) MLCS, AstraZeneca's Master List of Clinical Studies.

As stated above: D3562C00096, 4522IL/0096, NCT00240331, 2004-001741-15 and AURORA all refer to the same study. To get these codes, we need the internal database keys. That is, 10017 in IMPACT, and 24488 in MLCS. For each of these sources, there are primary and secondary columns to hold study codes (fig 2). These columns have been used differently over time. In most cases, the main code is stored in the primary column, i.e. MS_STUDY_NAME in MLCS, and TRIAL_ALIAS_CODE in IMPACT. The main code according to the current study code standard in AstraZeneca is the so-called D-code e.g. D3562C00096, D3562C00098 and D3560L00030 with D356 as the substance code for rosuvastatin. However, in some cases other codes, such as the "old Zeneca code" e.g. 4522IL/0096, 4522IL/0098 and 4522US/0011 with 4522 as the substance code and IL as the code for international studies and US for United States studies, have instead stored in the primary column. This is the case for the AURORA and the JUPITER studies, while the CORONA study has the D-code in both the primary column and the secondary column (fig 4).

MLCS				IMPACT		
		Primary Column	Secondary Column		Primary Column	Secondary Column
MS_STUDY_ACRONYM	MS_STUDY_ID	MS_STUDY_NAME	MS_STUDY_SYNONYMS	TRIAL_NO	TRIAL_ALIAS_CODE	SPECIAL_NAME
AURORA	24488	4522IL/0096	**D3562C00096**	10017	4522IL/0096	**D3562C00096**
CORONA	24489	**D3562C00098**	4522IL/0098	10765	**D3562C00098**	**D3562C00098**
JUPITER	24272	4522US/0011	**D3560L00030**	10181	4522US/0011	**D3560L00030**

Figure 4. Differences in the use of study code columns for the three trials. AstraZeneca D-codes in bold.

The use of the special character "slash" (/) in the study code standard used by Zeneca has caused problems with some technologies, hence we see that it is often removed or replaced (e.g. 4522IL0096 and 4522IL_0096). The opposite is also experienced, often we see the use of "slash" (/) between the different parts in the AstraZeneca D-code standard (e.g. D3562/C00096).

ACRONYMS NOT CONSISTENTLY RECORDED

Using acronyms to refer to clinical studies is now common practice. The term "acronymania" has been coined to describe this explosion in popularity [ref 1]. Some acronyms—like HEART, SMART and HOPE—have been used for many studies. We also see sequential acronyms such as SOLO-1 and SOLO-2. Many of the acronyms are quite creative: AURORA, part of the GALAXY program, refers to "**A** study evaluating the **U**se of **R**osuvastatin in patients requiring **O**ngoing **R**enal dialysis: an **A**ssessment of survival and cardiovascular events".

Despite the popularity of acronyms, we have encountered issues with how they are being recorded in both internal and public databases. Internally MLCS has a column for Acronym while IMPACT does not. Instead the acronym can be a part of the Study title or Study description in IMPACT. The public ClinicalTrials.gov registry has an Acronym column per study, and a Sponsor Code table for the different codes. However, these are free text fields that are not always consistently populated. For example, the

AURORA acronym is not listed on ClinicalTrials.gov as an acronym, only in the title: “AURORA: Crestor 10mg Versus Placebo in Subjects With End-stage Renal Disease (ESRD)”. The acronyms for CORONA and for JUPITER are listed as secondary ids, not as acronyms (fig 5).

NCT_ID	SPONSOR	ACRONYM	CODE_VALUE	CODE_TYPE
NCT00206310	AstraZeneca		4522IL/0098	org_study_id
NCT00206310	AstraZeneca		CORONA	secondary_id
NCT00206310	AstraZeneca		D3562C00098	secondary_id
NCT00239681	AstraZeneca		4522US/0011	secondary_id
NCT00239681	AstraZeneca		D3560L00030	org_study_id
NCT00239681	AstraZeneca		Jupiter	secondary_id
NCT00240331	AstraZeneca		4522IL/0096	org_study_id
NCT00240331	AstraZeneca		D3562C00096	secondary_id

Figure 5. On clinicaltrials.gov, none of the three trials have the ACRONYM field populated.

IMPLICIT CONNECTION BETWEEN SPONSOR AND PUBLIC IDENTIFIERS

The two AURORA public identifiers (`NCT00240331` and `2004-001741-15`) are stored in IMPACT per study (using the internal database keys i.e. `10017` for `AURORA`) as regulatory references per country using the two specially added “countries”: `COUNTRY_CODE = 'EU' AND COUNTRY_NAME = 'European Union, EudraCT Number'`, `COUNTRY_CODE = 'CTR' AND COUNTRY_NAME = 'Clinical Trial Registry'`. This means that there is only an implicit connection between the sponsors and public identifiers.

The NCT-number is the identifier in the ClinicalTrials.gov (CT.gov) database. The format is “NCT” followed by an 8-digit number, e.g.: NCT00000419. The identifier in the European Clinical Trial database (EudraCT) has the structure `YYYY-NNNNNN-CC`. `YYYY` is the year in which the number is issued, `NNNNNN` is a six-digit sequential number, `CC` is a check digit. Information about studies can be accessed via URLs using these database identifiers. The study registry URLs for AURORA are:

```
https://clinicaltrials.gov/ct2/show/NCT00240331
https://www.clinicaltrialsregister.eu/ctr-search/search?query=2004-001741-15
```

In both ClinicalTrials.gov and EudraCT the sponsors can register sponsor identifiers. The ClinicalTrials.gov examples above (fig 5) for the three CRESTOR studies show differences in how the two code types, secondary_id and org_study_id, have been recorded. This means that for public registries the connections between public identifiers and the sponsor identifiers are not straightforward.

DUPLICATED STUDY RECORDS

We have over 20.000 studies listed in MLCS, with about half of them also represented in IMPACT. A new, planned study is “born” upstream as an activity in the project planning system. It is then transferred to IMPACT and from IMPACT to MLCS. For old acquired studies the processes have been different over the years. There is also a legacy of studies that were conducted before CTMS solutions were electronic. Today there is a reference in MLCS back to the internal database key in IMPACT, but this reference did not exist when we started the project to integrate study information.

Hence, we needed to match the study records across the two systems. While doing this, we identified anomalies such as the same study code being used for more than one study record, and variances of study codes non-compliant with the study code standards.

ACQUIRED STUDIES

Yet another challenge is posed by the many different code standards used for drug products, compounds and studies part of the mergers and acquisitions in the last decade. The AURORA study is a quite straightforward example with only two codes, while for other studies we have seen up to five different sponsor codes. Also, there is the bigger problem of ensuring that all old acquired studies are listed in internal databases, mainly set up to support the planning and management of new studies. This problem was highlighted in a recent blog post from the company TrialScope: "Many organizations struggle to produce a definitive list of trials that are their responsibility. For industry sponsors the challenge is often tracking studies associated with acquisitions, so we recommend creating a comprehensive list of all acquired products and businesses, as well as their acquisitions." [ref 2]

ENABLERS

In 2016 we started the AZ ClinicalTrials project to simplify the time consuming and error prone work of manually gathering information about clinical studies using many different system's user interfaces, Excel sheets and SharePoint sites. We built an integration layer to simplify these activities by providing a user interface, integration with Enterprise Search, and an API for accessing clinical study data in human and machine-readable formats (fig1-3).

Below we describe practical components enabling the assignment of shared, global and persistent identifiers for clinical studies as one part of the AZ ClinicalTrials project. However, it is important to highlight the "bigger picture": such a proactive project required the following high-level enablers:

- Clear study data entity ownership
- Domain knowledge
- Small, agile team – close collaboration with business partners
- Monitoring and continuous improvement

In referring to the same study across systems, we needed to agree on how to identify a study across systems. A Uniform Resource Identifier (URI) identifies an entity/resource, such as a clinical study, at a global scale simplifying integration of information. The URI is a generalization of the Uniform Resource Locator (URL). Together with html, the http-based URL is the key enabler behind the World Wide Web. Using http-based URLs, each information resource on the web is assigned a web address and a location. A URI identifies any resource, whether accessible on the web or not. "A URL is simply a URI that happens to point to a physical resource over a network." [ref 3]. The use of URIs in life science has been referred to as the identifiers for the 21st century [ref 4] and is key in applying the linked data principles for clinical study data [ref 5].

The Study URI for the AURORA trial is: https://data.pharma.abc/clinicaltrial/D3562C00096 Such a Study URI consists of three parts: 1) Global namespace: `https://data.pharma.abc` 2) Resource type: `clinicaltrial`, 3) AstraZeneca Trial ID (AZT_ID) e.g. `D3562C00096`. Together, they form a shared, global identifier for the study. The AZT_ID is the connection point for the different codes, numbers and acronyms in source columns stored in two internal systems for clinical studies: Source ID `10017` in IMPACT's TRIAL_NO, and `24488` in MLCS's MS_STUDY_ID. Illustrated above is a detailed view of study codes and acronyms for the AURORA study integrated through the AZ Trial ID (AZT_ID) and accessed via the URL https://data.pharma.abc/clinicaltrial/D3562C00096/codes (fig 1).

The key enabler behind our URIs is a small table: `STUDY`, in a proxy layer above the source systems representing the study entities. In this table, we store the assigned AZT_IDs and link them to internal database keys (fig 6).

AZT_ID	MLCS_MS_STUDY_ID	IMPACT_TRIAL_NO	AZP_ID	CREATION_TIME	DEPRECATED
D3560L00030	24272	10181	CRESTOR	2018-02-23 18:49:20.552307	False
D3562C00096	24488	10017	CRESTOR	2018-02-23 18:49:20.725281	False
D3562C00098	24489	10765	CRESTOR	2018-02-23 18:49:20.727927	False

Figure 6. The STUDY table rows
for the three trials

The linking process is regularly performed by a code type pattern recognition logic and a cross system matching logic. Given a vertical structure of study codes and internal database keys from different source columns, the output is the recognized CODE_PATTERN study code (fig 7).

AZT_ID	SOURCE	SOURCE_ID	SOURCE_ID_COLUMN	SOURCE_COLUMN	CODE_VALUE	CODE_TYPE
D3560L00030	IMPACT	10181	TRIAL_NO	SPECIAL_NAME	D3560L00030	ASTRAZENECA_D_CODE
D3560L00030	IMPACT	10181	TRIAL_NO	TRIAL_ALIAS_CODE	4522US/0011	ZENECA_CODE
D3560L00030	MLCS	24272	MS_STUDY_ID	MS_STUDY_NAME	4522US/0011	ZENECA_CODE
D3560L00030	MLCS	24272	MS_STUDY_ID	MS_STUDY_SYNONYMS	D3560L00030	ASTRAZENECA_D_CODE
D3562C00096	IMPACT	10017	TRIAL_NO	SPECIAL_NAME	D3562C00096	ASTRAZENECA_D_CODE
D3562C00096	IMPACT	10017	TRIAL_NO	TRIAL_ALIAS_CODE	4522IL/0096	ZENECA_CODE
D3562C00096	MLCS	24488	MS_STUDY_ID	MS_STUDY_NAME	4522IL/0096	ZENECA_CODE
D3562C00096	MLCS	24488	MS_STUDY_ID	MS_STUDY_SYNONYMS	D3562C00096	ASTRAZENECA_D_CODE
D3562C00098	IMPACT	10765	TRIAL_NO	SPECIAL_NAME	D3562C00098	ASTRAZENECA_D_CODE
D3562C00098	IMPACT	10765	TRIAL_NO	TRIAL_ALIAS_CODE	D3562C00098	ASTRAZENECA_D_CODE
D3562C00098	MLCS	24489	MS_STUDY_ID	MS_STUDY_NAME	D3562C00098	ASTRAZENECA_D_CODE
D3562C00098	MLCS	24489	MS_STUDY_ID	MS_STUDY_SYNONYMS	4522IL/0098	ZENECA_CODE

Figure 7. The Study Code Sources with recognized code types
for the three trials

The process is implemented in Pandas (Python Data Analysis Library) [ref 7] using Jupyter notebooks for prototyping [ref 8]. These data science tools have proven to be excellent in managing the challenges described above. We are now in the process of generalizing parts of the code base and publishing it as open source. One user-case scenario is to perform study code pattern recognition of ClinicalTrials.gov‘s sponsor identifiers.

OPPORTUNITIES

By assigning URIs for clinical studies, we have integrated study metadata across systems. We have also integrated study related information, e.g. links to study datasets, summaries of biological samples and implemented new tables and user interfaces for reviews of informed consents and data sharing requests. The screen shots in the beginning of this paper show examples from the user interface of the AZ ClinicalTrials web application available internally.

The most used feature is a “Google-look-a-like” search box with autocompletion features. The resource is open access to all AZ employees and enables users to gain a comprehensive overview of study information using the study acronym, thereby not needing to know any study codes.

The API and the underlying database views can also be accessed programmatically from other applications and data science notebooks. An example is the Metadata Hub, an application that consumes information from AZ ClinicalTrials, displaying the study metadata together with annotations and structured information on study design. This is an innovation project and is described in more details in another PhUSE Europe 2018 paper [ref 6].

Another example of programmatic access is a notebook using the study code look-up endpoint to cross reference studies listed internally with the studies listed in ClinicalTrials.gov accessed via the Aggregate Analysis of ClincalTrials.gov (AACT) database provided by the Clinical Trials Transformation Initiative (CTTI) [ref 9]. This gives us an opportunity to compare data across AstraZeneca studies and acquired studies.

Investigative work has been undertaken on study codes specified as metadata in a large document management system, and in data folder paths on a large file system for clinical data. As an example, the study code pattern recognition function in combination with the study code look-up API enables us to list Informed Consent Forms (ICFs) documents across countries per study.

CONCLUSIONS

The focus on reuse of data and the application of machine learning has led to a renewed interest in longitudinal access to clinical study data. This requires the systematic combination of information about clinical studies and clinical study data from existing systems. Doing this leads to discovery of information gaps not covered by the data sources themselves, and these gaps also need to be systematically filled. A fundamental part of this process is ensuring that identifiers are consistent.

In this paper, we have outlined opportunities, enablers and challenges of implementing Study URIs as high-quality identifiers. Using the AURORA study conducted over 15 years ago, and still of immense interest, we have described the challenges in how the different study codes are being applied and how we have overcome these to best utilize this study information.

Future development includes moving beyond application specific use of URIs in favor of one master entity, and establishing a cross application policy to assign URIs to more master and reference entities such as product, projects and indications. Furthermore, an enterprise infrastructure is required to resolve URIs for master entities both internally and externally.

We are also in the process to implement open source code for the identifier logic used in data science tools. Before the conference presentation in early November, we will provide a code repository with more technical details.

Improving data quality across an organization is not a one-off exercise, but something that requires a long-standing commitment and close collaboration with stakeholders.

REFERENCES

1. C. Labos, “The Alphabet Soup of Clinical Trial Acronyms ACRONYM: An Albeit Cool but Really Obtuse Naming Strategy for Medical Studies,” Medscape http://www.medscape.com/viewarticle/857543 (last visited Sept 2018)

2. Trial Sponsors Take Control NOW! Two Steps to EU Results Disclosure Compliance. Blog post by T. Wicks, TrialScope, https://resources.trialscope.com/trialscope-blog/trial-sponsors-take-control-now (last visited Sept 2018)
3. https://en.wikipedia.org/wiki/Uniform_Resource_Identifier (last visited Sept 2018)
4. McMurray et al., Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data PLoS Biol. 29;15(6) [PMID:28662064]
5. T. Williams and A. Oliva, Breaking the Mould: Clinical Trials Data as RDF, PhUSE Europe Connect 2017, https://www.phusewiki.org/docs/Conference%202017%20TT%20Papers/TT02.pdf (last visited Sept 2018)
6. E. Nyman and K. Forsberg, MetaDataHub - Integrating, searching and presenting study descriptive information, PhUSE Europe Connect 2018
7. https://pandas.pydata.org/
8. http://jupyter.org/
9. https://www.ctti-clinicaltrials.org/aact-database

ACKNOWLEDGMENTS

Maria Benjegård, Åsa Ström, So Man, Elisabeth Nyman, Ian Dix

CONTACT INFORMATION

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

Kerstin Forsberg
AstraZeneca
Pepparedsleden 1
SE-431 83, Mölndal, Sweden
Work Phone: +46-31-776 53 18
Email: kerstin.l.forsberg@astrazeneca.com