PhUSE 2017

Paper DS10

Moving Towards a Round World: Linking Multi-dimensional Clinical Metadata Across the CDISC Standards

Chris Decker, d-Wise, Morrisville, USA
Scott Bahlavooni, d-Wise, Morrisville, USA

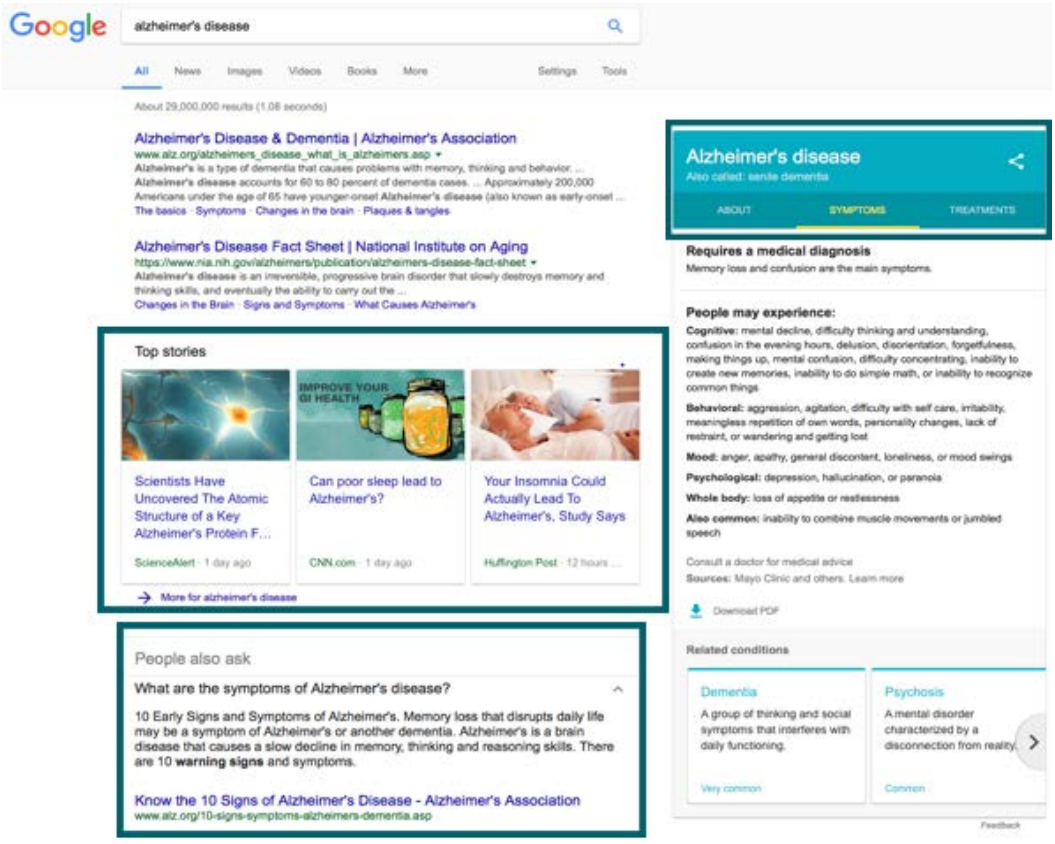
ABSTRACT

Industry and regulatory agencies continue to struggle implementing CDISC for both the study workflow and in support of the submission review process. Several factors contribute to these ongoing challenges including limitations within the CDISC standards themselves and the inability to represent complex relationships across clinical information in limited tools such as Excel. We collaborated with a customer to build a proof of concept based on innovative design concepts, modern technology, and semantic modeling techniques used within other industries which linked metadata across the data flow from Collection --> SDTM --> ADaM --> Analyses and TFLs. This enabled users to see true traceability understanding relationships across artifacts including the impact of changes in the data flow. We will describe the current problem, the focus on user centric design, the process of refactoring CDISC to enable traceability and how this approach will make the business more efficient and deliver higher quality data.

INTRODUCTION

Industry and regulatory agencies continue to struggle implementing CDISC for both the study workflow and in support of the submission review process. Several factors contribute to these ongoing challenges including limitations within the CDISC standards themselves and the inability to represent complex relationships across clinical information in limited tools such as Excel.

In our personal lives we live in a connected world where all our information in linked together (e.g. Facebook, LinkedIn). We take the availability of information for granted and don't realize what's under the covers to link that information together. If you search for information about a disease a family member has, e.g. Alzheimer's, you receive a LOT of interconnected information which helps you understand more about the disease and make better decisions about your family member.



Information within our clinical trials has the same dynamic relationships but unfortunately, we store our standards and data in 2 dimensions with no robust way of linking that information. Within this paper and presentation, we will describe the current problem in more detail, show how to connect the dots by taking a different approach, and proving out the concept for managing the relationships across the clinical information.

THE PROBLEM

The figure below is what we familiar at looking at in our daily work. Separate data sets with variables and values. What is the real problem with trying to pull this information together in a meaningful and clinically relevant way for a clinician who is trying to reach a conclusion. The reality is that because of, these connected and interrelated data points become disconnected and unrelated when we put them in this representation.

pr.xpt

Row	STUDYID	DOMAIN	USUBJID	SPDEVID	PRSEQ	PRLNKID	PRTRT	PRLOC	PRFAST	PRSTDTC
1	ABC123	PR	AD01-101	22	1	02	PET/CT	HEAD	Y	2012-05-22T09:30:00
2	ABC123	PR	AD01-102	22	1	04	PET/CT	HEAD	Y	2012-05-22T08:00:00
3	ABC123	PR	AD01-103	44	1	05	PET	HEAD	Y	2012-05-22T09:00:00

ag.xpt

Row	STUDYID	DOMAIN	USUBJID	AGSEQ	AGLNKID	AGTRT	AGCAT	AGDOSE	AGDOSEU	AGSTDTC
1	ABC123	AG	AD01-101	1	02	18F-Florbetapir	AMYLOID TRACER	370	MBq	2012-05-22T08:40:00
2	ABC123	AG	AD01-102	1	04	11C-PiB	AMYLOID TRACER	370	MBq	2012-05-22T07:20:00
3	ABC123	AG	AD01-103	1	05	FDG	GLUCOSE TRACER	400	MBq	2012-05-22T08:30:00

mo.xpt

Row	STUDYID	DOMAIN	USUBJID	SPDEVID	MOSEQ	MOLNKID	MOREFID	MOTESTCD	MOTEST	MOORRES	MOORRESU
1	ABC123	MO	AD01-101	16	1	02	1234	VOLUME	Volume	1.8	mL
2	ABC123	MO	AD01-101	16	2	02	1234	VOLUME	Volume	1.9	mL
3	ABC123	MO	AD01-101	16	3	02	1234	VOLUME	Volume	3.7	mL
4	ABC123	MO	AD01-101	16	4	02	1234	THICK	Thickness	3	mm
5	ABC123	MO	AD01-101	16	5	02	1235	VOLUME	Volume	864	mL

Whereas technologies such as Google and Facebook integrate these relationships inside their data, in our world the institutional knowledge in our heads is what connects the data points. There are no electronic links between the data and nothing that really provides traceability likes everyone claims.

In the famous story by Hans Christian Andersen “The Emperor’s New Clothes” two weavers convince the Emperor that he is wearing an invisible set of clothes that only the most privileged can see. However, in reality, he is naked and embarrassed in front of his whole kingdom.

This is the same as our industry providing ‘specifications’ or ‘metadata’ that supposedly describes what our data will look like. We have one piece of ‘clothing’ that looks like this

Variable Name	Variable Label	Type	Controlled Terms or Format	Origin	Role	Core	Source Domain	Source Variable	Derivation
STUDYID	Study Identifier	Char		Protocol	Identifier	Req	AE	STUDYID	
DOMAIN	Domain Abbreviation	Char	AE	Assigned	Identifier	Req			DOMAIN = 'AE'
USUBJID	Unique Subject Identifier	Char		Derived	Identifier	Req	AE	STUDYID SUBJID	USUBJID=strip(STUDYID)||'-'||strip(SUBJID);
SUBJID	Subject Identifier for the Study	Char		CRF Page 1	Topic	Req	DM	SUBJID	
SITEID	Study Site Identifier	Char		CRF Page 1	Record Qualifier	Req	DM	SITENO	
AESEQ	Sequence Number	Num			Identifier	Req			Sort records by STUDYID, USUBJID, AEDECOD, AESTDTC and assign AESEQ to 1 for the first record, increased by 1 for each record.
AEREFID	Reference ID	Char		CRF Page 112	Identifier	Perm	AE	AESEQ	
AETERM	Reported Term for the Adverse Event	Char		CRF Page 112	Topic	Req	AE	AETERM	Keep only records with AENONE = ' '.
AEDECOD	Dictionary-Derived Term	Char	AEDICT_F	Assigned	Synonym Qualifier	Req	AE	AEDECOD	Keep in mixed case.

And another one that looks like this

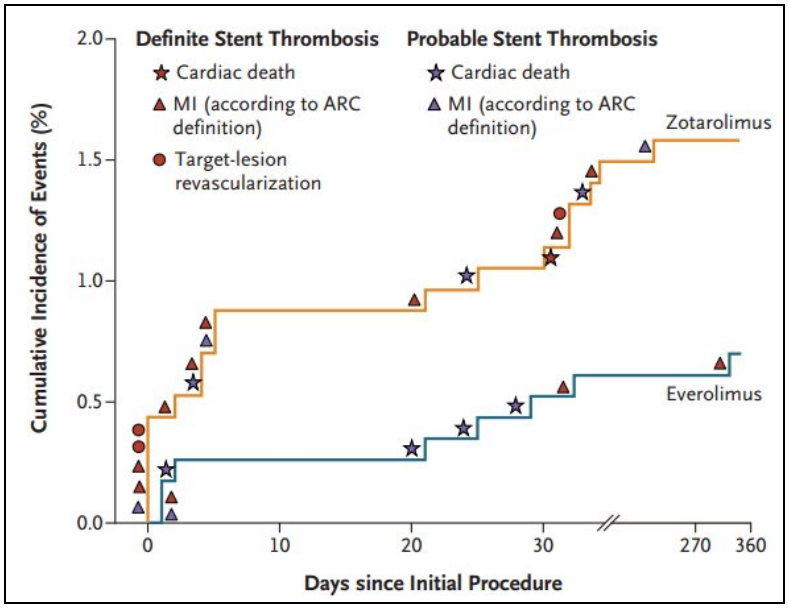
STUDYID	USUBJID	AETERM	AESTDTC	AESEQ	DOMAIN	AESPID	AEDECOD	AEBODSYS
CDISCPILOT01	01-701-1015	VERBATIM_0995	2014-01-03	1	AE	E07	APPLICATION SITE ERYTHEMA	GENERAL DISORDERS AND ADMINISTRATION SITE CONDITIONS
CDISCPILOT01	01-701-1015	VERBATIM_1126	2014-01-09	3	AE	E06	DIARRHOEA	GASTROINTESTINAL DISORDERS
CDISCPILOT01	01-701-1015	VERBATIM_1219	2014-01-03	2	AE	E08	APPLICATION SITE PRURITUS	GENERAL DISORDERS AND ADMINISTRATION SITE CONDITIONS
CDISCPILOT01	01-701-1023	VERBATIM_0300	2012-08-07	1	AE	E08	ERYTHEMA	SKIN AND SUBCUTANEOUS TISSUE DISORDERS
CDISCPILOT01	01-701-1023	VERBATIM_0300	2012-08-07	4	AE	E08	ERYTHEMA	SKIN AND SUBCUTANEOUS TISSUE DISORDERS
CDISCPILOT01	01-701-1023	VERBATIM_1549	2012-08-07	2	AE	E09	ERYTHEMA	SKIN AND SUBCUTANEOUS TISSUE DISORDERS
CDISCPILOT01	01-701-1023	VERBATIM_1650	2012-08-26	3	AE	E10	ATRIOVENTRICULAR BLOCK SECOND DEGREE	CARDIAC DISORDERS

Neither of which ever actually interacts with each other! This gives us this false sense of traceability or compliance that because we are checking the box that we have specifications we have better quality data. In reality, we are the ones wearing the Emperor’s clothing believe that our ‘specs’ give us quality data.

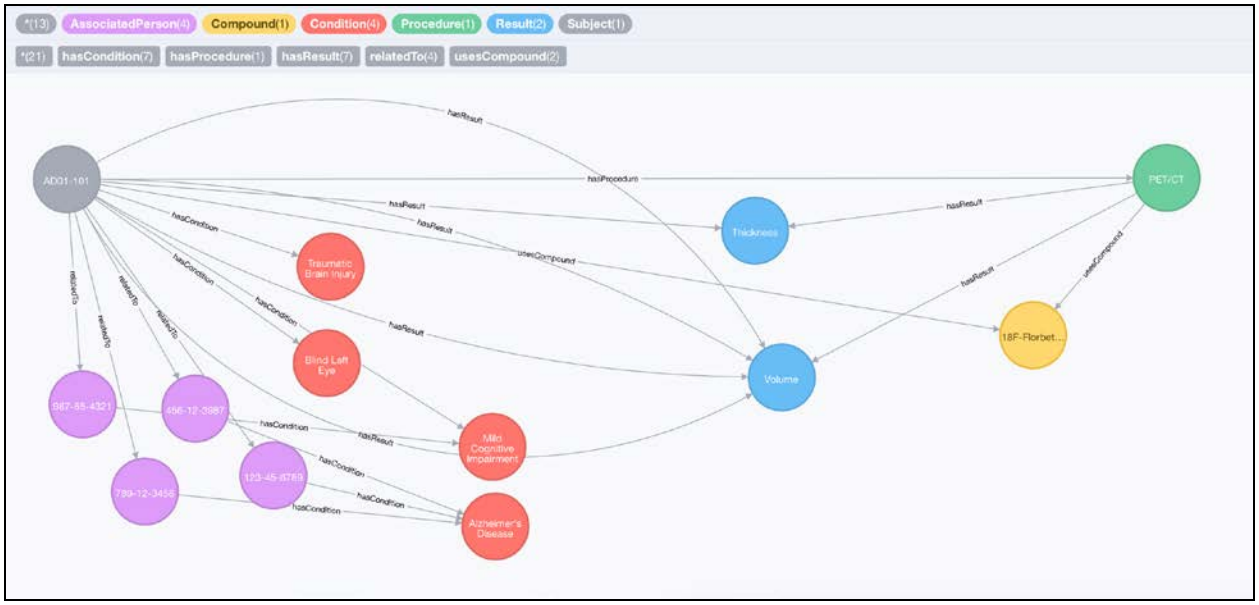
CONNECTING THE DOTS

The first step in connecting the dots is for the industry to stop using the word 'metadata'. I know the first response to this comment is "What?!?" Most industries out there are not even sure what the word metadata means or use it in a very different context. The reality is that that all the information we collect whether it is the value of a blood pressure, the name of a variable, or the length of numeric value is all data. Data that must be linked together in intelligent ways to really allow to use our data.

We can connect this information in the form of a graph. You probably hear the word graph and think of a fancy picture you produce to show lines, bars, or time to analysis like the figure below.



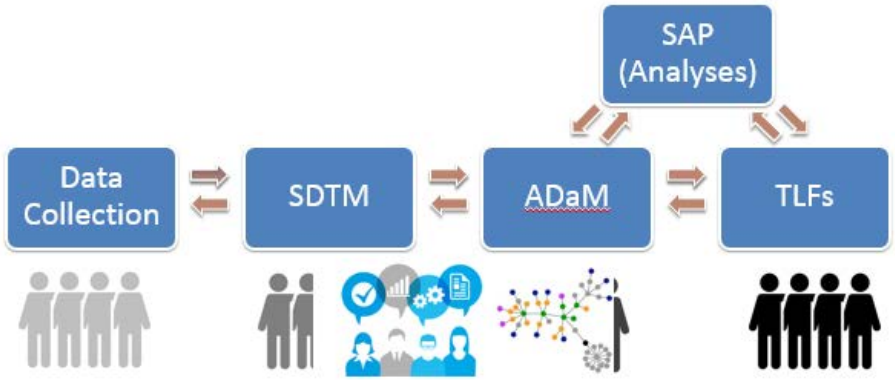
These are not the graphs we are talking about. The graphs we describe are databases that use graph structures for semantic queries with nodes, edges and properties to represent and store data. A key concept of the system is the graph (or edge or relationship) directly relates data items in the store and the relationships allow data in the store to be linked together directly, and in many cases retrieved with one operation. This contracts a relational database which forces a structure on that is hard to update if relationships change.



The figure above shows the information about the patient which includes values (nodes) and relationships (connections) and links all the patient information together. By baking the relationships into the database, institutional knowledge is not needed to connect the dots. In today's world information is duplicated across multiple datasets, it's takes brain power to connect the information every time, and no relationships really exist – Traceability is a farce and you are wearing the Emperor's clothing.

CASE STUDY: PROVING THE METADATA CONCEPT

At this point, you are probably saying "Another academic exercise". Over the last few years people have been presenting about semantics, linked data, and how it will change the world, but haven't really shown the value. What we will now show you is how we used these concepts to build a proof of concept for study specifications that links information across the CDISC metadata flow as well as reporting of the results.



The diagram shows a simplistic flow of our metadata and data through the lifecycle. How many of us have developed specifications for one the components in the diagram and can honestly say it was a "pleasurable" experience? This flow while seemingly simple, is fraught with problems. Specifications are created in silos by people who throw it over the wall to the next person in line. We use rudimentary tools like Excel which are not scalable, cannot provide change control, and cause people to duplicate information from silo to silo as well as within a silo. This siloed spreadsheet world has very little in the way of real connected relationships and required significant CDISC knowledge and years of experience to actual produce the artifacts.

What if we could…

Reimagine the way specifications are developed making it something people enjoy doing

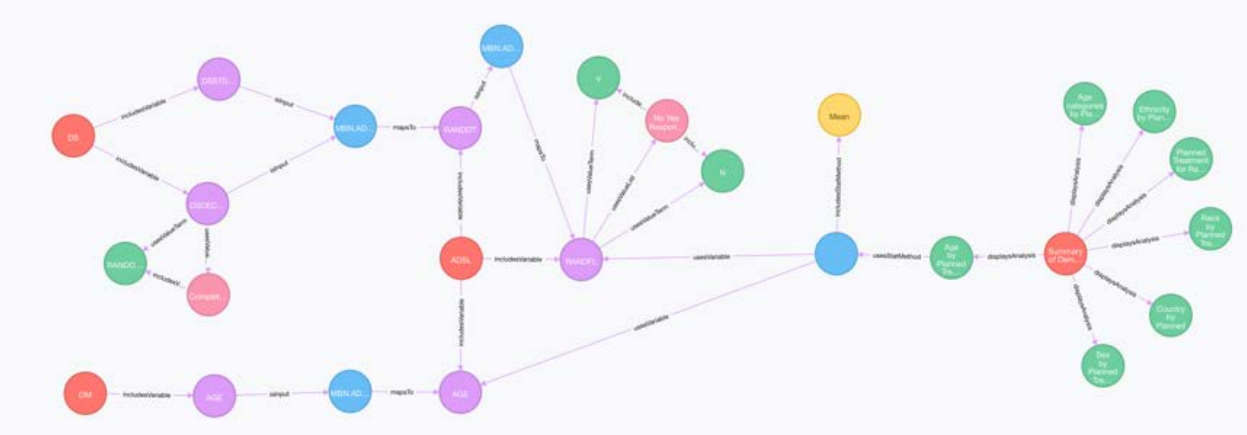
Intelligently link the specifications across the data flow in BOTH directions providing an end user the impact of making a change anywhere along the process

Have an underlying database that had the relationships as part of the data

Replace the siloed teams with collaborative cross functional teams that understand how their specific decisions impacted other along this data flow

Our proof of concept with the customer focused on tackling these "What ifs". We started with implementing a new methodology that put the user and the business need at the heart of the design. We focused on the user experience, iteratively went through the design and development, and used a newer technology stack not usually implemented in our industry. With regards to this presentation, we will focus on the use of a newer underlying database technology, graph databases, and how we leveraged to meet the business need.

We started with a set of standard CRFs, SDTM/ADaM/TFL specifications, an SAP, and a LOT of metadata buried in Excel with minimal connectivity. Our first step was to pull out all the valuable information in their Excel files contained within black boxes and refactored into an integrated graph database. This exercise was probably the bulk of the work as we needed to develop a new model (nodes and relationships that linked together the previous paragraphs of information. We iterated over the model throughout the POC adding pieces and parts as necessary to build that true integration. In some cases (analyses and displays), we had to build these relationships with no previous knowledge as CDISC does not address this type of metadata. After building the model we loaded the customer's metadata into the graph. The figure below is an example of that metadata.



After developing the metadata, using user centric design, we built out an interface and working prototype that could leverage this metadata focusing on:

- A user experience that people actually enjoy
- Real Traceability showing the flow of information from collection to analysis
- Impact analysis allowing a user to see what impact a change in their study will have on upstream or downstream artifacts
- Imbedding the institutional knowledge in our heads within the model

Since we were able to build the institutional knowledge into the graph, I don't need an army of specialists to create specifications nor do I need a plethora of people in my standards organization.

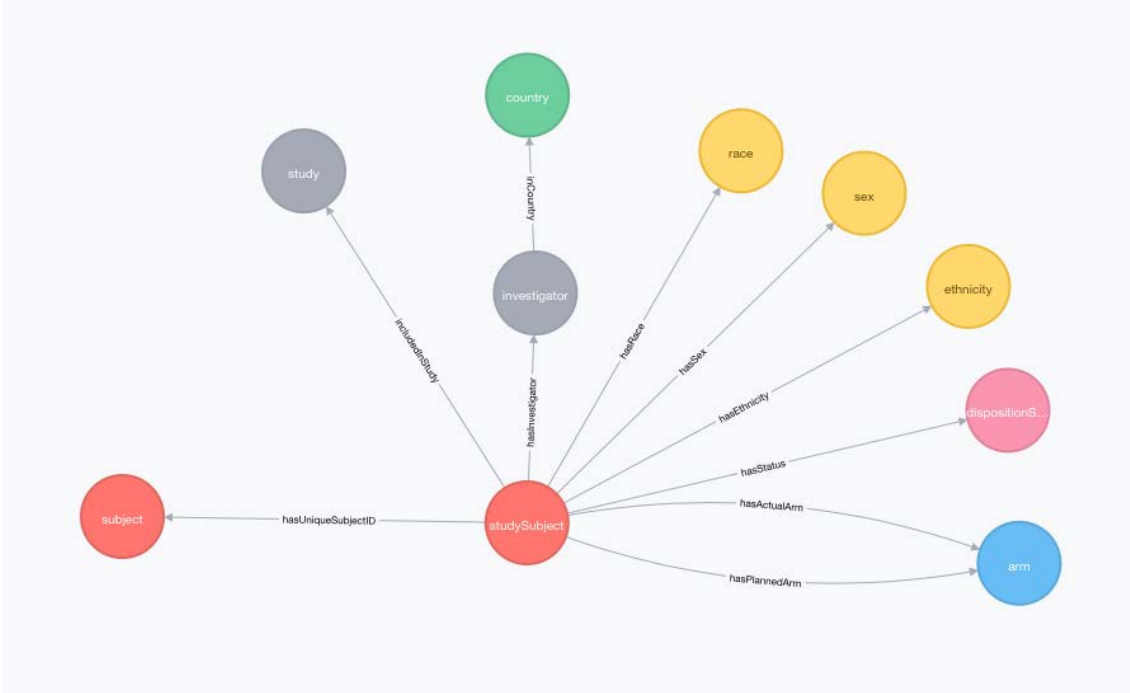
Other industries have been using this methodology and technology for years (e.g. Google, Facebook) and the industry is starting to explore the capabilities. It's time for us to move into 21st century and exponentially leap forward in our ability to capture and use our data.

WHY NOT THE DATA?

Above, we describe the ability to refactor the CDISC standards and build metadata that can provide true linking of information and traceability in your metadata. What if take even a bigger leap and put the actual data into this model throwing out the way we currently 'do' metadata.

At the 2017 CSS, one of the sessions focused on graph database technologies and challenged participants to think about clinical data differently. As part these sessions, participants were separated into groups, were provided an example of SDTM data from the Demographics (DM), Vital Signs (VS), and AE (Adverse Events) domain, and were asked to refactor the SDTM domains as a graph. At the end of this session, a representative from each team was asked to explain their modeling decisions and were asked about the value of representing the data using a graph.

While providing different versions of the graph, ALL participants were able to see the value of representing the data in this type of model. This was very encouraging given (a) approximately 50% of the 25 attendees were non-technical folks and (b) 75% had minimal exposure to graph technologies prior to participating in the session. The figure below is an example of a model for DM developed by one of the groups.



The session leaders created DM data for three studies, populated the model, and shared a number of use case. Below is one use case.

FDA has slightly different requirements to represent Screening Failures in the DM domain than the SDTM IG. PMDA follows the SDTM IG explicitly. The team was able to extract SDTM data the meets the CDISC/PMDA rules and FDA rules from the **SAME** graph database. The figure below shows the two data sets created from the same database.

SDTM DM domain conformant to CDISC/PMDA rules

STUDYID	USUBJID	SUBJID	RFSTDTC	RFENDTC	RFXSTDTC	RFXENDTC	RFXPDTC	SITEID	AGE	AGEU	SEX	RACE	ETHNICITY	ARMCD	ARM	ACTARMCD	ACTARM	COUNTRY	DMDTC	DMDY
CDISCPILOT01	01-701-1015	1015	1/2/14	7/2/14	1/2/14	7/2/14	2014-07-02T11:45	701	63	YEARS	F	WHITE	HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	12/26/13	-7
CDISCPILOT01	01-701-1028	1028	7/19/13	1/14/14	7/19/13	1/14/14	2014-01-14T11:10	701	71	YEARS	M	WHITE	NOT HISPANIC OR LATINO	Xan_Hi	Xanomeline High Dose	Xan_Hi	Xanomeline High Dose	USA	7/11/13	-8
CDISCPILOT01	01-701-1047	1047	2/12/13	3/29/13	2/12/13	3/9/13	7/28/13	701	85	YEARS	F	WHITE	NOT HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	1/22/13	-21
CDISCPILOT01	01-701-1118	1118	3/12/14	9/9/14	3/12/14	9/9/14	2014-09-09T13:28	701	52	YEARS	M	WHITE	NOT HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	2/27/14	-13
CDISCPILOT01	01-701-1130	1130	2/15/14	8/16/14	2/15/14	8/16/14	2014-08-16T13:10	701	84	YEARS	M	WHITE	NOT HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	2/9/14	-6
CDISCPILOT01	01-701-1133	1133	10/28/12	4/28/13	10/28/12	4/28/13	2013-04-29T10:13	701	81	YEARS	F	WHITE	NOT HISPANIC OR LATINO	Xan_Hi	Xanomeline High Dose	Xan_Hi	Xanomeline High Dose	USA	10/23/12	-5
CDISCPILOT01	01-701-1153	1153	9/23/13	4/1/14	9/23/13	3/16/14	2014-04-01T14:25	701	79	YEARS	F	WHITE	NOT HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	9/6/13	-17
CDISCPILOT01	01-701-1180	1180	2/12/13	3/23/13	2/12/13	3/18/13	4/7/13	701	56	YEARS	M	WHITE	NOT HISPANIC OR LATINO	Xan_Hi	Xanomeline High Dose	Xan_Hi	Xanomeline High Dose	USA	1/28/13	-15
CDISCPILOT01	01-701-1181	1181	12/5/13	12/12/13	12/5/13	12/9/13	5/23/14	701	79	YEARS	F	WHITE	NOT HISPANIC OR LATINO	Xan_Hi	Xanomeline High Dose	Xan_Lo	Xanomeline Low Dose	USA	11/26/13	-9
CDISCPILOT01	01-701-1188	1188	2/15/13	3/25/13	2/15/13	3/24/13	8/4/13	701	71	YEARS	M	WHITE	NOT HISPANIC OR LATINO	Xan_Lo	Xanomeline Low Dose	Xan_Lo	Xanomeline Low Dose	USA	2/3/13	-12
CDISCPILOT01	01-701-1234	1234	3/30/13	9/22/13	3/30/13	9/22/13	2013-09-22T09:25	701	69	YEARS	M	WHITE	NOT HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	3/20/13	-10
CDISCPILOT01	01-701-1239	1239	1/11/14	7/11/14	1/11/14	7/10/14	2014-07-11T15:40	701	56	YEARS	M	WHITE	HISPANIC OR LATINO	Xan_Hi	Xanomeline High Dose	Xan_Hi	Xanomeline High Dose	USA	12/28/13	-14
CDISCPILOT01	01-701-1275	1275	2/7/14	6/14/14	2/7/14	5/31/14	2014-06-14T12:35	701	61	YEARS	M	AMERICAN INDIAN OR ALASKA NATIVE	NOT HISPANIC OR LATINO	Xan_Hi	Xanomeline High Dose	Xan_Hi	Xanomeline High Dose	USA	1/25/14	-13
CDISCPILOT01	01-701-1345	1345	10/8/13	3/18/14	10/8/13	3/18/14	2014-03-18T14:12	701	63	YEARS	F	WHITE	NOT HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	10/1/13	-7
CDISCPILOT01	01-701-1386	1386					1/7/14	701	71	YEARS	M	WHITE	NOT HISPANIC OR LATINO	SCRNFAIL	Screen Failure			USA	12/28/13	

SDTM DM domain conformant to FDA rules

STUDYID	USUBJID	SUBJID	RFSTDTC	RFENDTC	RFXSTDTC	RFXENDTC	RFXPDTC	SITEID	AGE	AGEU	SEX	RACE	ETHNICITY	ARMCD	ARM	ACTARMCD	ACTARM	COUNTRY	DMDTC	DMDY
CDISCPILOT01	01-701-1015	1015	1/2/14	7/2/14	1/2/14	7/2/14	2014-07-02T11:45	701	63	YEARS	F	WHITE	HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	12/26/13	-7
CDISCPILOT01	01-701-1028	1028	7/19/13	1/14/14	7/19/13	1/14/14	2014-01-14T11:10	701	71	YEARS	M	WHITE	NOT HISPANIC OR LATINO	Xan_Hi	Xanomeline High Dose	Xan_Hi	Xanomeline High Dose	USA	7/11/13	-8
CDISCPILOT01	01-701-1047	1047	2/12/13	3/29/13	2/12/13	3/9/13	7/28/13	701	85	YEARS	F	WHITE	NOT HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	1/22/13	-21
CDISCPILOT01	01-701-1118	1118	3/12/14	9/9/14	3/12/14	9/9/14	2014-09-09T13:28	701	52	YEARS	M	WHITE	NOT HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	2/27/14	-13
CDISCPILOT01	01-701-1130	1130	2/15/14	8/16/14	2/15/14	8/16/14	2014-08-16T13:10	701	84	YEARS	M	WHITE	NOT HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	2/9/14	-6
CDISCPILOT01	01-701-1133	1133	10/28/12	4/28/13	10/28/12	4/28/13	2013-04-29T10:13	701	81	YEARS	F	WHITE	NOT HISPANIC OR LATINO	Xan_Hi	Xanomeline High Dose	Xan_Hi	Xanomeline High Dose	USA	10/23/12	-5
CDISCPILOT01	01-701-1153	1153	9/23/13	4/1/14	9/23/13	3/16/14	2014-04-01T14:25	701	79	YEARS	F	WHITE	NOT HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	9/6/13	-17
CDISCPILOT01	01-701-1180	1180	2/12/13	3/23/13	2/12/13	3/18/13	4/7/13	701	56	YEARS	M	WHITE	NOT HISPANIC OR LATINO	Xan_Hi	Xanomeline High Dose	Xan_Hi	Xanomeline High Dose	USA	1/28/13	-15
CDISCPILOT01	01-701-1181	1181	12/5/13	12/12/13	12/5/13	12/9/13	5/23/14	701	79	YEARS	F	WHITE	NOT HISPANIC OR LATINO	Xan_Hi	Xanomeline High Dose	Xan_Lo	Xanomeline Low Dose	USA	11/26/13	-9
CDISCPILOT01	01-701-1188	1188	2/15/13	3/25/13	2/15/13	3/24/13	8/4/13	701	71	YEARS	M	WHITE	NOT HISPANIC OR LATINO	Xan_Lo	Xanomeline Low Dose	Xan_Lo	Xanomeline Low Dose	USA	2/3/13	-12
CDISCPILOT01	01-701-1234	1234	3/30/13	9/22/13	3/30/13	9/22/13	2013-09-22T09:25	701	69	YEARS	M	WHITE	NOT HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	3/20/13	-10
CDISCPILOT01	01-701-1239	1239	1/11/14	7/11/14	1/11/14	7/10/14	2014-07-11T15:40	701	56	YEARS	M	WHITE	HISPANIC OR LATINO	Xan_Hi	Xanomeline High Dose	Xan_Hi	Xanomeline High Dose	USA	12/28/13	-14
CDISCPILOT01	01-701-1275	1275	2/7/14	6/14/14	2/7/14	5/31/14	2014-06-14T12:35	701	61	YEARS	M	AMERICAN INDIAN OR ALASKA NATIVE	NOT HISPANIC OR LATINO	Xan_Hi	Xanomeline High Dose	Xan_Hi	Xanomeline High Dose	USA	1/25/14	-13
CDISCPILOT01	01-701-1345	1345	10/8/13	3/18/14	10/8/13	3/18/14	2014-03-18T14:12	701	63	YEARS	F	WHITE	NOT HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA	10/1/13	-7
CDISCPILOT01	01-701-1386	1386					1/7/14	701	71	YEARS	M	WHITE	NOT HISPANIC OR LATINO	SCRNFAIL				USA	12/28/13	

Why is this important? It shows that you can create SDTM data from data stored in a graph. Second, in the current environment a user would need to create separate analysis datasets for the FDA and PMDA if they wanted some level of traceability. If the analysis created based on the data in the graph, there is 100% traceability between my **single** data source (i.e. the graph) and my analyses. Less work, higher quality, data integrity.

CONCLUSION

At the beginning of this paper, we described how Industry and regulatory agencies continue to struggle implementing CDISC for both the study workflow and support of the submission review process. As we described in the paper, one of the primary reasons we have struggled is the limitations of the underlying models and the current technology used

by industry to describe the multidimensional nature of clinical information. We will continue to struggle if we don't look to embrace new ways of modeling our clinical information and really answering the questions we have in the clinical development process.

In our personal lives, we live in a connected world where all our information in linked together (e.g. Facebook, LinkedIn), yet we don't take that simple step of realizing how we could represent the information in our clinical research world in the same way.

In conclusion, we should stop trying to build 'traceability', 'governance', or 'linkages' in a world where the underlying models and existing technology can't support it. Instead we do the best we can now to get the deliverables out the door and focus our energies on the next generation.

CONTACT INFORMATION

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

Author Name: Chris Decker
Company: d-Wise
Email: Chris.Decker@d-wise.com

Author Name: Scott Bahlavooni
Company: d-Wise
Email: Scott.Bahlavooni@d-wise.com