Peter L. Elkin, MD, MACP, FACMI, FNYAM
Professor and Chair
Department of Biomedical Informatics &
Professor Internal Medicine
Department of Biomedical Informatics
State University of New York at Buffalo

Session 7: Data Standardization in Data Warehousing

Hello everybody. My name is Dr. Peter Elkin. I'm from the University of Buffalo, and I'm here to talk to you about the Center for Leading Innovation and Collaboration's 2021 Insights to Inspire Informatics, the Journey to interoperability. I'm a professor of Biomedical Informatics and Internal Medicine at the University of Buffalo. I'm excited to be able to talk to you about our journey to interoperability. I'll be sharing some best practice advice on data warehousing and standardization.

Objective - So the idea is to talk to you about the importance to the consortium for data standardization in research informatics and in data warehousing. A topic near and dear to many of our hearts and informatics.

During this session you will learn about – Some of the best practices in clinical research data warehousing; What changes we have made to improve data standardization; What are the most significant challenges in implementing standards; Recommendations for addressing standardization challenges; Examples of how we have been able to address data standardization at our hub, who we involved and what resources were required in order to do this properly.

Some of the best practices are in clinical research data warehousing - So what are some of the best practices in clinical data warehousing research? Well, first you need to develop and implement a *strong data governance strategy*. This has to do with *understanding what each of your data elements* means and how you're coding that meaning, in a standardized way, but is consistent across your academic enterprise. This *requires collaboration of many people*, including clinical leaders, informaticians, and also the individual practitioners that are on the front lines who are interested in using the data that they create in their clinical practices for research. You must employ an *implementation strategy* that *preserves data provenance*. Data provenance is how you know where the data's been and what transforms have been done to the data as it's made its way from its initial discovery and documentation into your research data warehouse. You have to *develop mission and vision statements*, your data warehouse, and often with *use cases* to support the principal mission. You need to *employ fair data principles*. If you have multiple data sources without a common identifier, you will need a *strong master patient index program* to ensure that the patient's data are assigned to the right patient. You must *build into your data warehouse an evaluation program* and *plan for a continuous quality*

The University of Rochester Center for Leading Innovation and Collaboration (CLIC) is the coordinating center for the Clinical and Translational Science Awards (CTSA) Program, funded by the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH), Grant U24TR002260.

improvement program as well. You should think of how to teach and support your users. So this includes data science education, giving your research community a *sense of the limitations of your data warehouse* and trying to bring them into the fold. So as you develop it, it'll be something that they're comfortable and familiar with using.

What changes we made to improve data standardization - So what changes have, have we made to improve data standardization? And what advice do we have? Well, first there's a bottom up approach and just data standardization. So all the data elements per the data governance process needs to be both formally and systematically defined. So formal definitions are description based logic definitions using standardized ontology and systematic definitions are English like definitions that build off one another, like you might see in a dictionary. You should use a strongly typed system where data types are explicit so that you don't have trouble with cross data types in different parts of your data warehouse incoming feeds. You should use the correct terminology for the right task. So if you're using SNOMED CT, for things like diagnosis and findings. If you used RX NORM for clinical drug names, LOINC for sections of the clinical record and laboratory result test names, and the gene ontology for genetic content and pathway names and the human phenotype ontology for phenotyping constructs from this. There are some new things that are coming down the pike. We'll talk about a little bit later.

These should be stored along with the structured data in a standardized observational data model, the commonly used ones today are OMOP, i2b2, and PCORnet. If you're not familiar with any of these, you need to talk to your Informatics Lead, and they should bring you up to speed. You need to differentiate in your data warehouse 1) things that are codified that are true assertions, 2) things that the clinician said are true, 3) from things that they've said are false, 4) from things that are being worked up, that you don't know yet are true or false. You should use a system capable of generating post coordinated expressions and therefore semantic triples or compositional expressions. And the reason for this is that the level of granularity of the standardized coding systems and the clinical nomenclature that you use in your notes and reports, and even in your problem lists vary. And in order to reconcile this, sometimes it takes multiple codes to code up a single utterance. You must account for longitudinality in your models. If you look at the data that we have, there's a lot of missing data, and there's a lot of data that's quite sparse, and we're trying to compare data for people over periods of time that may not be quite equivalent. So we need models for how to create longitudinality in ways that actually represent the data well. Mistakes in this area are common reasons for studies to produce invalid results. Standardized access models, help to standardize the researcher's view of the data. So it's not only how you define it, but how you provide access to it that help your researchers understand how best to use your data repository. One standard that is used quite a lot is the FHIR standard for HL7 and SMART on FHIR for keeping session information.

What significant challenges that remain in implementing standards. -What are the most significant challenges in implementing data standardization? Well, data science is hard. It's a lot of work. Standards are complex in their nature. So it takes a while to learn them. So you need people on your team who understand that. There are many standards to choose from, of course. And so you need a guide to help you through what choices there are. They have to be implemented consistently. So it's not just that you use the same standards, but you implement them in the same way to create the standardization that you want across both your enterprise and then for the consortium across all of the hubs. Standards have an update cycle. And so just when you think that it's all done, when you've implemented your standards, they change on at least an every six month basis. And you have to have a model for how you're going to take data that's already in patient records and point it to codes that have now been marked obsolete. And then those codes are now pointed to the new code that should be used in its place. And trailing through that historically will empower your enterprise to be able to find all the data that's relevant for a particular data point. Standards need a change management program. This is very important because standards do change over time and you need to filter through the rhetoric to get the best practices. So there's a lot of standards development organizations that are pushing their

The University of Rochester Center for Leading Innovation and Collaboration (CLIC) is the coordinating center for the Clinical and Translational Science Awards (CTSA) Program, funded by the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH), Grant U24TR002260.

standards. And you really want someone on your team that understands standardization so that they can help advise what is most commonly used and why. Be aware of mapping between terminologies. You may end up with apples and oranges because maps often don't take into account all of the nuances that were intended by the original standards developer and they rarely have a formal definitions associated with the mapping. Consider shared cohort definition. So as you start to build cohorts for types of patient populations that you wish to study, sharing those in a central place within your organization. So that the next researcher who wants to do research on a similar population can start from that cohort definition and try to build consistency. The whole idea behind standardization is to get consistency across all of your enterprise and makes for a much more accurate and reproducible data repository and data environment. There's no one solution to rule them all right now. You have to pick and choose lots of standards. Some are working on it. The Department of Veterans Affairs has created a terminology called Solar, which is integrated into its logic SNOMED CT, LOINC and RxNorm, which can be a step in the right direction. And they've created a metadata standard, which has been passed in HL7 called the Associated Normal Former (ANF). And so they're at the present time trying that out to see if that could help people to use one entity and be able to be successful.

Recommendations for addressing standardization challenges – I've broken it up into two categories. The first is organizational. First convince your leadership that you need a strong data governance program, then implement data provenance, which is again, I'll remind you a way to track everything that's done with the data, so that you can get reproducibility. And, you do this behind the scenes when they're not looking, even if they don't ask for it, because at the end of the day, when you ask yourself why did the data show that you can trace it right back to the source. Invite cinical leadership to take part in the process, especially for their specialty. There's no question that better definitions come from the specialty themselves. You should hire an ontologist and have one on board. They will get you out of thorny problems. When the current standards don't support what you currently need. From a technical perspective, start by using what's already been well formed solutions by others, try them out, see what you like, and don't like about them and that can give you an idea of what you have to build yourself. If you want to build your own system, compare your work with others on a similar problem so that you can get a sense of how well you're doing. Pick an observational database model and stick with it, jumping around between them cost you money and diverts your attention away from the real process thorny problems that need to be solved. Make sure that you can codify your data and at a minimum with SNOMED and RxNorm. How you access the data can be just as important as how you build the clinical data warehouse. So give very clear instructions with very clear usability tested interfaces to your clinical research population. It will empower your community. Get your EHR vendor to turn on their FHIR appliance. This will allow you to both access the data in your clinical records, and also to write back to it when you want to use the data to translate that research into a better clinical practice. The FHIR appliances can write to the banner bars and give alerts to your physicians in ways that can make their practice better and create guard rails around clinical practice in a way that creates safer and more effective care.

Examples of how we have been able to address data standardization at our hub and who we involved - We started data governance and involved our health system and they embraced it. Matter of fact, they seem to love it. This gives them control over what goes into the data warehouse in a way that they really haven't had before. We show value-add for the health system. They need the codes back for their own projects, even managing the hospital. So providing SNOMED, LOINC, and RxNorm codes can help the hospital to actually manage their own business in a more accurate way and can convince their physicians that the level of quality of the research that management's doing is consistent with best practice. We employed natural language processing to codify our textual data into problem lists, med list, lab tests names, and at the present doing clinical notes and reports. We created a viewer to facilitate data-driven observational research and data-driven recruitment to clinical trials. And this allows people who don't understand data science to do

data science work on the database without having to become informaticians. We call it the BMI, biomedical informatics investigator. We also built a cell phone application for recruitment to clinical trials that are being registered within our clinical trial management system.

So who do we involve? Well, we need to involve everybody from every service line, at least somebody from every service line. We need to involve administration, and we have to get your CTSA PI on board and interested in this. And I know there'll be interested if it helps to empower research at your hub, apply for more grants and to get preliminary data for interesting prospective trials.

The resources were required in order to do this properly? So what resources were required to increase data standardization? Well, excellence isn't easy, nor is it cheap. We budgeted for what we think that a perfect environment would be, should be about $5 million a year, not counting the hardware from our chief information officer at the university who by the way, is all going on this with us and helped create the plan together with myself, our CTSA PI, Tim Murphy, and our Dean, Michael Kane. This includes a director of the program, three programming resources, data governance folks to work across the university in each clinical entity, ontology support, natural language processing support. And we're lucky we have a clinical informatics fellowship, which is a somewhat cost sharing, the clinical informatics fellows can get quite a lot of work done in terms of data governance and data provenance that can help us to be successful. Also on the training side, they can be quite helpful with our faculty and, and even our trainees. So we saved money also by having students help with the work, student projects can be on this kind of an infrastructure. And this helps both get work done for the university and also helps us to give very interesting and meaningful and publishable projects to our students. And then we try to provide a return on investment for the work. You know, there are a lot of ways that quality assurance, utilization review projects and quality improvement projects can be done on the same data that we use for research. And this can provide a return on investment for everybody.

Takeaway Points – 1) Standardization and interoperability is definitely worth the effort. 2) It can empower your organization to be its best. 3) It can and does improve your school and university's rankings. 4) It can attract better faculty and better students. 5) It can position you to leverage your data toward next generation predictive analytics, which can pay you back in terms of FDA-approved utilities and can create an ROI for your standardized clinical data warehouse. 6) Beyond that, it'll make you happy to do the right thing and to create an interoperable clinical data warehouse. Not when you start, of course, you'll have to work through the thorny problems, but certainly when you have completed it, and even when you're on the journey, it will become very rewarding.

We encourage you to view the following webcasts as they provide foundational information for the rest of the series.

Session 1 – Introduction to Insights to Inspire
Session 2 – Language of Informatics
Session 3 – Introduction to Informatics
Session 4 – Introduction to Maturity Models
Session 5 – Importance of Interoperability

After that, please view the remaining webcasts in the order of your choice. The webcasts are available on the CLIC website: https://clic-ctsa.org/ or you can access them by searching CLIC_CTSA on Vimeo or YouTube.

Thank you for viewing Session 7 ***– Data Standardization in Data Warehousing***.
Please join us for the rest of the ***Informatics: The Journey to Interoperability series***.

The University of Rochester Center for Leading Innovation and Collaboration (CLIC) is the coordinating center for the Clinical and Translational Science Awards (CTSA) Program, funded by the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH), Grant U24TR002260.