CURATING MEDICAL KNOWLEDGE TO ENABLE SYSTEMATIC LEARNING OF DIAGNOSTIC EVIDENCE FROM ELECTRONIC HEALTH RECORD SYSTEMS

Professor Brendan Delaney, Imperial College London, UK. Dr Jean-Karl Soler, Mediterranean Institute of Primary Care, Malta. Dr Derek Corrigan, Royal College of Surgeons in Ireland, Ireland. Dr Vasa Curcin, King's College London, UK. Mr Erik Mayer, Imperial College London, UK.

Description

Diagnostic error results in significant harm to patients. Our aim is to create a scalable infrastructure to support learning of diagnostic predictors and their presentation via an EHR-integrated decision support system. The relevant information artefacts are, concepts of disease, symptoms, signs and investigations and the mathematical relationships between them.

In order to achieve a common set of concepts there are problems of classification and of granularity that need to be managed across the healthcare system:

- Terminology is not sufficient to support learning as we require unambiguous and complete concepts in a classification (such as ICD10 or ICPC2) in order to derive stable relationships.
- In addition, the size of the classification and its granularity needs to be defined such that the clinical detail supported and the size of the data available match, or there will be too many categories with little data.
- Unfortunately ICD10 does not define symptoms and signs and SNOMED-CT is not a classification.

Our solution is:

- To define a core ontology of restricted, collaboratively curatable, code-sets in SNOMED-CT that can support learning for diagnosis of chosen subsets of conditions. (4)*
- To make the ontology available as a web-service. (4-5)
- To use the ontology to drive a smart coding interface linked to the EHR. (1)
- To gather Reasons for Encounter and diagnoses organised in Episodes of Care, linked over time and setting. (2)
- To use Machine Learning to derive diagnostic associations linking symptoms, signs, test results and demographics with diagnoses. (3)
- The associations are also stored in the ontology, driving a Decision Support Interface. (5)
- This closes the learning loop, creating a LEARNING HEALTH SYSTEM for diagnosis (Figure 1)

*See figure 1

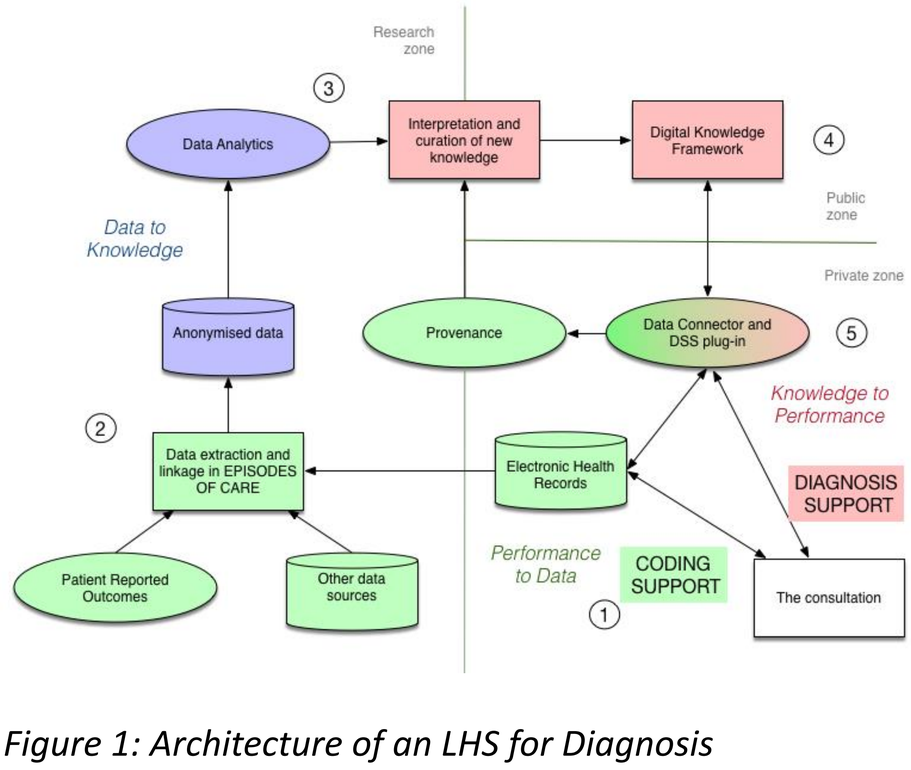
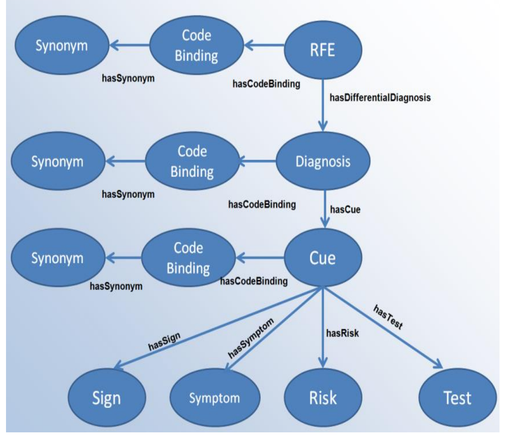


Figure 1: Architecture of an LHS for Diagnosis



SNOMED-CT expressions are flexible

- Concepts
- Terms
- Relationships (polyhierarchical)
- Synonyms

Can be combined by pre-coordination

Can allow negation

243796009|situation with explicit context|:

{408729009|finding context|=410516002|known absent|,

THE ONTOLOGY MANAGES THE CLASSIFICATION OF UNIQUE CONCEPTS – which SNOMED CT cannot.

Figure 2: Ontology for diagnosis based on Weed's Episode of Care Model.

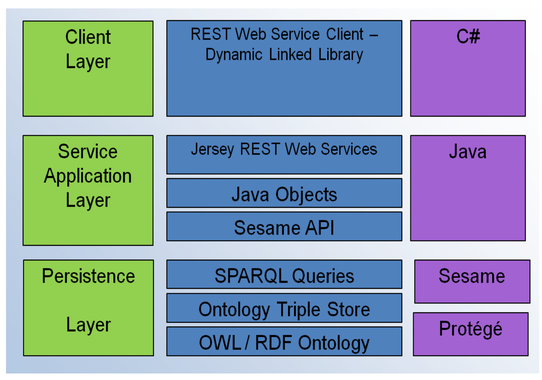


Figure 3: Service stack

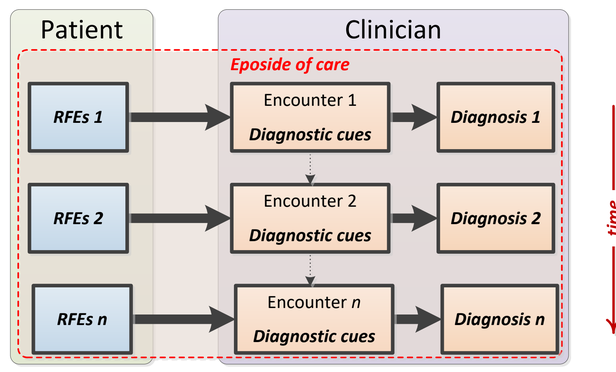


Figure 4: Reasons for Encounter linked to Episodes of care for analytics

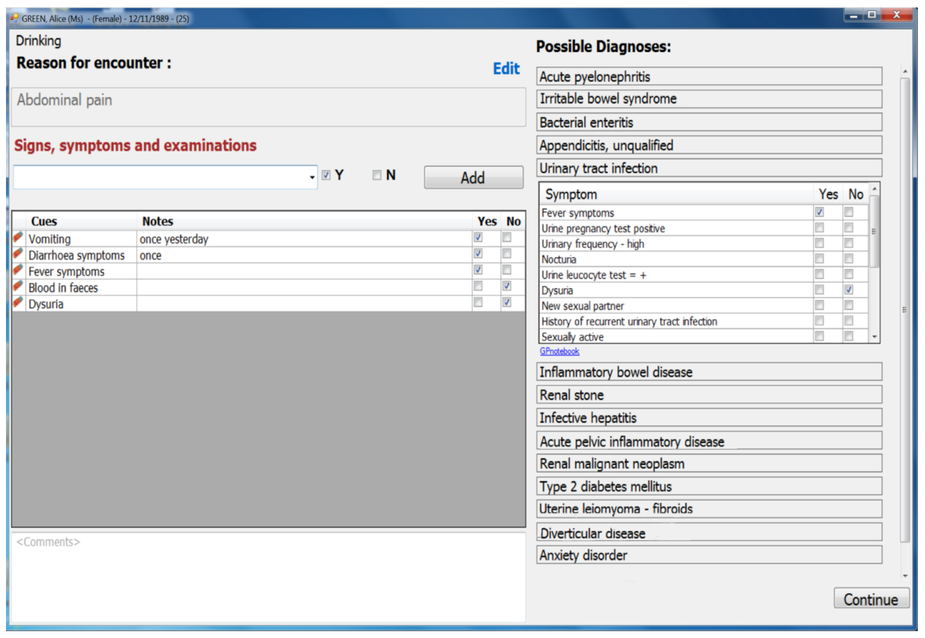


Figure 5: Data capture and coding interface linked to EHR.

1.Olga Kostopoulou, Andrea Rosen, Round Thomas, Ellen Wright, Brendan Delaney, and Abdel Douiri. Early diagnostic suggestions improve accuracy of GPs: an experimental study. BJGP 2014: 65 (630), e49-e54

2.Kostopoulou, Olga; Lionis, Christos; Angelaki, Agapi; Ayis, Salma; Durbaba, Stevo et al. Early diagnostic suggestions improve accuracy of family physicians: a randomized controlled trial in Greece. 2015 ePub ahead of print Fam. Pract. p. cmv012

3.Brendan C. Delaney, Vasa Curcin, Anna Andreasson, et al., "Translational Medicine and Patient Safety in Europe: TRANSFoRm—Architecture for the Learning Health System in Europe," BioMed Research International, vol. 2015, Article ID 961526, 8 pages, 2015. doi:10.1155/2015/961526

4.Porat T, Kostopoulou O, Woolley A, Delaney BC, Eliciting User Decision Requirements for Designing Computerized Diagnostic Support for Family Physicians, Journal of Cognitive Engineering and Decision Making, ISSN: 1555-3434 SAGE Publications; 2016;10(1):57–73

5.Kostopoulou O, Porat T, Corrigan D, Mahmoud S, Delaney BC. Diagnostic accuracy of GPs when using an early intervention decision support system: a high fidelity simulation. Br J Gen Pract 2017; DOI: 10.3399/bjgp16X688417 .

6.Corrigan D, Munnelly G, Kazienko P, Kajdanowicz T, Soler J-K, Mahmoud S, Porat T, Kostopoulou O, Curcin V, Delaney B. Requirements and validation of a prototype learning health system for clinical diagnosis. Learn Health Sys. 2017;e10026. https://doi.org/10.1002/lrh2.10026

7.T Porat, B Delaney, O Kostopoulou. The impact of a diagnostic decision support system on the consultation: perceptions of GPs and patients. BMC Medical Informatics and Decision Making 2017 (1), 79

The programme was initially funded by the European Commission under FP7 247787 TRANSFoRm (2010-2016), and is currently funded by Cancer Research UK Project Grant No C37891/A25310. Contact: brendan.delaney@imperial.ac.uk

Identification of common data elements in disease registries to inform clinical decision support development

Craig S. Mayer, MS[a,b] , Vojtech Huser, MD Ph.D[a]
[a] Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD
[b] ICF Incorporated., Rockville, MD

Introduction

Research disease registries often collect more detailed data elements (DE) compared to routine healthcare records. We assumed that the most useful registry DEs can later be adopted as input for decision support in a research or clinical environment. We worked with registry data dictionaries (DDs) as opposed to trial DDs because we expect registries will have a higher level of DE curation.

This poster presentation is particularly relevant to developers considering decision support in research settings

Methods

Our goal was to identify Common Data Elements (CDEs) in a convenient sample of HIV patient registries. We acquired registry DDs that were either publicly available or available upon simple request.

We defined a CDE as a data element (DE) present in at least 80% (4/5) of analyzed registries. The analysis was done by taking a set of benchmark DEs from the registry with the least amount of DEs (Australian HIV Registry) and combining similar DEs using EHR conventions to generate a set of benchmark DEs. The registry with the least amount of DEs was chosen is it was the likeliest to limit the possibility of a DE being present in the requisite amount of registries.

We then semantically mapped each benchmark DE to the other four registries. Identifying synonymous terms that are equivalent to that of the benchmark DE.

Example of Semantic Mapping

.

Australia	EuroSIDA	Emory	Minnasota	California
VIRAL LOAD	RNA_V	RESULT_ORIGINAL	tst_test_definition_code	vl_recent_value
DATE OF VIRAL LOAD	RNA_D	LAB_DATE	tst_date	vl_recent_dt
EXPOSURE	Mode		cln_risk	transx_categ

Type of DE Grouping

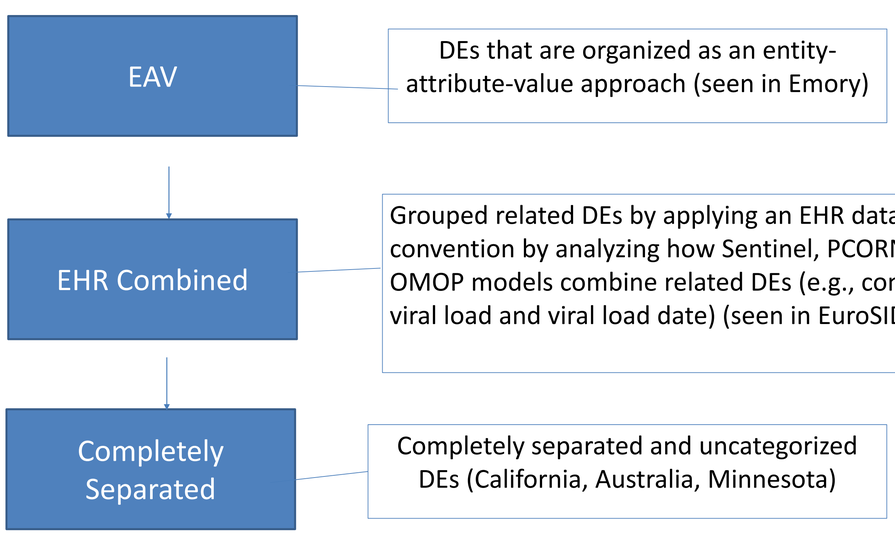


Table 1:Common Data Elements

100% CDE	80% CDE
Date of Birth	Name
Sex	Country of Birth
Clinic Identifier	Residence
Date of HIV Diagnosis	Exposure
Viral Load (value and date)	
CD4 (value and date)	
Date of Last Contact	
Date of Death	

Conclusions

With 12 CDEs out of the 22 EHR convection combined DEs the majority of the DEs in the benchmark registry are common across the other registries.

While other strategies for identifying CDEs can be used, such as data type mapping, permissible value matching, and computerized text matching, manual semantic mapping is a viable method for identifying CDEs in a small sample size of DDs and the use of a benchmark DD.

Results

There were a total of 1120 DEs across the 5 registries ranging from 43 to 665 DEs in a given registry.
The Australian National HIV Registry (benchmark) had the smallest amount of DEs at 43 and reduced to 37 DEs when nation specific elements were removed.

The benchmark DEs further reduced to 22 when EHR conventions were applied.

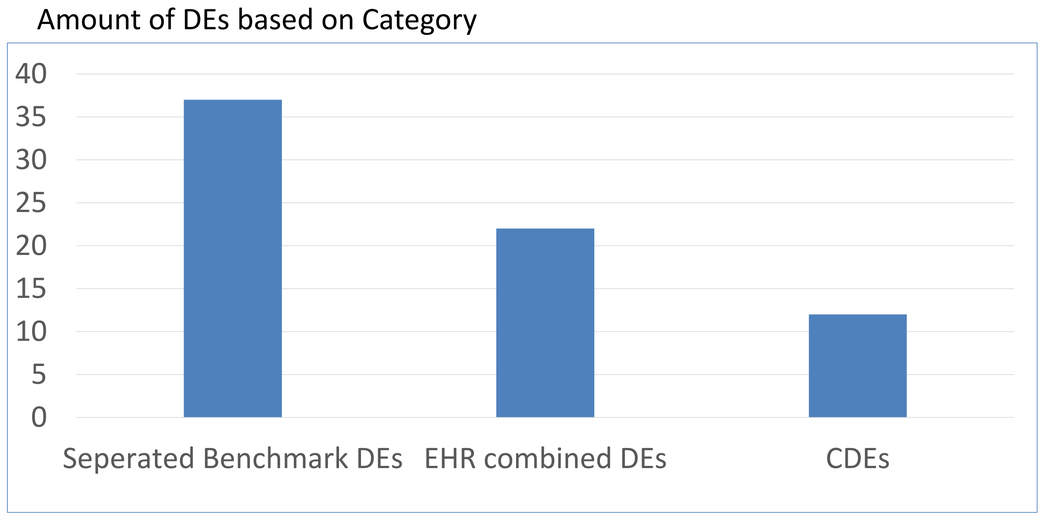
Table 1 shows the 12 Common data elements

A review of data type shows that date is the most common data type that is present in the set of CDEs.

Only the Emory CFAR Registry used popular vocabularies such as ICD, LOINC, and NDC.

Title of Registry
Emory Center For AIDS Research (CFAR) Registry
EuroSIDA: European Patients Infected With HIV
California HIV/AIDS Case Registry
The Australian National HIV registry
Minnesota HIV Surveillance Registry

Amount of DEs based on Category



Number of completely separated data elements

Registry	# of DEs
The Australian National HIV registry	43
EuroSIDA Study on Clinical and Virological Outcome of Patients With HIV (NCT02699736)	102
Minnesota HIV Surveillance Registry	125
Emory Center For AIDS Research (CFAR) Registry	199
VA Clinical Case Registry: HIV	206
California HIV/AIDS Case Registry	665

The project repository at https://github.com/lhncbc/CDE/tree/master/hiv/registries contains additional methods and results.

Acknowledgements: This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC) and by NIH Office of AIDS Research. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services. **Contact:** vojtech.huser@nih.gov

NIH U.S. National Library of Medicine

ICF

Bayesian network transfer learning to improve re-usability of computable biomedical knowledge for public health

Ye Ye[1], Michael Wagner[1], Gregory Cooper[1], Fuchiang Tsui[2], Jeremy Weiss[3], Per Gesteland[45], Jeffrey Ferraro[45], Peter Haug[45]
[1]University of Pittsburgh, Pittsburgh, PA; [2]Children's Hospital of Philadelphia, Philadelphia, PA; [3]Carnegie Mellon University, Pittsburgh, PA; [4]University of Utah, Salt Lake City, UT; [5]Intermountain Healthcare, Salt Lake City, UT
yey5@pitt.edu

Introduction

This study aims to improve the re-usability of computable biomedical knowledge. In particular, we focus on computable biomedical knowledge encoded in probabilistic formalism, such as Bayesian network models. Models can be re-used *in toto*, or the knowledge they encode can be 'adapted' for settings or time periods other than those in which the model was developed. Whether a model can be re-used *in toto* depends on its performance in a new setting and also whether there are fundamental mismatches between the data required by the model and the data available in the new setting.

We previously studied *in toto* model re-use of Bayesian network (BN) models of diagnosis, which we have been using for the automatic detection of cases of infectious disease.[1,2] As depicted in Figure 1A, we used training data and K2 Bayesian network learning[3] to create these BN models (called classifiers) that detect cases of infectious respiratory diseases, such as influenza, from electronic medical record (EMR) data. The training data included variables such as Unified Medical Language System (UMLS) concepts, which are represented by their Concept Unique Identifiers (CUIs). The CUIs were extracted from patient charts automatically using natural language processing (NLP). We machine-learned BN models at Intermountain Health (IH) in Utah. We first tested whether IH models could be re-used *in toto* at the University of Pittsburgh Medical Center (UPMC) in Pennsylvania.

We found that the re-used model lost performance, which motivated the development of a transfer learning algorithm. Transfer learning is a kind of machine learning. It explicitly distinguishes between a source setting, which has the model that we would like to re-use and sometimes the data from which the model was learned/derived, and a target setting, which has data insufficient for deriving a model. We created the Bayesian Network Transfer Learning (BN-TL) algorithm[4] (see methods). As depicted in Figure 1B, BN-TL combines the information in a source BN model with available, local target data to create a target BN model. We then compared its performance versus *in toto* re-use of influenza BN models (see results).

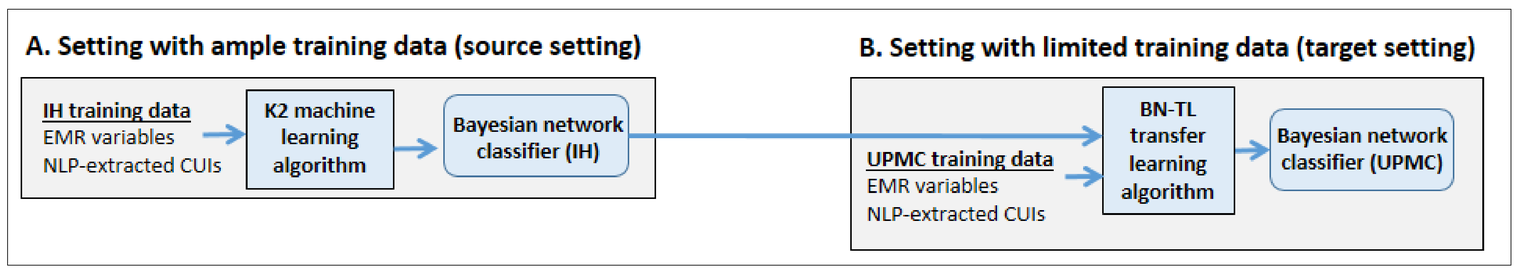


Figure 1. Regular machine learning (A) and transfer learning (B).

Methods

The main idea underlying BN-TL is to modify a source BN model's structure and parameters (conditional probabilities) to more accurately perform classification tasks in a new (target) setting. Figure 2 shows how the BN-TL algorithm starts with a source BN model with six features (colored light blue and labelled *S* to distinguish them from the target features to be added in later steps). BN-TL then applies a sequence of operations: cutting the source model, conducting recurring grow–prune refinement cycles, and assigning conditional probabilities to complete the target model. Briefly, by treating a source model as a prior model, the BN-TL algorithm is able to integrate source knowledge with target data.

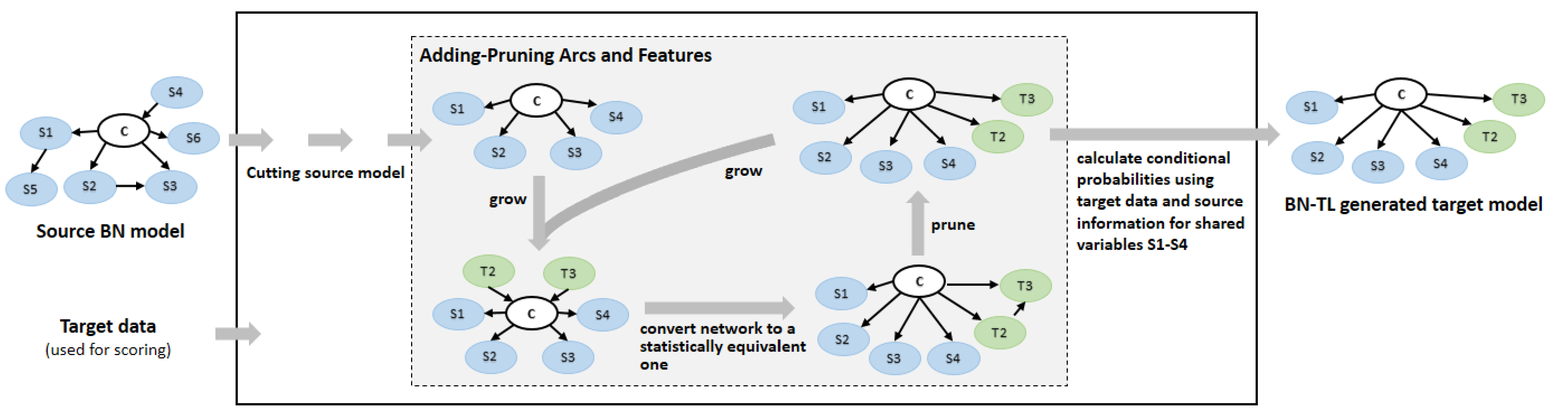


Figure 2. Schematic of the BN-TL algorithm. BN-TL takes as input a source BN model and data from the target setting. It first cuts source features that are not available in the target setting or that have little impact on predicting the class variable. It then enters a grow–prune refinement cycle that ends when there are no "valuable" new features to be added. Finally, BN-TL calculates conditional probabilities to complete the target BN model.

Results

We studied the ability of BN-TL to perform transfer learning from UPMC source models to create IH target models, and vice versa, during the 2014-15 season. We assumed that neither IH nor UPMC had historical data available from previous seasons. Table 1 shows the cumulative counts of positive and negative cases in the early weeks of the season. We used these cases to develop models for each location.

Table 1. Cumulative counts of IH and UPMC ED visits during 2014-15 influenza season

Week	Setting	Data	Laboratory confirmed influenza	Laboratory confirmed influenza negative	Dates
1	UPMC	Train	7	93	2014-11-14 to 2014-11-20
	IH	Train	6	114	2014-11-15 to 2014-11-20
2	UPMC	Train	23	166	2014-11-14 to 2014-11-27
	IH	Train	18	236	2014-11-15 to 2014-11-27
4	UPMC	Train	71	388	2014-11-14 to 2014-12-11
	IH	Train	94	550	2014-11-15 to 2014-12-11
8	UPMC	Train	455	1,311	2014-11-14 to 2015-01-08
	IH	Train	396	1,385	2014-11-15 to 2015-01-08
Greater than 8	UPMC	Test	215	1,965	2015-01-09 to 2015-04-14
	IH	Test	204	3,627	2015-01-09 to 2015-05-05

Table 2 compares the Area under the Receiver Operating Characteristic Curve (AUCs) for the UPMC→IH re-use scenario for the following three approaches to obtaining IH models: (1) learning an IH model from IH data (labeled *Target* in Table 2), (2) *in toto* re-use of the UPMC model (*Source*), and (3) using BN-TL to adapt the UPMC model ($BN\text{-}TL_{UPMC \rightarrow IH}$). The $BN\text{-}TL_{UPMC \rightarrow IH}$ models performed significantly better than IH models at week 2. $BN\text{-}TL_{UPMC \rightarrow IH}$ models also performed significantly better than UPMC models at weeks 1, 2, and 4.

In the other direction (IH→UPMC, Table 3), the $BN\text{-}TL_{IH \rightarrow UPMC}$ models performed significantly better than UPMC models at week 2 and 4 and significantly better than IH models at week 1.

At the early stage of an influenza outbreak, BN-TL-adapted models performed better in the target setting than either *in toto* re-use of source models or target models learned from limited target data.

Table 2. Influenza model performance when sharing from UPMC to IH

Week	Target: BN_{IH}	Source: BN_{UPMC}	$BN\text{-}TL_{UPMC \rightarrow IH}$
1	0.64 (0.60-0.68)[2]	0.58 (0.54-0.62)[13]	**0.67 (0.63-0.70)[2]**
2	0.67 (0.63-0.71)[23]	0.61 (0.57-0.65)[13]	**0.71 (0.68-0.75)[12]**
4	**0.74 (0.71-0.77)[2]**	0.66 (0.62-0.70)[13]	0.73 (0.69-0.76)[2]
8	**0.76 (0.73-0.80)[23]**	0.70 (0.67-0.74)[1]	0.71 (0.68-0.75)[1]

Cells list AUC and confidence intervals for visits in IH (greater than 8 weeks).
The highest value for each week is bolded.
1. significantly different from BN_{IH}
2. significantly different from BN_{UPMC}
3. significantly different from $BN\text{-}TL_{UPMC \rightarrow IH}$
Statistical comparisons using De-Long tests (0.05 as the cutoff for p-value).

Table 3. Influenza model performance when sharing from IH to UPMC

Week	Target: BN_{UPMC}	Source: BN_{IH}	$BN\text{-}TL_{IH \rightarrow UPMC}$
1	**0.60 (0.55-0.64)[2]**	0.56 (0.52-0.60)[13]	**0.60 (0.55-0.64)[2]**
2	0.59 (0.55-0.64)[3]	0.60 (0.56-0.64)	**0.62 (0.58-0.66)[1]**
4	0.60 (0.56-0.65)[23]	**0.66 (0.62-0.70)[1]**	0.64 (0.60-0.68)[1]
8	**0.67 (0.64-0.71)**	0.66 (0.62-0.70)	0.66 (0.63-0.70)

Cells list AUC and confidence intervals for visits in UPMC (greater than 8 weeks).
The highest value for each week is bolded.
1. significantly different from BN_{UPMC}
2. significantly different from BN_{IH}
3. significantly different from $BN\text{-}TL_{IH \rightarrow UPMC}$
Statistical comparisons using De-Long tests (0.05 as the cutoff for p-value).

Conclusions

This is the first study of applying transfer leaning techniques to improve the re-usability of computable biomedical knowledge across institutional boundary. Our Bayesian transfer learning algorithm successfully adapted a Bayesian network model for a second healthcare system.

Acknowledgements

This study was supported by R01LM011370 Probabilistic Disease Surveillance from the National Library of Medicine and Andrew Mellon Fellowship.

References

1. Ye Y, Wagner MM, Cooper GF, et al. A study of the transferability of influenza case detection systems between two large healthcare systems. PLoS one. 2017;12(4):e0174970.

2. Ferraro JP, Ye Y, Gesteland PH, et al. The effects of natural language processing on cross-institutional portability of influenza case detection for disease surveillance. Applied Clinical Informatics. 2017 Feb;8(02):560-80.

3. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine learning. 1992;9(4):309-347.

4. Ye Y. Transfer Learning for Bayesian Case Detection Systems (Doctoral Dissertation, University of Pittsburgh). 2019.

Computer-Interpretable Guideline Processing Tools using Microservice Architectures

Martin Chapman[1], Brendan Delaney[2], Vasa Curcin[1]
[1]King's College London, London, UK; [2]Imperial College London, London, UK

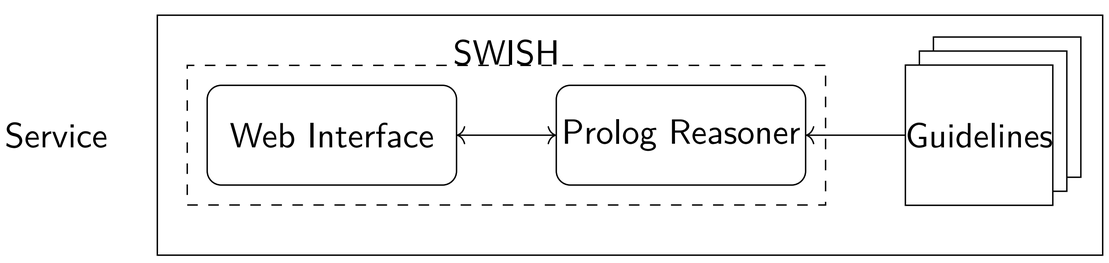
Abstract

Several tools exist that are designed to process computer interpretable guidelines (CIGs), each with a distinct purpose, such as detecting interactions or patient personalisation. While it is desirable to use these tools as part of larger decision support systems (DSSs) doing so is often not straightforward, as their design does not often support external interoperability or account for the fact that other CIG tools may be running in parallel, a situation that will become increasingly more prevalent with the increased adoption of CIGs in different parts of the health system. This results in an integration overhead, system redundancy and a lack of flexibility in how these tools can be combined.

To address these issues, we define a blueprint architecture to be used in the design of guideline processing tools, based on the conceptualisation of key components as REST-ful microservices. In addition, we define the types of data endpoints that each component should expose, for both the communication between internal components and communication with external components that exist as a part of a DSS. To demonstrate the utility of our architecture, we show how an example guideline processing tool can be restructured according to these principles.

Challenge

CIG processing tools consist of a guideline store and a reasoner, e.g.



Traditional, monolithic architectures like this:

- Are geared for human interaction through UIs, rather than autonomous interaction.
- Do not implement standard communication between CIG formalisms (stores) and reasoners. This results in
  - Data redundancy.
  - Limited interactivity across different tools.
- Are difficult to deploy and have limited resilience and scalability.

As such, they do not integrate well with decision support systems.

Benefits

CIG processing tools designed under our architecture:

- Increase interoperability through well-defined services and endpoints.
- Facilitate autonomous communication by providing RESTful endpoints with machine-processable responses.
- Are resilient by having an interaction proxy to the reasoner.
- Scale, by allowing components of the system, such as the reasoner, that may receive heavy load to be replicated.
- Integrate with other processing tools also designed under the architecture
  - Same store, multiple reasoners; reusability.
  - Different stores, same reasoner; technological heterogeneity.

References

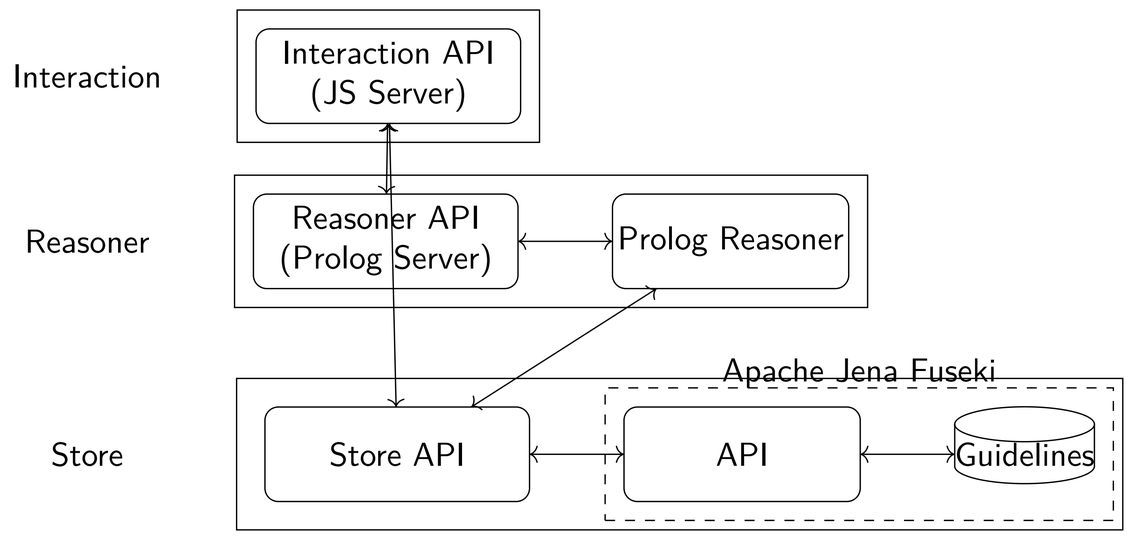
References

[1] Veruska Zamborlini, Jan Wielemaker, Marcos Da Silveira, Cedric Pruski, Annette Ten Teije, and Frank Van Harmelen. SWISH for prototyping clinical guideline interactions theory. In *CEUR*, volume 1795, 2016.

Architecture

Our tool is based on *RESTful microservices*, which separate the features of a system into individual services, and the functionality of each service is invoked using URIs that are defined as a part of an Application Programming Interface (API).

Our architecture defines distinct types of microservices, e.g.



And distinct types of RESTful endpoints, e.g.

Microservice	Endpoints
Interaction	/guideline/create/ /guideline/add/ /guideline/drug/add /guideline/transition/add /guideline/belief/add /guideline/.+/delete /guideline/drug/get /guideline/drug/effect/get /guidelines/interactions
Reasoner	/guidelines/.+
Store	/guideline/.+

Deployment

Tools designed under our architecture have been integrated with two DSSs:

1. *CONSULT*, assisting stroke patients in self-managing their treatments: `https://consult.kcl.ac.uk/`.
2. *ROAD2H*, providing access to healthcare in low and middle-income countries: `http://www.road2h.org/`.

Calls are made to the (redesigned) CIG processing tool by each DSS's decision-making component.

BUILDING A NATIONAL HUMAN PHENOME PLATFORM

Dr Vasa Curcin, King's College London, UK. Prof Jim Davies, Oxford University, UK. Prof Georgios Gkoutos, University of Birmingham, UK. Dr Emily Jefferson , University of Dundee, UK. Dr Helen Parkinson, European Bioinformatics Institute, UK. Dr Gabriella Rustici, University of Cambridge. Mr Daniel Thayer, Swansea University. Dr Spiros Denaxas, University College London, UK.

Background

Despite the widespread use of EHR data for observational and interventional research, current approaches and resources for creating, validating and representing EHR- derived phenotypes and tools are siloed, fragmented and lack a common standard. Unlike the genome, the human phenome lacks a go-to resource for researchers and clinicians to deposit or obtain algorithms, tools, methods and training material. From >1800 UK primary care EHR studies, only 5% included sufficiently-reproducible algorithms. Simultaneously, significant heterogeneity is observed in creating/validating phenotypes (one review reported 66 asthma definitions). Most current methods focus on curating lists of terminology terms rather than complete phenotypes, e.g. implementation logic, evaluation evidence etc.

Health Data Research UK, a national programme for health data science, is developing an open-access, UK- wide approach to standardisation, methods, tools, and validations for defining disease and health-related conditions in a consistent, reproducible fashion across data modalities (structured, text, imaging, sensors, wearables). We anticipate such phenotyping algorithms kitemarked by HDR will become the de-facto authoritative source of consistent, transparent, reproducible and useful definitions of human disease in the United Kingdom.

Research question

Current approaches for creating, validating and representing phenotype definitions from multimodal data (EHR, imaging, text, wearables) are fragmented and lack a common approach. No national, authoritative resource exists for the community to deposit, discover and share phenotyping methods, tools and algorithms. There is a recognized need to establish an open- access resource to enable the community to interact, disseminate best practices and enable reproducible research at scale while maximizing benefits from the data. Consistent and replicable phenotype definitions will enable us to define and redefine disease with multimodal data at national scale, fit for an era in which there is a grand convergence of care, technology, and research to deliver patient benefit.

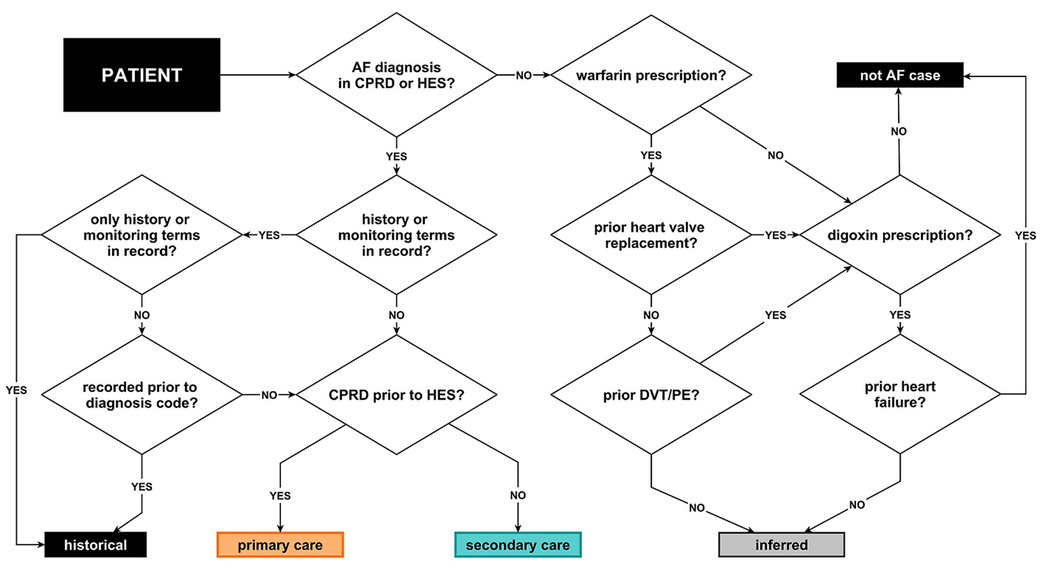


Figure 1: Atrial fibrillation definition in UK CALIBER repository

Objectives

Scoping and Prototyping: To landscape existing national/international approaches for creating, validating and curating multimodal disease phenotypes; gather requirements through stakeholder engagement; define a phenotype presentation metadata standard, and deliver a prototype showcasing exemplars.
Phenome Portal: To build and curate an online, open-access, standards-driven library of complex phenotypes enabling their dissemination, re-use, evaluation, and citation.
Computable Phenotype Model and Tooling: To evaluate computable phenotype representation approaches and build data management tools for common UK EHR datasets.
Training and Capacity Building: To develop and deliver cross-disciplinary training on phenotyping, reproducible science, scientific software development at undergrad, postgrad and continuous professional development (CPD) levels.
Community and Engagement: To ignite and evolve the user community by incentivizing usage and on-going meaningful engagement across stakeholders.

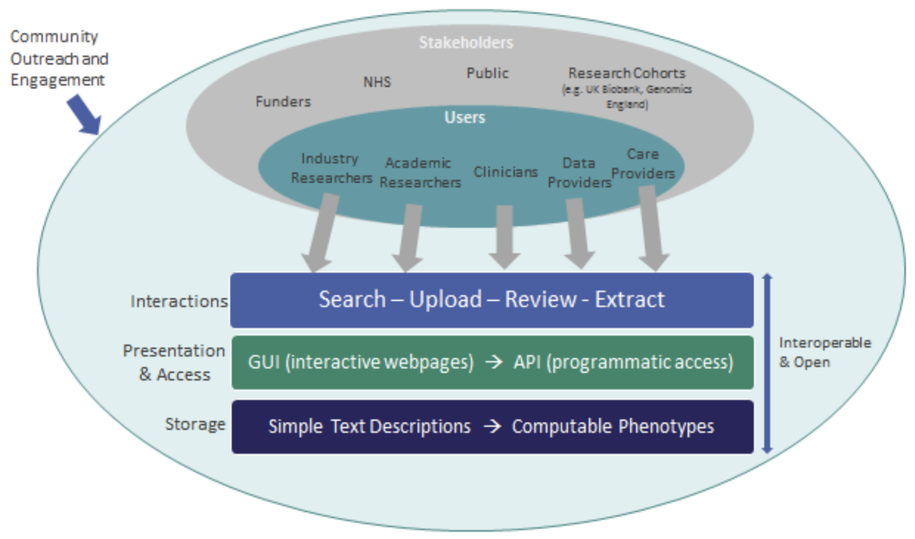


Figure 2: Conceptual overview of the Human Phenome Portal

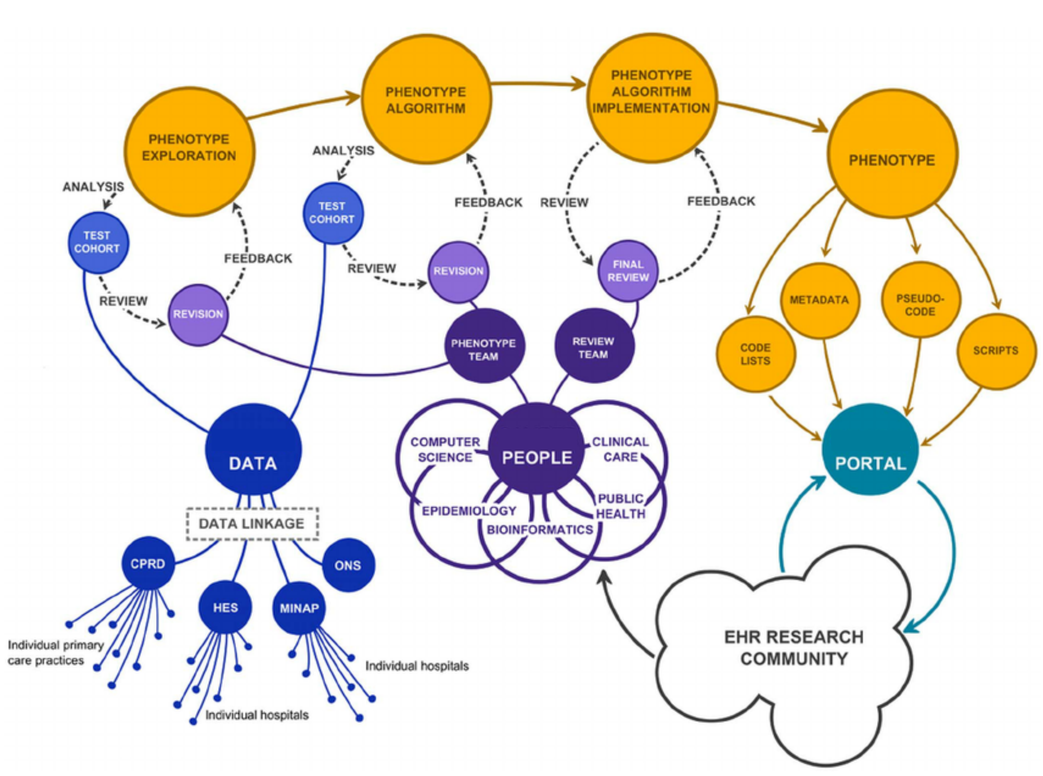


Figure 3: Phenotype lifecycle on the portal

The programme is funded by Health Data Research UK. Contact: vasa.curcin@kcl.ac.uk

Visual representation for knowledge representation:
Graphics libraries to support anatomy terminologies and ontologies

Melissa Clarkson PhD MDes MA
mclarkson@uky.edu

Steve Roggenkamp
steve.roggenkamp@uky.edu

Institute for Biomedical Informatics University of Kentucky

Background

Visual representations are crucial for communicating some types of biomedical knowledge, particularly in the domain of anatomy. Methods for integrating standardized visual representations into knowledge representation schemes are largely unexplored, leaving users dependent upon textual descriptions or ad-hoc visuals.

Objective

We introduce SVG-based graphics libraries as a tool for standardizing visual representations of biomedical knowledge. We use a library for craniofacial anatomy to demonstrate the potential of this approach, but this work is applicable to any domain of biomedical knowledge that can benefit from visual representation.

Design requirements for graphics libraries

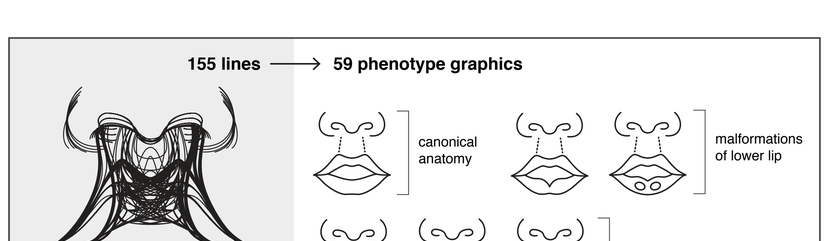
Four characteristics will facilitate use of graphics libraries as standards in knowledge representation and allow the graphics to be customized for use in biomedical informatics applications.

SVG: Scalable Vector Graphic is an XML-based standard. With SVGs, application developers can apply CSS to control the appearance of elements and incorporate interactivity through JavaScript.

Composable: By designing graphics as compositions of graphic elements, the flexibility of a library is maximized while minimizing the number of elements to be produced.

Extensible: Additional graphics will need to be added based on user needs.

Identifiable: In order to serve as standards for knowledge representation, the graphics should have unique identifiers. This will allow authors of terminologies and ontologies to map to the graphics that serve as visual representations of terms and classes.



Demonstration of the "composable" design of the graphics library for cleft lip phenotypes

Implementing web-accessible graphics libraries

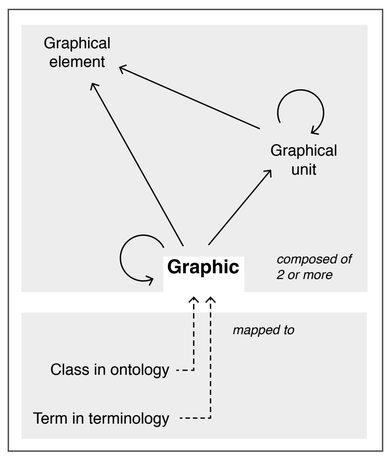
Components

Our approach to composing graphics uses three types of components:

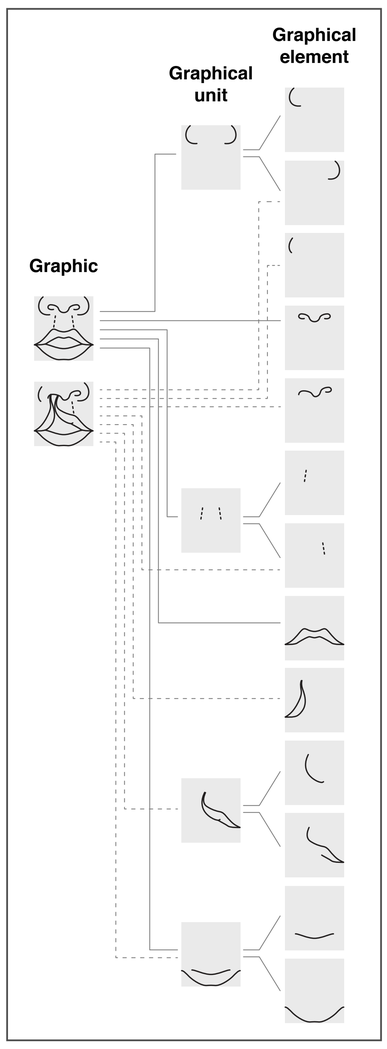
Graphical element: An SVG element of type <path>, <line>, <ellipse>, <polygon>, <polyline>, or <rect>

Graphical unit: Two or more graphical elements or graphical units

Graphic: A complete "ready-to-use" representation composed of two or more graphical elements, graphical units, or graphics.



Scheme for composing graphics from graphical units and graphical elements



Example from the graphics library for cleft lip phenotype

Identifiers and tags

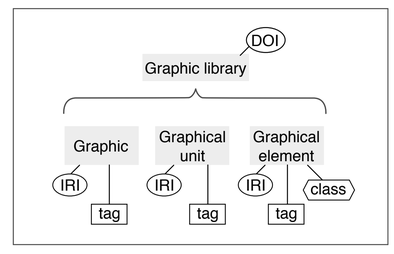
Four types of tags and identifiers are used:

DOIs: A graphic library will be identified by a DOI.

IRIs: Each graphic, graphical unit, and graphical element receives a unique IRI.

Tags: Text tags can be applied to graphics, graphical units, and graphical elements to aid in retrieval through the web interface.

Classes: Classes are applied to the graphical elements to control appearance with CSS.



Scheme for identifiers and tags

Visit our web application at
endlessforms.info

We thank Jim Brinkley and José (Onard) Mejino of the Structural Informatics Group at the University of Washington for support during the early phases of this project.

Using a Knowledge Repository to Integrate Clinical Documentation Metadata and Reference Standards

Karen M. Bavuso RN, MSN,[1] Saverio Maviglia MD, MS,[1,2,3] Melinda Wong BS Pharm, RPH,[1]
Emily Man Wai Chan, PharmD,[1] Roberto A. Rocha MD, PhD[1,2,3]

[1] Semedy, Inc., Waltham, MA; [2] Brigham and Women's Hospital, Boston, MA; [3] Harvard Medical School, Boston, MA

Introduction

- Consistent representation and integration of clinical documentation entities promotes data integrity, facilitates point-of-care data usage, and enables secondary data retrieval for research and quality improvement. [1-2]
- Lack of consistency in managing clinical documentation metadata has detrimental consequences to organizations, including disruption of operational processes and suboptimal clinical outcomes.
- We illustrate how metadata, version history, and reference terminology bindings support successful management and retrieval of documentation entities and related artifacts.

Methods

- Development of a Clinical Knowledge Management System (CKMS) that provides domain-specific properties, extensible metadata and provenance, and detailed revision history.
- Design of domain models to enable the representation of structural interdependencies across forms, data elements, value sets, and terminology concepts.
- Integration of documentation entities from different sources.
- Utilization of reference terminology bindings allowing for a wide-range of semantic interconnections.
- Creation of customized semantic queries to identify content mappings.
- Implementation of presentation templates to enable visualization of clinical documentation content details.

Results

Metadata	Instance values
Lifecycle state	Work in progress
Revision number	2
Source	LOINC Code System
Creator	Karen Bavuso
Date created	October 27, 2018 5:41:48 PM
Date last modified	July 10, 2019 11:38:43 PM

Form model	Instance values
Long name	Skin assessment panel
Data element declaration	Body site panel
	Color of skin
	Moisture of skin
	Temperature of skin
	Turgor of skin
	Skin integrity
	Pressure points examined of skin
	Mucous membrane integrity
	Skin assessment
Purpose	Used for routine skin assessment data collection
Topic	Skin assessment panel (LOINC)
Source	LOINC Code System

Data element model	Instance values
Long name	**Color of skin**
Datatype	Enumerated
Topic	Color of Skin (LOINC)
Value domain	Color of skin

Value set model	Instance values
Long name	**Color of skin (Value set)**
Value set members	Normal Cyanotic Mottled Jaundice Pale Flushed Lividity Red (erythematous)

Figure 1: Sample properties and values across clinical documentation entities

- Essential metadata details captured and stored across models.
- Custom domain models to represent clinical documentation entities.

Property name	7/11/2019 (Rev. 2)	10/27/2018 (Rev. 1)
Lifecycle state	Under review	Published
Long name	Color of skin (Value set)	Color of skin (Value set)
Value set members	Normal	Normal
	Cyanotic	Cyanotic
	Mottled	Mottled
	Jaundice	Jaundice
	Pale	Pale
	Flushed	Flushed
	Pink	
	~~Lividity~~	Lividity
	Red (erythematous)	Red (erythematous)

Figure 2: Sample visual of content changes across revisions

- Ability to track revision history and visualize content changes.

- Unique properties allowing for reference terminology binding, content categorization, and mapping.

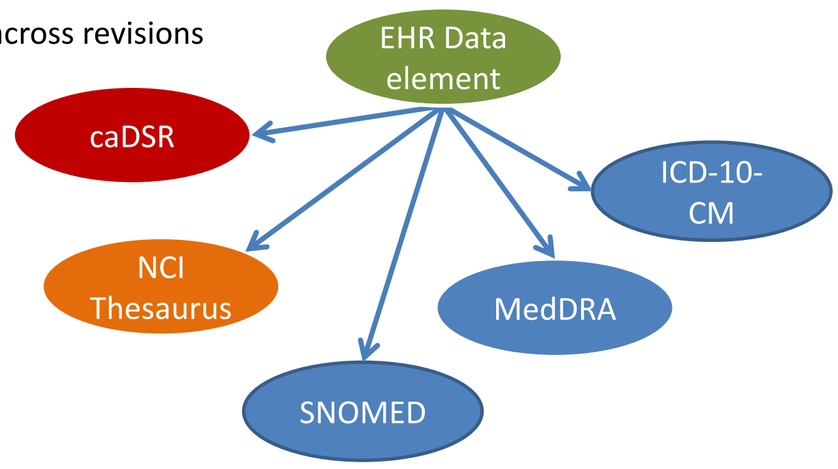


Figure 3: Clinical documentation entity with semantic relationships

Conclusions

- We have successfully integrated documentation entities from different sources.
- The repository provides an opportunity for stakeholders to search, visualize, annotate, and track the evolution of documentation entities and related artifacts.
- The integrated repository becomes an essential resource for managing interconnections and downstream dependencies, expediting troubleshooting efforts and preventing content malfunctions.
- Uniform domain models and bindings to reference terminologies (e.g. NCIt, MedDRA, ICD-10-CM, SNOMED CT), enable semantic queries to support research, analytics, and data interoperability initiatives.

References

1. Rocha RA, Maviglia SM, Sordo M RB. Clinical knowledge management program. In: Greenes R, ed. *Clinical Decision Support - The Road to Broad Adoption (Second Edition).* Burlington; 2014:773-817.
2. Cusack CM, Hripcsak G, Bloomrosen M, et al. The future state of clinical data capture and documentation: a report from AMIA's 2011 Policy Meeting. *J Am Med Informatics Assoc*. 2012:134-140. doi:10.1136/amiajnl-2012-001093

Interoperable Knowledge Modeling in Emergency Care:
Application of the HL7 Emergency Care Domain Analysis Model

James C. McClay, MD MS. University of Nebraska Medical Center, Omaha, NE
Laura Heermann Langford, RN, PhD, Intermountain Health Care, Salt Lake City, UT

Abstract

The Emergency Care System (ECS) in the United States provides immediate unscheduled care to an increasing number of patients annually. Every ECS patient encounter generates data in a myriad of unlinked systems at regional emergency services and hospitals. Often, patients' prior health records are not available in a timely manner, fragmented, or not found. To address availability and interoperability of data relevant to emergency care, the Health Level Seven (HL7) Emergency Care Workgroup collaborates to ensure evolving health data interoperability standards incorporate ECS considerations. Foundational to that effort is a comprehensive model of Emergency Care information, the Emergency Care Domain Analysis Model (EC-DAM).
The EC-DAM interlinks models of controlled vocabulary, information artifacts, health record functions and emergency department workflow. This poster describes the design considerations and the impact of the EC-DAM on related information standards such as FHIR resources in prehospital care, trauma care, disease surveillance, and clinical research. The EC-DAM provides an integrated, standardized platform for creation of interoperable EC related computable knowledge tools. This poster presentation is particularly relevant to developers of health information standards and health information systems that involve the ECS.

Background

- The U.S. Hospital Based Emergency Care system manages 140 million annual visits in a high risk, low information environment.
- The Emergency Care (EC) system greatly benefits from shared knowledge resources.
- Health Level 7 (HL7) is the ANSI accredited Standards Development Organization for healthcare information interoperability specifications.
- The HL7 Emergency Care Workgroup (ECWG) provides oversight to information standards impacting EC.
- The ECWG developed the EC Domain Analysis Model (EC-DAM) to coordinate standards for sharing knowledge resources impacting EC practice.
- The EC-DAM provides a model framework of EC data, workflow and information systems functions for use in implementing EC related knowledge resources such as protocols, clinical scoring, FHIR resources, and clinical guidelines.
- We describe the EC-DAM and present a simple example of use.

Methods

- HL7 Standards development through open, consensus processes, procedures and policies based on established an Governance and Operations Manual
- HL7 Domain Analysis Modeling Process outlined in reference 1
- The data model for the EC-DAM is the Data Elements for Emergency Department Systems (McClay 2015)
- The information model is compiled from HL7 Clinical Elements Models and the Version 3 Reference Information Model
- The EC Business Process Model is the result of consensus input from EC informatics experts, ED IS vendors and HL7 modelers
- The Emergency Department Information System Functional Profile is derived from the HL7 Electronic Health Record Functional Model (Hammond 2008)
- The components of the EC-DAM were utilized to model capture of the Glasgow Coma Scale in stroke assessment protocols for this demonstration

Results

- HL7 Emergency Care Domain Analysis Model integrates information standards for data, information modeling, workflow and information system functions
- The EC-DAM framework supports interoperable modeling of Information System requirements to support clinical processes involving hospital based emergency departments.

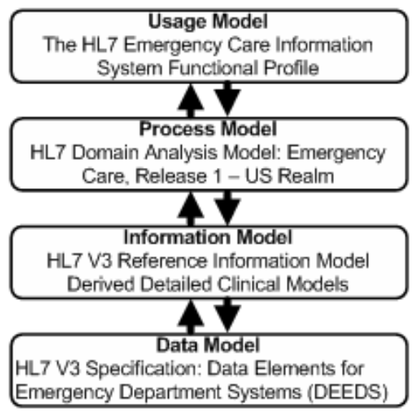


Figure 1: Component Models of EC-DAM:
Integration of standards for represent EC information.

Acknowledgements:
This work is supported by the many volunteers serving the HL7 community, especially the participants in the HL7 Emergency Care Workgroup

"The nice thing about standards is that you have so many to chose from"
- Andrew S. Teanenbaum, Computer Networks, 1981, p168

The Emergency Care Domain Analysis Model

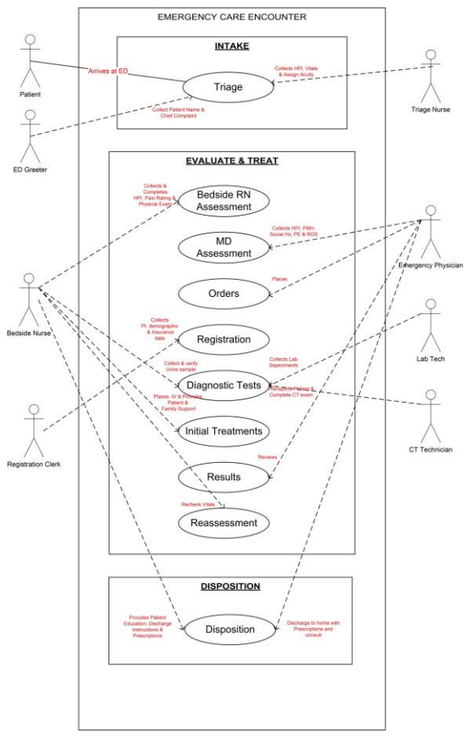


Figure 2: Example EC-DAM Use Case Diagram
Showing actors and activities in emergency care encounter

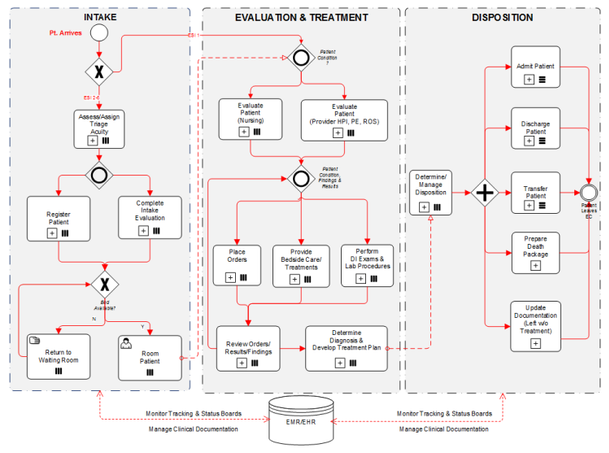


Figure 3: Business Process Model of Emergency Care Encounter
Generalization of EC processes indicating typical flow and options during an encounter. Each box is expanded in document to indicate data required and information system functions. Actors are indicated in use case diagrams

Application of EC-DAM in stroke protocol

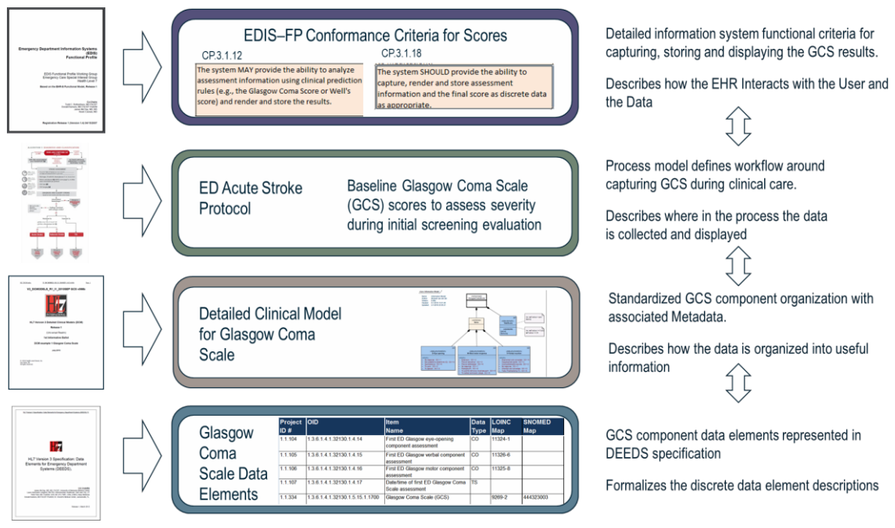


Figure 4: EC-DAM components for GCS in stroke: Inter-relationship of components of EC DAM in creating a scoring tool for Glasgow Coma Scale. Data Elements from DEEDS, information model from CIMI, workflow model for stroke, information system components for tool.

Discussion

- Large scale deployment of interoperable knowledge artifacts depends on a shared set of models for data, process and usage.
- The HL7 Domain Analysis Modeling Process provides a framework for integrating and extending existing clinical interoperability standards across different clinical domains
- The EC-DAM represents one of a family of HL7 domain analysis models representing a spectrum of clinical conditions.
- The open, consensus development process allows free adoption and reuse of standards by the knowledge management community

References

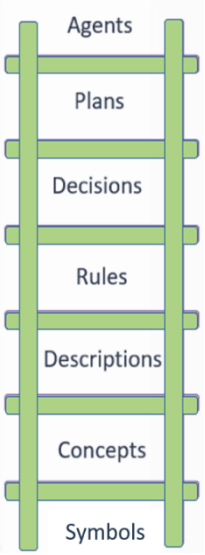
HL7 Clinical Interoperability Council, "HL7 Specification: Domain Analysis Model Specifications and Requirements - Canonical Definition, Release 1" (2017) http://www.hl7.org/implement/standards/product_brief.cfm?product_id=463 (accessed 6/29/2019)

HL7 Emergency Care Workgroup, "HL7 Domain Analysis Model: Emergency Care, Release 1 – US Realm" (2016) http://www.hl7.org/implement/standards/product_brief.cfm?product_id=421 (accessed 6/29/2019)

McClay JC, Park P, Janczewski M, Langford L (2015); Standard for Improving Emergency Information Interoperability: The HL7 Data Elements for Emergency Department Systems (DEEDS). *J Am Med Inform Assoc*; DOI: http://dx.doi.org/10.1093/jamia/ocu040

Hammond, W. E. (2008). eHealth interoperability. Studies in health technology and informatics, 134, 245.

Artificial intelligence and knowledge engineering offer the next revolution in medical digital publishing: www.OpenClinical.net



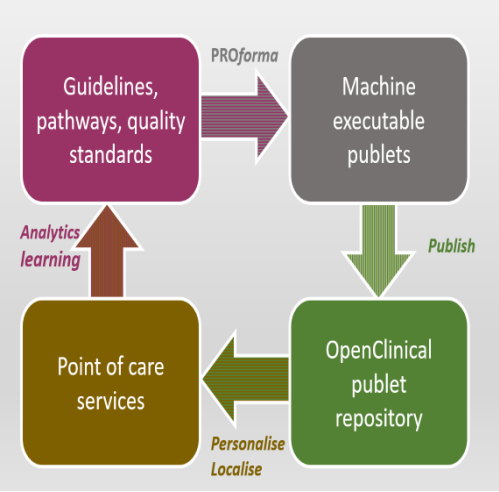
Countless medical organisations around the world publish clinical guidelines and other standards of best practice as a basis for improving quality and safety of care and reducing costs. Major publishers include AHRQ's *National Guideline Clearing House* in the USA and NICE, the *National Institute for Health and Care Excellence* in the UK. Most clinical guidelines use traditional text documents, checklists and flow diagrams in order to set out recommendations for care.

Unfortunately, healthcare professionals rarely have time to read and absorb clinical guidelines and they do not achieve their potential benefits. **OpenClinical has developed a novel technology and platform for creating and publishing best practice as executable models of care.**

AI and knowledge engineering

The OpenClinical method for modelling clinical guidelines as machine-executable services is based on the *knowledge ladder* (above left) using the PRO*forma* guideline modelling language (*JAMIA*, 2003). Experience has shown that OpenClinical tools are simple and practical for building many kinds of decision-support systems, care planning and workflow management services in many clinical settings and specialties. **There are currently more than 50 examples of executable services on the OpenClinical repository and many successful published trials (table right).**

Clinical tasks	Trials and evaluations
Routine prescribing **by GPs**	Walton et al *British Medical Journal* 1996
Mammographic screening	Taylor et al *Medical Image Analysis* 1999
Genetic risk assessment	Emery et al *British Medical Journal* 1999, 2000
Genotyping and prescribing **antiretrovirals**	Tural et al, *AIDS* 2002
Chemotherapy prescribing **for ALL**	Bury et al, *British J Haematology* 2005
Early referrals **of suspected cancer**	Bury et al PhD thesis (2006)
Cancer risk assessment, investigation	Patkar et al, *British J Cancer* 2006
Hospitalisation decisions**: asthma**	Best Practice Advocacy Centre, NZ 2009
Genetic risk assessment, treatment planning	Glasspool et al, *J Cancer Education* 2010
Multidisciplinary decision making	Patkar et al, *BMJ Open* 2011
Investigation, diagnosis **of thyroid nodules**	Peleg et al, *Endocrine Practice*
Guideline compliance **in stroke**	Ranta et al, *Neurology* 2015
Shared decision-making**, chemotherapy**	Miles et al, *BMJ Open*
Diagnosis **of hyponatremia**	Gonzales et al *Int J Med Informatics* 2017
Detection **and** Diagnosis **of ophthalmic disease**	G Chandrasekaran PhD Thesis (2017)
Kidney donor-patient eligibility	Knight et al *Transplantation medicine* (in review)



Publets: a new concept in knowledge-publishing

PRO*forma* models are the *content* of a new class of AI product called *publets* • OpenClinical provides model researchers and developers with the tools to create, test and distribute publets for use at the point of care • When in use publets capture structured patient data for subsequent analytics, research and machine learning as illustrated in the *knowledge-data lifecycle* (left) • **Publets can be tested and trialled on the OpenClinical server, then published and potentially licensed to 3rd parties for routine clinical deployment.**

More information about the OpenClinical concept and technical details are available from the OpenClinical team (info@openclinical.net)

SRDR 2.0

Make systematic review information more clear and accessible for decisionmakers

Ian Saldanha, Randy Jap, Bryant Smith*, Jens Jap, Birol Senturk, Joseph Lau, Ethan Balk. (* presenter)

BROWN School of Public Health

AHRQ Agency for Healthcare Research and Quality

USER NEEDS

- Systematic reviews underpin health decisions, but are voluminous documents that are not easily accessible for daily decision making
- The Systematic Review Data Repository 2.0 (SRDR 2.0) is a platform for sharing and aggregating summary review data digitally and interactively
- SRDR 2.0 can make outcomes, results, and other information readily accessible to decisionmakers

VALUE PROPOSITIONS

- FREE and OPEN ACCESS
- Interactive and user-friendly
- Ability to filter by:
  - ◇ population factors (e.g., age)
  - ◇ intervention aspects (e.g., dose)
  - ◇ outcomes
  - ◇ study designs (e.g., randomized trials)
  - ◇ settings (e.g., outpatient)
- Promote data transparency and interoperability

UP NEXT…

- User testing of the SRDR 2.0 prototype
- Broaden the platform's support for different systematic review formats
- Promote use of SRDR 2.0 for guideline development, clinical decision support, and decision-making in health

DO YOU HAVE IDEAS?

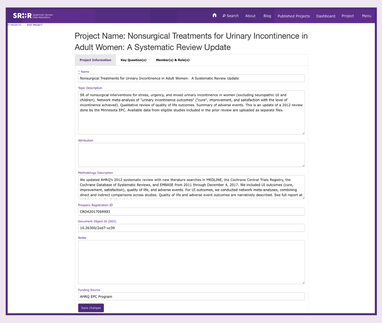
- Do you have ideas for additional features that we should incorporate?
- Would you use SRDR 2.0 for accessing systematic review info?

ian_saldanha@brown.edu

@saldanha_ian

SRDR 2.0 Workflow

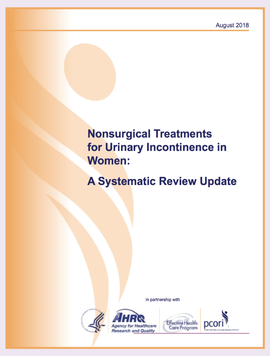
SRDR+ Project*



Contributor Submission

+

Published Report

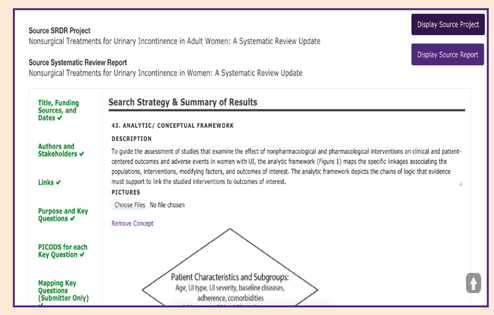


→ **NLP Aided Extraction**

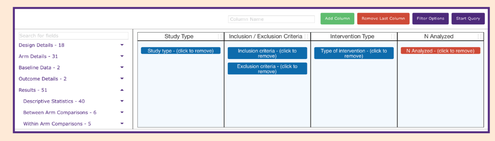
Structured Data in SRDR 2.0

End User Access

View and Share Structured Data



Export Custom Evidence Tables



* SRDR+ is a free, online, collaborative platform for extracting and archiving systematic review data.

OUR SOLUTION

1. Standardize data

All systematic review and report data are categorized into key elements to improve organization and searchability

- **Level 1**: • Meta data • Documentation (Executive Summary, Protocol, etc.)
- **Level 2**: • Populations • Outcomes • Interventions • Designs • Comparisons
- **Level 3**: • Results • Study Quality

2. Make Data Processable

Includes computable data structures and alleviates data extraction processes by applying natural language processing to systematic review data

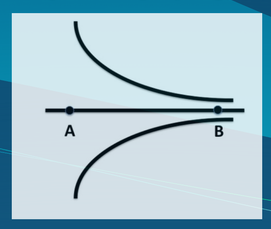
SRDR Systematic Review Data Repository

LEARN MORE AT

https://srdrplus.ahrq.gov

Medication Indication Formalization: A Preliminary Model

Stuart J. Nelson, MD, FACP, FACMI
George Washington University
Biomedical Informatics Center
Stuart.james.nelson@gmail.com



Mark S. Tuttle, FACMI
marktuttle@pacbell.net

How do we represent the indications for a medication in a computable manner?

INDICATION FROM LABEL:

REDACTED is indicated for the treatment of patients with cystic fibrosis (CF) aged 12 years and older who are homozygous for the F508del mutation or who have at least one mutation in the cystic fibrosis transmembrane conductance regulator (CFTR) gene that is responsive to tezacaftor/ivacaftor based on in vitro data and/or clinical evidence [see Clinical Pharmacology (12.1) and Clinical Studies (14)]. If the patient's genotype is unknown, an FDA-cleared CF mutation test should be used to detect the presence of a CFTR mutation followed by verification with bi-directional sequencing when recommended by the mutation test instructions for use. REDACTED is a combination of tezacaftor and ivacaftor, indicated for the treatment of patients with cystic fibrosis (CF) aged 12 years and older who are homozygous for the F508del mutation or who have at least one mutation in the cystic fibrosis transmembrane conductance regulator (CFTR) gene that is responsive to tezacaftor/ivacaftor based on in vitro data and/or clinical evidence. (12.1, 14) If the patient's genotype is unknown, an FDA-cleared CF mutation test should be used to detect the presence of a CFTR mutation followed by verification with bi-directional sequencing when recommended by the mutation test instructions for use.

QUESTIONS:
Is this helpful? Can you remember it? There are 3000+ NDAs, 10,000+ labels

EARLIER FORMALIZATION ATTEMPTS - DRUG-DISEASE ASSOCIATIONS:

REDACTED – Cystic Fibrosis

QUESTIONS:

Helpful?	Maybe
Necessary?	Yes
Sufficient?	For a significant minority of drugs (E.g., insulin and Type I DM)
Could I find this in an EHR?	For most of the minority

NEW MODEL:

DrugID	IndicationID	Condition	Condition_Value	Use	Special	Note
3	2	Disease	Cystic Fibrosis	treatment		

Indication ID	2	Context	Mutation	Context_Value	F508del
		Context		Context_Value	

DrugID	IndicationID	Condition	Condition_Value	Use	Special	Note
3	3	Disease	Cystic Fibrosis	treatment		

Indication ID	3	Context	Mutation	Context_Value	CFTR gene
		Context	Lab result	Context_Value	susceptibility

QUESTIONS:

Is this more helpful?	Your thoughts here
Why?	Enables clinical complexity management
Sufficient?	For humans, maybe.
Could I find this in an EHR?	No

PRELIMINARY OBSERVATIONS:

1. Model Fits 97% or more of indications
2. Many contextual expressions highly abstract (not likely represented in EHR)

NEXT STEPS:

1. Learn what in label content and in label model is represented in HER standard terminologies.
2. Address abstractions not currently findable using standard terminologies.

REFERENCE:

Nelson SJ, Oprea TI, Ursu O, Bologa CB, Zaveri A, Holmes J, Yang JJ. Mathias SL, Mani S, Tuttle MS. Formalizing drug indications on the road to therapeutic intent. *J Am Med Inform Assoc*, ocx064, https://doi.org/10.1093/jamia/ocx064

SUPPORT: NIH/ NLM grant number R21LM012929

REDACTED brand name is Symdeko

The NCI Cancer Research Data Commons (CRDC)

Erika Kim, PhD, Tanja Davidsen, PhD, Juli Klemm, PhD, Allen Dearry, PhD, Tony Kerlavage, PhD

Cancer Informatics Branch, Center for Biomedical Informatics and Information Technology, National Cancer Institute, National Institutes of Health, Rockville, MD

Abstract

Basic and clinical research in cancer is increasingly focused on generation of rich datasets to identify the molecular basis for disease and to match targeted therapies that factor in each patient's unique biology. To progress towards this goal, the cancer research community will need to access, integrate, and analyze many different types of data, including genomics, proteomics, multi-resolution, multi-modality imaging data, cancer models, clinical treatment and outcomes, population-based data, and data contributed by health care providers and patients themselves. Investment in the informatics infrastructure to fully leverage these diverse data types is imperative and was called out as a priority by the Cancer Moonshot℠ Blue Ribbon Panel. To this end, NCI has initiated development of a Cancer Research Data Commons (CRDC) to provide access to interoperable data repositories, analysis tools, and workspaces. The vision for the NCI CRDC is a virtual, expandable infrastructure that provides secure access to diverse data types, allowing users to analyze, share, and store results, leveraging the storage and elastic compute of the cloud. This poster presentation is particularly relevant to community stakeholders with interests in providing input to CRDC growth or learning more about developing data commons to serve CBK community.

NCI Cancer Research Data Commons

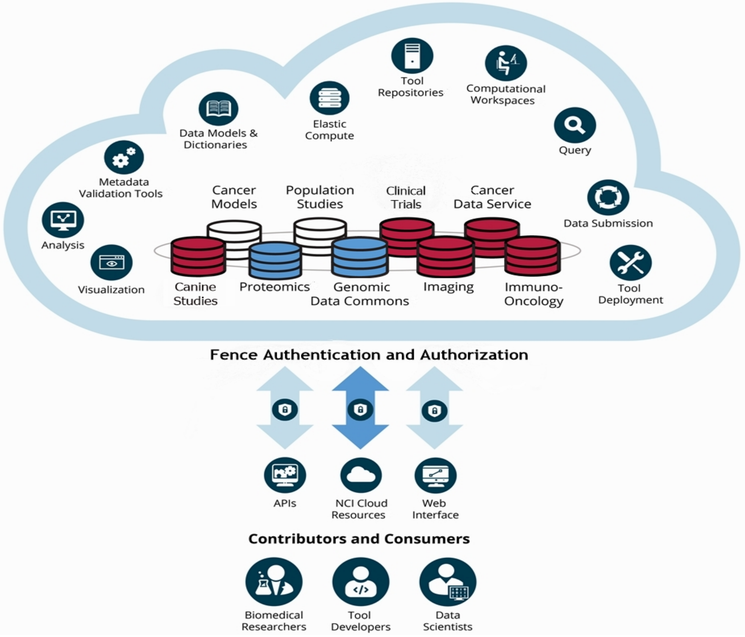


Figure 1: Cancer Research Data Commons. The CRDC is a virtual, expandable infrastructure that supports collaboration among stakeholders. Blue cylinders represent nodes that are operational and red cylinders represent nodes that are currently under development. White cylinders represent examples of anticipated future nodes.

NCI CRDC Goals

- Enable the cancer research community to share diverse data types across programs and institutions
- Provide secure access to data, regardless of where it is stored
- Provide mechanisms for innovative tool discovery, visualization, analysis using elastic compute
- Help NCI Data Coordinating Centers sustain and share their data publicly

CRDC Components

Data Commons Framework

The DCF is a framework of modular and reusable core components that will provide common approaches to functions that are required across the CRDC.

Fence A&A | API & Container Environment | Consistent Data Models | Metadata Validation Tools | Collaborative Workspaces

Data Nodes

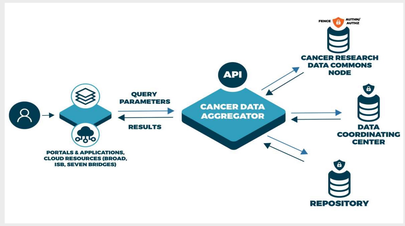
A node is a repository in the CRDC containing related data that have been harmonized and stored in a format that is ready for analysis by the research community. The data are brought together with infrastructure for security, interoperability, and elastic compute capability. Operational nodes include Genomic Data Commons and Proteomic Data Commons.

Center for Cancer Data Harmonization

The Center for Data Harmonization will have a mandate to provide tools, guidance, and training for harmonization for use of multi-modal research and clinical data. The CCDH will:

- Create a harmonized data model and cross-map across nodes
- Provide semantic concierge services to CRDC nodes and NCI Data Coordinating Centers related to data models, metadata, and terminology.
- Create, adapt, and disseminate data harmonization tools
- Develop new terminology, metadata, mappings, and models to support data aggregation through the CRDC

Cancer Data Aggregator



A federated query engine consisting of an API layer to enable queries across:

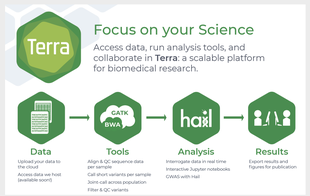
- NCI Cancer Research Data Commons (CRDC)
- NCI Data Coordinating Centers (e.g. HTAN DCC)
- Additional Repositories (e.g. KidsFirst DRC)
- Proposals due by August 15th, 2019: https://go.usa.gov/xyKe3

What is a Data Commons? Data Commons co-locate data, storage and computing infrastructure, and frequently used tools for **analyzing** and **sharing data** to create an **interoperable** resource for the research, clinician, and patient communities.

NCI Cloud Resources

BROAD INSTITUTE

The Broad Institute's Cloud Resource, FireCloud, was built on the Broad Institute's FireHose data analysis pipeline that has been used by over 1000s of researchers. FireCloud harnesses the elastic compute capacity of Google Cloud Platform to enable users to perform large-scale genomic analyses akin to those available through FireHose. Key advantages include running Broad's best practice tools and pipelines now powered by Terra. FireCloud users can also access curated open and controlled-access TCGA workspaces, upload their own data, and share workspaces.



Broad Institute	
	• PI: Anthony Philippakis • Google Cloud • 100s of workflows • http://firecloud.org, terra.bio

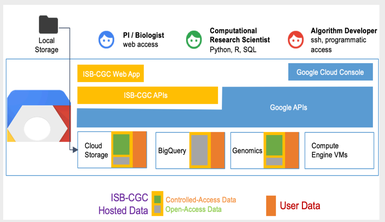
CANCER GENOMICS CLOUD SEVEN BRIDGES

Currently, over 2,000 researchers representing over 40 countries are using SBG to analyze hosted genomic data, or their own data . This Cloud Resource hosts over 11,000 cases of TCGA data along with CCLE, TARGET, TCIA, Simons Genome Diversity Project data, and over 5,000 applications added by users. They offer both an API and a user-friendly visual interface, to ensure that the Cloud Resource is accessible to computer scientists, bioinformaticians, bench-scientists, and clinicians alike. Users can deploy their own containerized tools, use hosted tools, and build analysis pipelines.

Seven Bridges Genomics	
	• PI: Brandi Davis-Dusenbery • Amazon Web Services • > 405 tools in the Public Apps library • http://www.cancergenomicscloud.org

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

The Institute for Systems Biology Cloud Resource runs on Google Cloud Platform and offers an interactive web-based application designed for biologists and clinicians, and hosts APIs, such as the Global Alliance for Global Health (GA4GH) API, for computational scientists and algorithm developers.



ISB-CGC takes advantage of Google Cloud Platform's built-in resources such as BigQuery, Compute Engine, App Engine, Cloud Datalab, and Google Genomics. Researchers can use BigQuery to explore clinical, biospecimen, and level-3 open access TCGA data. The ISB platform was used to analyze several datasets from the recent TCGA PanCan analyses.

Institute for Systems Biology	
	• PI: Ilya Shmulevich • Google Cloud including BigQuery • Interactive visualization and analysis • http://cgc.systemsbiology.net/

#NCICommons https://datascience.cancer.gov/data-commons

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contracts HHSN261200800001E, HHSN261201400006C, HHSN261201400007C, and HHSN261201400008C. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Placing Clinical Guidelines into the Electronic Medical Record

Barry H Ginsberg, MD, PhD, Team Leader, Digital Health Information Program, American Diabetes Association

Introduction: Many medical professional organizations produce standards of care and care pathways (Guidelines) for treatment of patients with specific diseases. These Standards are not widely used.[1,2] Getting an appropriate Standard or pathway in front of a physician at the time of a clinical visit increases the probability that the Standard/pathway will be used and will produce better medical outcomes[3-6].

Methods: To this end, we will form an association of representatives of the leading professional medical organizations, the EMR providers, hospitals, health care providers, IT professionals, regulators. Governmental entities and payors, to develop a standard system to migrate these clinical guidelines from any organization into any EMR. Appropriate guidelines could be selected by medical provider organizations such as hospitals, health groups, clinics or medical offices to improve care at their facility.

- **Voluntary Health Organizations (VHO) and Professional groups develop Guidelines**:
  - Recommendations
  - Care Pathways
- **Used by Lawyers, Payers and some Specialists**
- **Not widely used by PCPs[1,2]**
  - Not readily available to them
  - Large and complex documents (ADA Standards are 193 pages)
  - Not prioritized
  - Not available at Point of Care

- **Remind the Health Care Provider(HCP) of Guidelines that may help their patient[3]:**
  - Timely, at the point of care
  - Specific, for their patient / condition
  - Simple, no added effort to get reminders
  - Limited, only display the most important reminders
- **Incorporate it into the EMR the HCP is already using**
- **Overcome Therapeutic Inertia[7]**

Vision Statement (Overview)

We want to develop an "Agnostic" system:

- **Define a "Data Container" to digitize Guidelines**
- **Voluntary Health Organizations and Professional Societies will digitize and own their Guidelines**
- **Provide a standard for an interface with HCP**
- **EMR Vendor "connects" to interfaced guidelines**
- **Medical Organization may select the Guidelines that best fit their system's needs**

EMR Interoperablity: A Problem[8,9]

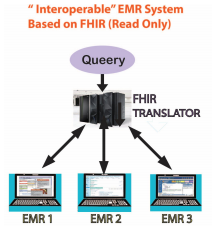
- There is no standard for how to keep clinical data in an EMR
- The same information is kept in different places with different names in different databases

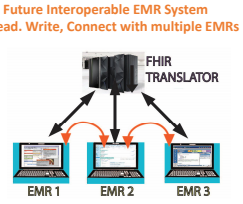
Visionary EMR	HardWorking EMR	EasyToUse EMR
Vascular System	8/17/18	Foot
Foot	Foot	Vascular System
Pulses	Vascular System	Pulses
R Dorsal Pedal	Pulses	R Dorsal Pedal
9/15/12	R Dorsal Pedal	9/15/12
8/17/18	[17045-R]	8/17/18
[VFDPR]		[FVP-R21]

- **So in EMR1 a digital pulse is stored in location VFDPR in EMR2 in 17045-R and in EMR3 in FVP-R21.**
- **Currently, there is no consistency!**

Interoperability:[10]

- HL-7 talks to all EMRs ▸ Provides laboratory data
- Many EMRs can talk to others using the same vendor
- HL-7 Developed FHIR (Fast Healthcare Interoperability Resource)
  - Talks to most EMRS, Can read most fields (often not in real time.)
- Serves as a platform for Personal Health Record (Like AppleHealth)
- Also need Syntatic Interoperability (and perhaps natural language processing)





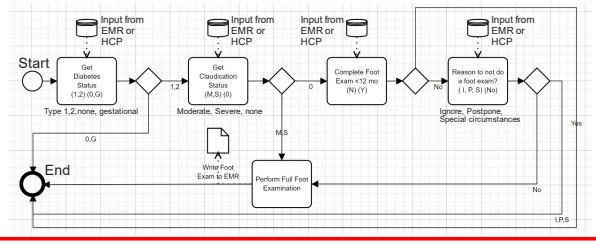
Future of Interoperability

- All EMRs are required to be "interoperable" by 2020[10].
- Generally using FHIR or add-ons like SMART or CDS Hooks[11,12]
- Full interoperability will take much longer
- Smartest 'play' is to use FHIR for its limited interoperability today but plan for full interoperability in the future
- HSPC is working toward data pointers in the cloud

Plan for our Digital Health Information Program

1. Digitize our guidelines using a standardized modeling program (like BPMN)[13]
   a) Standardization allows consistency
   b) Modeling languages, such as BPMN can be tested before coding
   c) Standardized input allows better coding.

BPMN Model of Diabetic Foot Exam

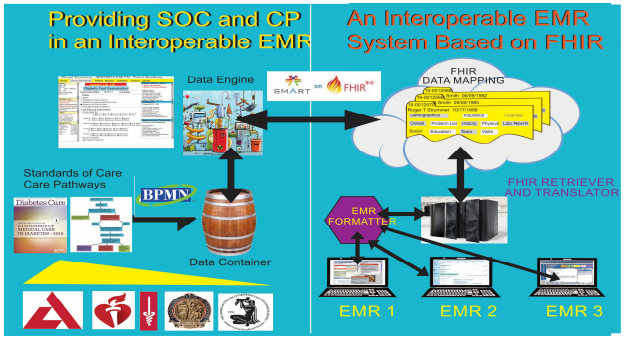


Plan for our Digital Health Information Program

2. Code to integrate with EMR
   a) Write items that need attention
      i. Urgent items especially urgent items: main screen
      ii. Prioritized specialty items show up in problem list frames
   b) Use FHIR to get needed data from EMR
   c) Screens for items needing attention or desired by HCW
      i. Ideally in frame of EMR (Future goal)
      ii. More likely as a new browser window
      iii. GUI formatted to mimic EMR for usability
   d) Write data back to EMR
      i. FHIR where possible
      ii. Keep data in cloud (maintained by DHIP)

Plan for our Digital Health Information Program

3. Connect to EMR using FHIR
   a) Limited connectivity now
      i. In most cases a single hospital/clinic/group
      ii. In some cases, multiple sites using same vendor
   b) Connectivity provided by site or EMR vendor
   c) Improved interoperability will allow greater connection



Progress: Step 1

- Gather Stakeholders. These have expressed interest:
- **Content Providers:**

American Diabetes Assn	American Heart Assn
American Cancer Society	National Kidney Foundation
American College of Surgery	Am Col. Of Obstetr & Gynecol
Am. Col Emergency Physicians	Amer Assn of Diabetes Educators

- **EMR and IT Vendors/groups**

Epic	Cerner
Allscripts	EHRa
Dorsata	Trisotech
HSPC	BPM+

- **Medical Institutions:**

VA	Parkland
University of Iowa	Health Partners

- **We are interested in others in these areas plus**
  - Government (FDA, NIH, CMS)
  - Digital Health Industry
  - Payors
  - Pharmaceutical and Medical Device companies

EHR Data Entry Systems Impact Clinical Data Quality

Jose-Franck Diaz-Garelli, Ph.D., Brian J. Wells, M.D. Ph.D, Roy Strowd, M.D.,
Tamjeed Ahmed, M.D, Rebecca Merrill, Umit Topaloglu, Ph.D.
Wake Forest Baptist Medical Center, Winston Salem, NC

Background

- Learning Healthcare System => Reuse of Clinical Data
- Diagnosis info is used to build patient sets in data warehouses..
- Diagnostic (DX) data are often inaccurate or imprecise
- *In practice*, DX codes often support the retrieval of EHR data for secondary use.

EHR-Added Data Complexity

- EHR system features further contribute to DX variability.
  - Many EHRs provide *multiple textual DX descriptions for each individual DX code* -> Make codes more "Clinician-Friendly"
- Consequences:
  1. More DX descriptions, more potential inaccurate DX selection
  2. Multiple descriptions per DX code, more potential variability
  3. Descriptions may have varying levels of semantic precision
- Example:
  - ICD-10 Code -> C71.1 - **Malignant neoplasm of frontal lobe**
  - **Descriptions Found:**

Anaplastic oligodendroglioma of frontal lobe

Bifrontal or frontal glioblastoma multiforme

Anaplastic astrocytoma of frontal lobe

Oligoastrocytoma of frontal lobe

Methods

Goal: Estimate intrinsic data quality through quantifying the change in variability and the number of inaccurate DX records (*i.e., wrong tumor type or wrong anatomical site*) in oncological EHRs and understand its underlying causes via semi-structured qualitative interviews.

Setup: Comparing EHR-extracted DX data before and after a Biopsy (BX) report in patients with confirmed brain neoplasms (ICD-10 ->C71.*)

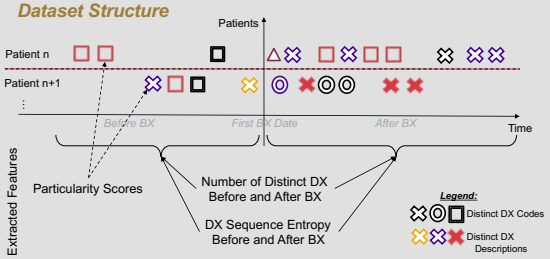
Setup Strengths:

- Limited list of specific DX codes (C71.0-C71.9)
- Definitive histopathology is available
- Annotated dataset available to corroborate secondary data extraction

Methods:

- Statistical Modeling (Binomial & Count Regressions)
- Semi-Structured Physician interviews

Dataset Structure



Results

Table 1 – Descriptive statistics

Measures\Timeframe	Overall	Pre-BX	Post-BX
Distinct Patients	31	19	31
Number of DX Records	1,643	188	1,455
Number of Inaccurate DX Records	212	1	211
Distinct ICD-10 Codes	6	3	6
Distinct DX Descriptions	42	22	40
Distinct Providers	131	41	121
Distinct Hospital Department	33	13	31
Days from Biopsy (Mean±Std.Dev.)	251±312	254±221	316±259

Table 2 – Regressions on DX Variability and Inaccuracies

Model	Term	Estimate (ß)	Ratio (exp(ß))	Ratio Confidence Interval (95%)		p-value
DX Variability	Entropy (DX Sequence)	2.26	9.58	1.53	90.7	0.0259
	Distinct Provider Count	0.203	1.23	1.07	1.45	0.00769
Wrong DX Code	PostBX	2.991	19.904	-0.362	6.577	0.048
	Number of Distinct DX	3.426	30.759	1.139	6.267	<0.0001
	Number of Providers	0.937	2.553	0.297	1.721	<0.0001
	Number of Distinct DX * Number of Providers	-0.197	0.821	-0.375	-0.049	<0.0001

Table 3 – Semi-Structured Interview Findings.
Items represent agreement between both interviewees

Items	Finding
Typical DX Selection Process	Clinicians select the broadest diagnosis possible (e.g., 'non-small cell lung cancer', 'brain neoplasm') because it takes too much mental effort to find a more specific diagnosis.
Interface Issue 1	The cognitive load to search for diagnoses is high because the search function looks different depending the EHR interface module.
Interface Issue 2	Clinicians were frustrated with the diagnosis search feature because it returns a list of hundreds of possible options, which aren't organized in a clinically meaningful way,
Additional Barrier	Time is also a barrier to entering specific diagnosis information due workflow constraints and current lack of data entry support.

Conclusion

- Our results suggest the existence of a link between EHR interaction design and clinical data quality.
- The burden of precise DX selection currently falls on the clinician.
- Further research is needed to uncover the links between clinical practice, EHR interaction design and clinical data quality.

Acknowledgements

The work is partially supported by National Cancer Institute to the Comprehensive Cancer Center of Wake Forest Baptist Medical Center (P30 CA012197) and by the National Institute of General Medical Sciences' Institutional Research and Academic Career Development Award (IRACDA) program (K12-GM102773). The authors acknowledge use of the services and facilities, funded by the National Center for Advancing Translational Sciences (NCATS) (UL1TR001420) as well as the support of the Wake Forest Baptist Comprehensive Cancer Center Qualitative and Patient-Reported Outcomes Shared Resource, supported by the National Cancer Institute's Cancer Center Support Grant award number P30CA012197.

References

1. Farzandipour, M., Sheikhtaheri, A. & Sadoughi, F. Effective factors on accuracy of principal diagnosis coding based on International Classification of Diseases, the 10th revision (ICD-10). *International Journal of Information Management* **30**, 78–84 (2010).
2. Hsia, D. C., Krushat, W. M., Fagan, A. B., Tebbutt, J. A. & Kusserow, R. P. Accuracy of Diagnostic Coding for Medicare Patients under the Prospective-Payment System. *http://dx.doi.org/10.1056/NEJM198802113180604* (2010). doi:10.1056/NEJM198802113180604
3. Wei, W.-Q. *et al.* Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc* **23**, e20–e27 (2016).

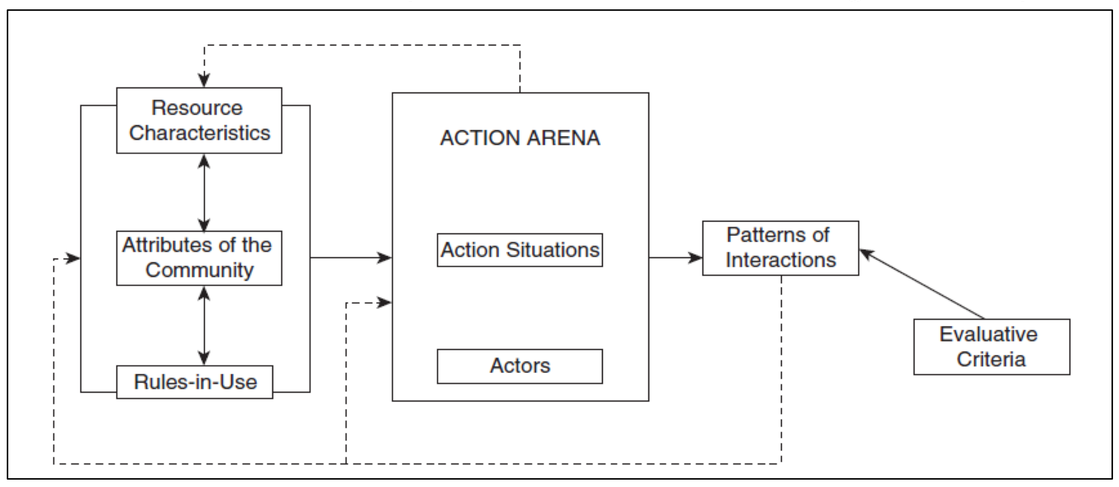
Examining the Theories of "Knowledge Commons" and Applications in Learning Health Systems

[1]Joshua E. Richardson, PhD, MS, MLIS; [2]Jodyn Platt, PhD; [3]John Fox, BSc, PhD; [4]Anthony Solomonides, PhD; [5]Apurva D. Desai, LL.M, M.S., PMP; [6]Philip D. Walker, MLIS, MS; [7]Robin Ann Yurk, MD, MPH; [8]Blackford Middleton, MD, MPH, MSc and The MCBK Policy Trust Workgroup

[1]RTI International, Chicago, IL; [2]University of Michigan, Ann Arbor, MI; [3]Oxford University, Oxford, UK; [4]NorthShore University Health System, Evanston, IL; [5]Department of Veterans' Affairs, Washington, DC; [6]Vanderbilt University, Nashville, TN; [7] MCBK Trust and Policy Workgroup, [8]Apervita, Inc., Chicago, IL

Introduction

"Knowledge commons" is defined as, "the institutional approach (commons) to governing the management or production of a particular type of resource (knowledge)" and has multiple living examples including news wire services, open source software, and genomic data commons.[1] Knowledge commons research includes concepts and tools that are used to address questions around governance, intellectual property, and sustainability. The Governing Knowledge Commons framework (GKC) is a way of identifying the key elements and relationships within knowledge ecosystems, and we are working to understand how its concepts can inform policy, trust, and governance within the computable biomedical knowledge landscape.



Methods

The MCBK Community's Policy and Trust Workgroup (PTWG) is charged with addressing gaps and issues for policies and governance that can impact the trustworthiness of CBK in learning health systems. We are identifying "attributes" and "resource characteristics" and "rules" within the biomedical knowledge ecosystem per Knowledge Commons literature:

1. *Community Attributes: Who are the community members and what are their roles?*
2. *Resource Characteristics: What are the characteristics of the resources?*
3. *Rules in Use: What are the explicit and implicit governance mechanisms? Key policies?*

The PTWG has identified a convenience sample of 31 LHS-related organizations ("community members") and is capturing information about their roles within an LHS: (1) Descriptive information about the organization and knowledge commons (e.g., name, disease domain(s)), (2) Governance and policy, (3) Business model and sustainability, and (4) Culture. For the current analysis, the PTWG is focusing on 10 community members that create public, private, and public-private goods that can be described along the spectrum of Boxwala et al.'s "Four Layers" of knowledge abstraction and specification. The 4 layers include: Layer 1 (L1, narrative), Layer 2 (L2, semi-structured), Layer 3 (L3, structured), and Layer 4 (L4, executable).[2]

Results

The PTWG has sought to establish principles for the identification and classification of knowledge commons, building on existing work and an analysis of case studies. We have identified and collected information about roles, resources, and rules for 30 community members and have closely analyzed 10 (displayed below). Of the 30 members, 15 manage publicly accessible biomedical knowledge, whereas 7 provide access to biomedical knowledge on a paid subscription basis, and 5 operate a mixed public/private model (3 are indeterminate). Seven community members manage knowledge at L4, 3 at L3, 8 at L2, 3 at L1, and 9 whose level of knowledge management is indeterminate.

Commons	Organization	Goods Type	Layer	Governance
KGRID Digital Library	University of Michigan School of Medicine	Public/Private	L3	Applies FAIR principles. Shared responsibility with users; users are responsible for the quality of the metadata.
Platform	Apervita	Private	L4	Complies with client's governance model. EULA and BAA available.
Cochrane Library	Cochrane Collaborative	Public/Private	L2	Complex editorial and management structure across different elements of the organization.
EBMG & EBMEDS	Duodecim	Private	L2	Duodecim extracts and governs the content from its publications.
Repertoire	OpenClinical	Public/Private	L4	CAFE principles: Community, Accessibility, Fairness, Empowerment
UpToDate	Wolters Kluwer	Private	L1	Team of CMOs for peer review with team of authors and editorial board.
Trust™ Guideline Repository	ECRI	Public	L1	TRUST and NEATS standards / IOM (2011) Guidelines
VA/DoD CPG	VA/DoD Guidelines	Public	L2	Directed by VA/DoD Health Executive Committee subject to public charter.
Clinical Knowledge Manager	OpenEHR	Public	L3	Governance instituted for domains and subdomains, but work in progress is not governed. Maintains data on users and contributors.
IMO 2.0 ETP	IMO Intelligent Medical Objects	Private	L4	Tight control of the core vocabulary to maintain quality. Subscription and collaboration models. When deployed, complies with client governance model.

Discussion and Conclusion

We are finding that the GKC Framework is useful for identifying community member attributes and the characteristics of the computable biomedical knowledge that the members manage.The GKC is also a helpful guide for developing a convenience sample of community members, for which we intend to develop detailed case studies. Research questions for case studies that we are considering include: 1) Do members' access to knowledge artifacts change as evidence gets translated from L1 to higher levels?; and 2) How might any knowledge commons policies need to address different layers of knowledge?

References

1. Strandburg KJ, Frischmann BM, Madison MJ, editors. Governing Medical Knowledge Commons [Internet]. 1st ed. Cambridge University Press; 2017 [cited 2019 Jul 15]. Available from: https://www.cambridge.org/core/product/identifier/9781316544587/type/book
2. Boxwala AA, Rocha BH, Maviglia S, Kashyap V, Meltzer S, Kim J, et al. A multi-layered framework for disseminating knowledge for computer-based decision support. Journal of the American Medical Informatics Association. 2011 Dec;18 Suppl 1:i132-9.

Automated Near Real-time Detection of the Overuse of Dental Diagnostic X-rays via Compliance with Best Practices Guidelines

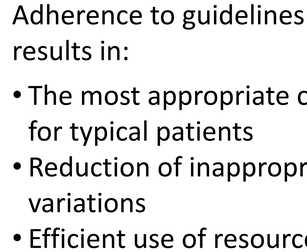
Todd Pollack, DMD, Elizabeth Runfola, Yoemy Waller, MS, MBA, and Franklin Din, DMD, MA

Healthcare Costs - Out of Control

The cost of healthcare is skyrocketing. A huge part of the cost is due to losses to Fraud, Waste, and Abuse (FWA).

Scope of problem: 100,000 providers x $100 FWA per week x 50 weeks per year =

$500 million per year loss



Current approaches to dealing with FWA are unable to deal with small dollar, high frequency, high prevalence abusive billing practices because of the manual nature of current FWA detection.

Use of computable guidelines for anti-Fraud, Waste, and Abuse, improved small dollaG FWA detection by 5%

Guidelines to Rationalize and Standardize

Interest in clinical guidelines has its origin in variations in medical service delivery and outcomes. Variations have resulted in inappropriate care, overuse or underuse of services, suboptimal outcomes. This is contrary to the fundamental desire of providers to offer, and of patients to receive, the best care possible.

Clinical practice guidelines are "statements that include recommendations, intended to optimize patient care, that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options" (Consensus Report, IOM, Mar 2011). The review results in a set of recommendations.

Compliance vs. non-Compliance

Adherence to guidelines results in:

- The most appropriate care for typical patients
- Reduction of inappropriate variations
- Efficient use of resources
- Foundation for clinical quality assessment

Non-Adherence to guidelines results in:

- Wide variations in care delivery and outcomes
- Patient harm
- Excessive treatments and costs
- The ability to measure clinicians against a standard

ADA Full Mouth X-ray Guideline

This is a snippet from the guideline. This is not computable as is.

Table 1.

TYPE OF ENCOUNTER	PATIENT AGE AND DENTAL DEVELOPMENT/			
	Child with Primary Dentition (prior to eruption of first permanent tooth)	**Child with Transitional Dentition** (after eruption of first permanent tooth)	**Adolescent with Permanent Dentition** (prior to eruption of third molars)	**Adult Parti**
New Patient* being evaluated for oral diseases	Individualized radiographic exam consisting of selected periapical/occlusal views and/or posterior bitewings if	Individualized radiographic exam	Individualized radiographic ex posterior bitewings with panor	

For more on guideline use and better program integrity and FWA detection

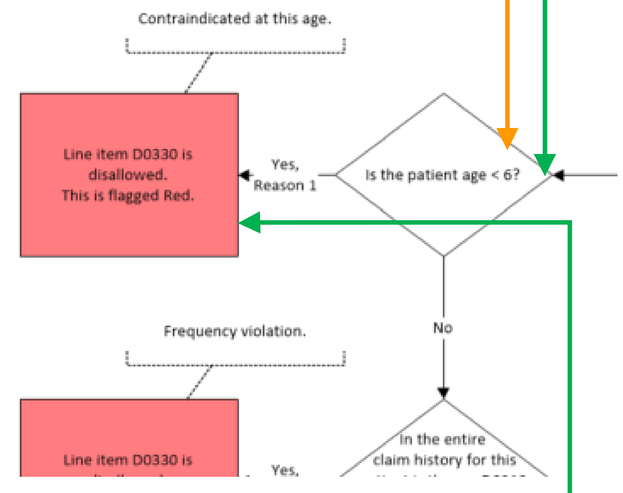
Deconstruct the Guideline into Computable Steps

Deconstruct the narrative into a Step-by-Step

Step	Guideline	Computable Deconstruction
1	Child with Primary teeth	Age < 6 years old
2	Child with Transitional Dentition	Age >= 6 years old and
2	Child with Transitional Dentition	Age <= 14 years old

Reconstruct and Build into a Model

Then, reconstruct and model from the Step-by-Step



Run and Validate Against Manual Review

Dentist	Pat ID	date of service	proc code	paid money	age	reason
Dr1	pat1	12/8/2014	D0330	$130.50	3	Patient is < 6 years old

Data Element Commonality
across
Clinical Trial, Observational and Electronic Health Records
that describe HIV infected Patients in the USA

Nick Williams, Ph.D & Kin Wah Fung, MD

Lister Hill National Center for Biomedical Communications
National Library of Medicine at the National Institutes of Health
With ICF International

Introduction

Several data collection efforts document clinical experiences of HIV infected patients but the data integration potential of these efforts remains under described.

Methods

We collected a ***convenience sample*** of HIV clinical trial, observational study and EHR patient level records from various sources. We classified data tables as containing primarily one of ten possible data element classes. We then counted the number of **Data Elements(DE)** and **Item Responses(IR)** present in the classified record tables and produce two metrics of comparison across data sources: (1)data element class commonality and (2)item response volume by data element classes across data collection effort types.

RESULTS

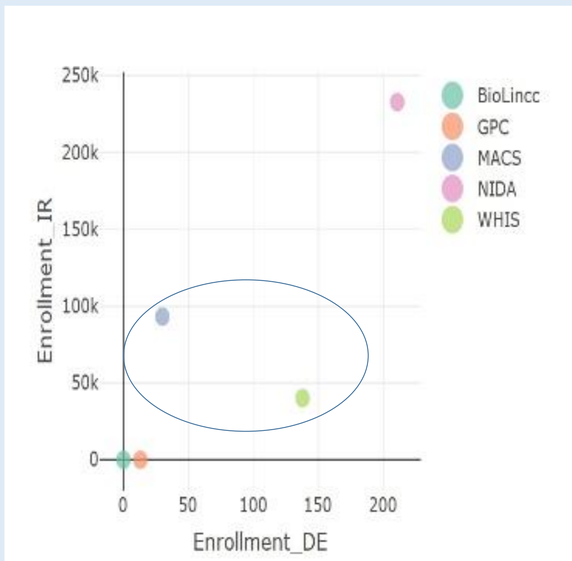
We evaluated five data sources with a total of **27,372** data elements and **204,976,009** item responses. Distributions of data by type are not evenly distributed across DE types. Common, usually well structured DE include DX, RX and Labs (RED); observational studies have similar separation from non-observational HIV studies for Survey and Enrollment (BLUE).

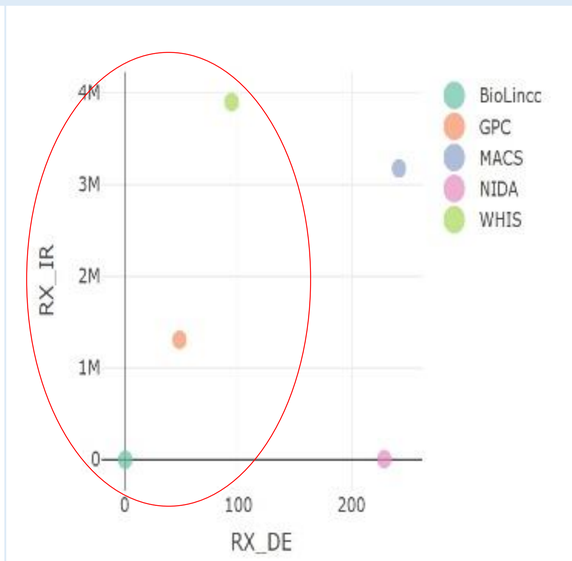
DISCUSSION

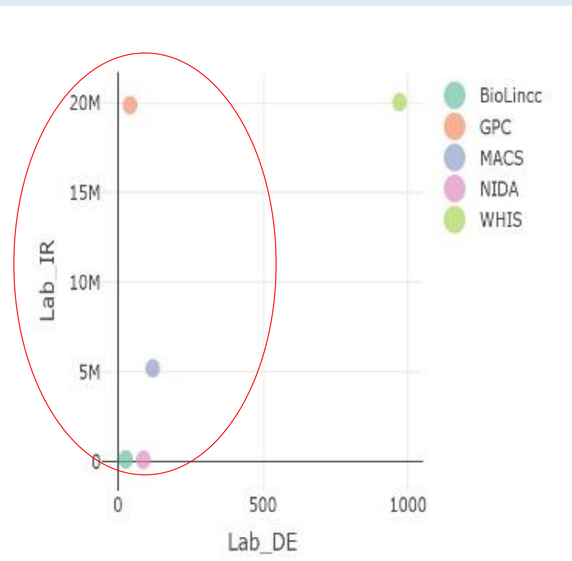
Different data source types have different data element and item response density.

This is most likely due to the underlying:

a. Case report structure(s)
b. Case natural history (Kinship or not)
c. Repeat Measure vs. 'ask every encounter'
d. Retrospective medical record extraction presents unique data element challenges; DX or Survey?

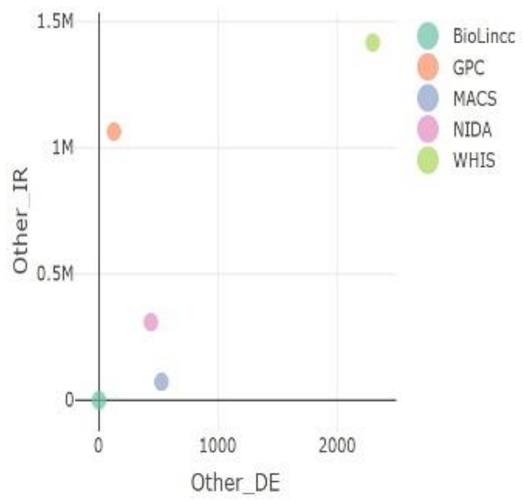


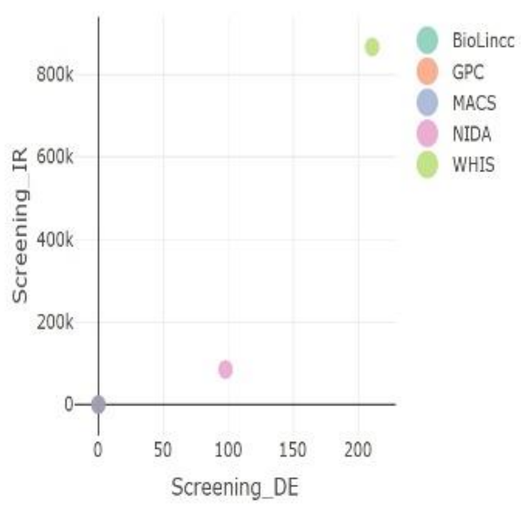


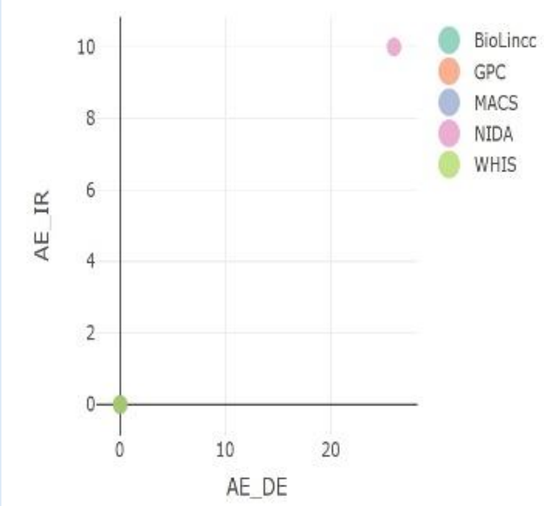


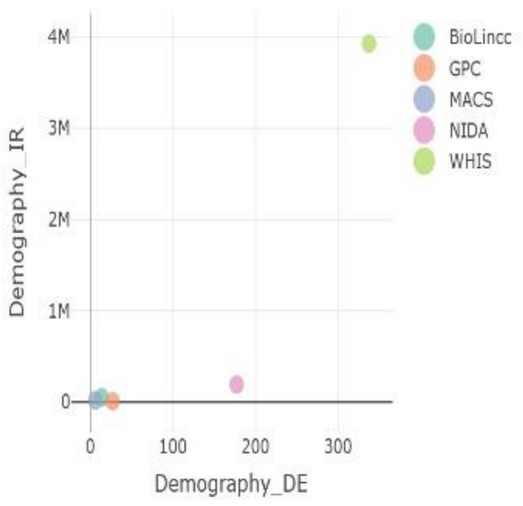
Different kinds of HIV patient level records have different clinical information.

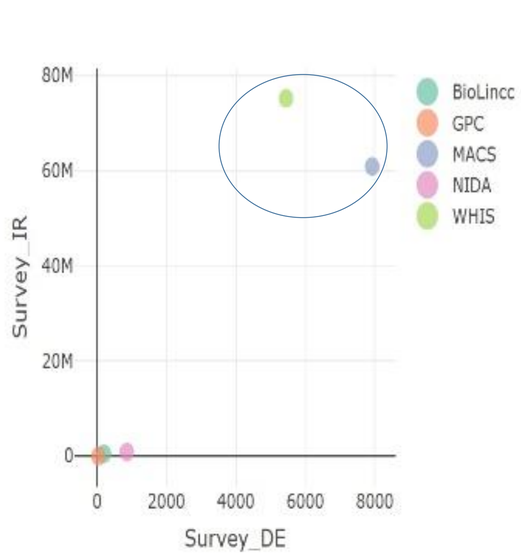
Data Reuse Studies *beware*!

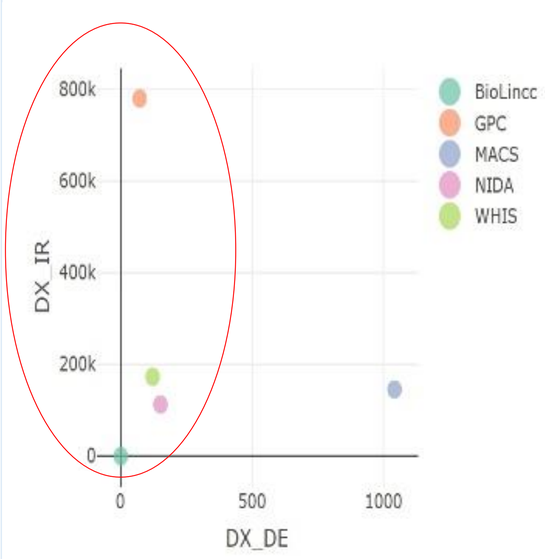












Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC) and by NIH Office of AIDS Research. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services.

Utilizing electronic dental record data to monitor periodontal disease progression

J. S. Patel[1, 2], A. Zai[1], T. Thyvalikakath[1, 2, 3]

[1]Dental Informatics Core, Indiana University School of Dentistry, [2]Indiana University School of Informatics & Computing, [3] Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN

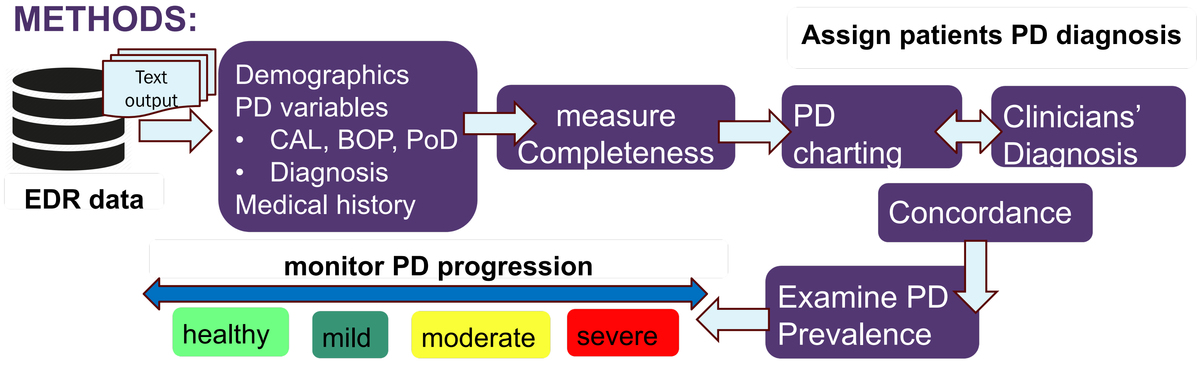
BACKGROUND:

- Periodontal disease (PD) remains a major cause of tooth loss and impairs quality of life.
- Due to wide use of electronic dental record (EDR) data for research during last decade, this data could be utilized to study PD, its risk factors, and disease progression.
- However, many challenges exist with using EDR data such as quality of data, and missing data. Spurious data generate spurious results.

OBJECTIVE:

- Develop measures to evaluate the quality of EDR data that indicate PD findings and diagnosis.
- Examine if longitudinal EDR could help in monitoring patients' PD progression.

METHODS:



INITIAL RESULTS:

Completeness of selected periodontal disease clinical variables in the EDR (total number of patients=25,317)

Clinical Variables	Number of patients (Unique) (%)	Numbers of patients: more than one visit info present (%)
Total Number of Patients	25,317 (100)	N/A
Date of Birth	25,317 (100)	N/A
Gender	25,243 (99.7)	N/A
Race	17,872 (71)	N/A
Bleeding on Probing	16,923 (67)	143 (0.6)
Clinical Attachment Loss	20,045 (80)	5 (0.0)
Pocket Depth	20,024 (79)	4 (0.0)
Clinicians' Assigned PD Diagnosis	11,130 (44)	5,732 (23)

Age distribution

Range	Numbers (%)
18-27	4,675 (18)
28-37	4,521 (18)
38-47	4,298 (17)
48-57	5,267 (21)
58-67	3,831 (15)
68-77	19,48 (8)
78-87	694 (3)
88 and older	83 (0.3)
Total	25,317 (100)

Gender distribution

Category	Numbers (%)
Female	13,632 (54)
Male	11,603 (46)
Unknown	74 (0.3)
Transgender	8 (0.0)
Total	25,317 (100)

Race distribution

Race	Numbers (%)
Caucasian	12,237 (48)
African American	3,311 (13)
Hispanic	1,630 (6)
Asian	501 (2)
Other	139 (0.5)
Multiracial	39 (0.2)
American Indian	13 (0.1)
Pacific Islander	2 (0.0)
Missing	7,418 (29)
Unknown	27 (0.1)
Total	25,317 (100)

CONTACT: jsp2@iupui.edu, tpt@iu.edu

Inferring patient instructions (signatura) from pharmacy dispensation data

Vojtech Huser, MD PhD, Nick Williams, PhD

Lister Hill National Center for Biomedical Communications, National Library of Medicine, NIH
vojtech.huser@nih.gov

Introduction

Some medication-focused decision support logics may require structured representation of patient instructions (signatura; or 'sig') specified in a prescription (e.g., "take every 8 hours"). As part of a larger project around common data elements in HIV patients, we investigated methods of inferring signatura from 2016 Medicare pharmacy dispensation claims of known HIV+ Medicare Part-D Patients.

Methods

First, in aggregate population analysis, we analyzed QPD ratio (quantity per day; quantity divided by days of supply [DOS]) for drugs taken by HIV patients (206,661 patients; 12,108 NDCs) who had 12 months of Part-D enrollment in 2016. In the Virtual Research Data Center (VRDC) from Center for Medicare and Medicaid Services (CMS), the following two drug dispensation data fields are available and defined in documentation [1] as:

(1) DAYS_SUPLY_NUM: this field indicates the number of days' supply of medication dispensed by the pharmacy and consists of the amount the pharmacy enters for the prescription.

(2) QTY_DSPNSD_NUM: This field indicates the number of units, grams, milliliters, or other quantity dispensed in the current drug event. If the Part-D event was for a compounded item, the quantity dispensed is the total of all ingredients. If the Part-D event was for a partial fill, the quantity is the total amount prescribed, not the portion covered by the partial fill. *Comments*: The values for this field are highly variable and depend on the form of the drug that was dispensed (e.g., liquids or tablets).

Second, for some drugs, we compared inferred signatura with separate HIV cohort for which we had Electronic Health Record data (in PCORNet model) obtained via the Great Plains Collaborative (utilizing PRN_FLAG column in ORDERING table). Third, to evaluate data quality of days of supply field (assuming medication compliance and no stockpiling), we analyzed dispensation pattern over time: D(n+1) dispensation followed by approximately DOS(n) days after D(n).

Results

Exhibit 1 shows for selected medications (on NDC level) the QPD ratio(QDR). For full list, see project repository (file S01_NDCs-and-quantity-per-day-ratio.csv). Related prior work was conducted by Bodenreider et al. [2]. Exhibit 2 shows Select medication dispensation events(Y) by Days Of Supply(X) for HIV, asthma, lipids, seizure and anti-biotic drugs common among the Medicare Part-D HIV+ population in 2016.

Conclusion

We concluded that inferring signatura is feasible for a subset of drugs and we piloted machine reasoning logic flagging irregular dispensation data.

The project repository contains additional results:
https://github.com/lhncbc/CDE/tree/master/hiv/ehr/signatura

References

1. ResDAC, Part D Drug Event file documentation (for Medicare data). Available at https://www.resdac.org/cms-data/files/pde/data-documentation [Accessed: May 24,2018]
2. Bodenreider O, Rodriguez LM. Analyzing U.S. prescription lists with RxNorm and the ATC/DDD Index. AMIA Annu Symp Proc. 2014;2014:297–306. Published 2014 Nov 14. PMID: 25954332

Acknowledgements:
This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC) and by NIH Office of AIDS Research. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services.

Exhibit One:
25 select National Drug Code Drug names with QDP Ratio (QDR), Population Days of Supply Average (POP DOS) and Population Quantity Dispensed Average (POP QUANT MEAN)

qdr	pop dos mean	pop quant mean	concept name
1	32.33	31.96	emtricitabine 200 MG / Tenofovir disoproxil fumarate 300 MG Oral Tablet [Truvada]
1	33.49	33.48	efavirenz 600 MG / emtricitabine 200 MG / Tenofovir disoproxil fumarate 300 MG Oral Tablet [Atripla]
1	31.46	31.59	darunavir 800 MG Oral Tablet [Prezista]
1	31.76	31.77	abacavir 600 MG / dolutegravir 50 MG / Lamivudine 300 MG Oral Tablet [Triumeq]
1	32.04	32.05	cobicistat 150 MG / elvitegravir 150 MG / emtricitabine 200 MG / tenofovir alafenamide 10 MG Oral Tablet [Genvoya]
1	32.53	32.54	abacavir 600 MG / Lamivudine 300 MG Oral Tablet [Epzicom]
1	21.17	25.06	Sulfamethoxazole 800 MG / Trimethoprim 160 MG Oral Tablet
1	32.46	32.45	emtricitabine 200 MG / tenofovir alafenamide 25 MG Oral Tablet [Descovy]
1	31.73	31.74	cobicistat 150 MG / elvitegravir 150 MG / emtricitabine 200 MG / Tenofovir disoproxil fumarate 300 MG Oral Tablet [Stribild]
1	31.89	31.94	cobicistat 150 MG / darunavir 800 MG Oral Tablet [Prezcobix]
1	32.09	32.09	emtricitabine 200 MG / Rilpivirine 25 MG / Tenofovir disoproxil fumarate 300 MG Oral Tablet [Complera]
1	32.38	32.53	Zolpidem tartrate 10 MG Oral Tablet
1	34.54	34.53	efavirenz 600 MG Oral Tablet [Sustiva]
1	45.41	44.97	Rosuvastatin calcium 10 MG Oral Tablet [Crestor]
1	32.32	32.35	emtricitabine 200 MG / Rilpivirine 25 MG / tenofovir alafenamide 25 MG Oral Tablet [Odefsey]
1	8.74	8.59	Levofloxacin 500 MG Oral Tablet
1	45.06	45.19	Hydrochlorothiazide 25 MG Oral Tablet
1	49.80	51.27	Lisinopril 5 MG Oral Tablet
1	56.00	56.37	Amlodipine 10 MG Oral Tablet
1	41.44	41.37	atorvastatin 20 MG Oral Tablet
1	45.80	45.82	atorvastatin 20 MG Oral Tablet
1	46.43	46.11	Rosuvastatin calcium 20 MG Oral Tablet [Crestor]
1	8.71	8.56	Levofloxacin 500 MG Oral Tablet
1	49.46	49.50	Hydrochlorothiazide 25 MG Oral Tablet
1	47.08	47.01	atorvastatin 40 MG Oral Tablet

Exhibit Two: Dispensations (Y) by Days of Supply(X) for select Medications

The Legal Interoperability of Biomedical Research Data:
Promoting Open Science by Defining and Describing Data Licensing Policies, Data Sharing and Use Agreements, and other Intellectual Property Frontiers in Select Data NIH Data Repositories

Paije Wilson, MLIS. Associate Fellow, NLM
Rebecca Goodwin, JD. Policy Analyst and Data Science and Open Science Specialist, NLM
Dina Paltoo, PhD, MPH. Assistant Director for Policy Development, NLM

NIH NLM

Abstract

With the emergence of Open Science initiatives there is increased focus on the potential to reuse scientific data, especially with regards to combining multiple datasets to facilitate new discoveries. However, legal restrictions or ambiguity relating to data reuse have remained persistent obstacles for researchers. For this study, the investigators wanted to determine what data reuse-related licensing and other intellectual property characteristics are associated with the data repositories on the list generated by the Trans-NIH BioMedical Informatics Coordinating Committee (BMIC).

Background

What is a Data License?

Definitions for "Data Licenses" are frequently ambiguous, and often differ between repositories. For the purpose of this study, a **Data License** will be defined as a legal contract that enables the use of legally protected data. Data licenses typically impose conditions and restrictions which govern the use of data, and are frequently found within data use agreements.

What are the Challenges in Data Licensing?

Though data licenses may serve as an extra layer of protection for data creators, they can present significant challenges for data reuse. Challenges tend to stem from complexity of data licenses, poor discoverability of licensing terms, and incompatible licensing provisions. For researchers attempting to reuse data, these issues create undue confusion and even, in many cases, serve as deterrents for reusing data, due to fears of legal repercussions should researchers misinterpret or fail to identify licensing terms.

Methods

The project was conducted in 3 parts:

1. A literature review was conducted to determine the landscape of issues facing legal data reuse.

2. 9 individuals with expertise in data licensing agreements, reuse of scientific data, or the NIH BMIC data repositories were interviewed on issues facing the legal reuse of data. Interviewees included individuals from the:

   - National Library of Medicine
   - (Re)useable Data Project
   - NIH Office of Extramural Research
   - NIH Office of Science Policy
   - Research Data Alliance
   - NIH Office of Technology Transfer

3. 20 of the 80 NIH data repositories from the BMIC* list were examined and compared for their data licensing, use agreements, and other stipulations relating to data reuse. The first 10 repositories were chosen on the basis of their domains (i.e. if a repository contained data from a domain that was not yet represented in the sample, the repository would be included into the analysis). Once unique domains were no longer found, the investigators analyzed repositories in the order in which they appeared on the BMIC list. The large number of genetics repositories in the sample reflects the disproportionate representation of the genetics domain in the BMIC list; this may be due in part to the longstanding culture of sharing in the genetics research field. A visualization of the repository domains represented in the sample may be seen in **Figure 1**.

***Note: The Trans-NIH BioMedical Informatics Coordinating Committee (BMIC) maintains a list of over 80 NIH-supported data repositories, along with their data sharing policies, data submission, and access processes.**

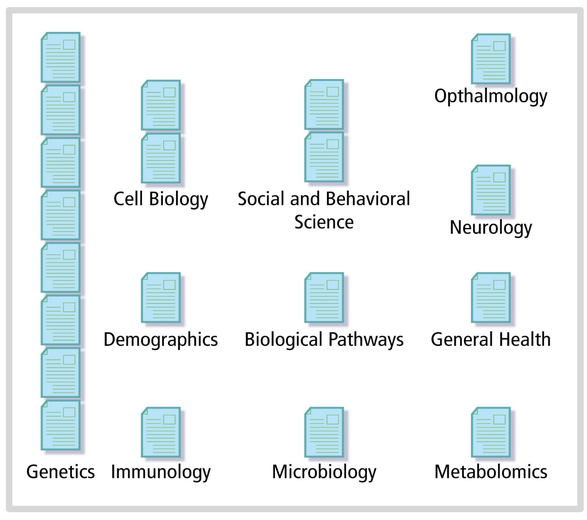


Figure 1 shows the scholarly domains that were represented in the sample. Genetics comprised 8 of the repositories; cell biology 2, social and behavioral science 2, and the remaining domains had 1 repository each within the sample. Visualization complements of Donny Bliss.

Results: Issues Facing Data Reuse

"Licenses on top of licenses"

- A repository may have a single, overarching license for all of its datasets, but the datasets, themselves, may have their own individual licenses with terms that may conflict with the overarching repository agreement
- Merging two datasets can result in convergence of two incompatible licenses.

Poor discoverability of licensing terms

- License terms may be scattered across multiple locations, have nonsensical or unstandardized page locations, or may even be missing.

Variability between and within license types

- No federal law explicitly defines data ownership. This causes confusion for data creators over what constitutes data ownership, and which parts of their data may be legally protectable.
- Licenses that are under the same standardized name may have different terms and conditions associated with them.
- Customized licenses, which are becoming increasingly prevalent, typically lack standardization and machine readability.
- A single repository may have different licenses for different types of datasets, with no justification or standardization for which characteristics warrant additional legal restrictions.
- Data creators and users frequently lack expertise in interpreting and adhering to data licenses.

Table 1 (above) shows a cumulative summary of the issues facing data reuse identified in the literature review and the interviews

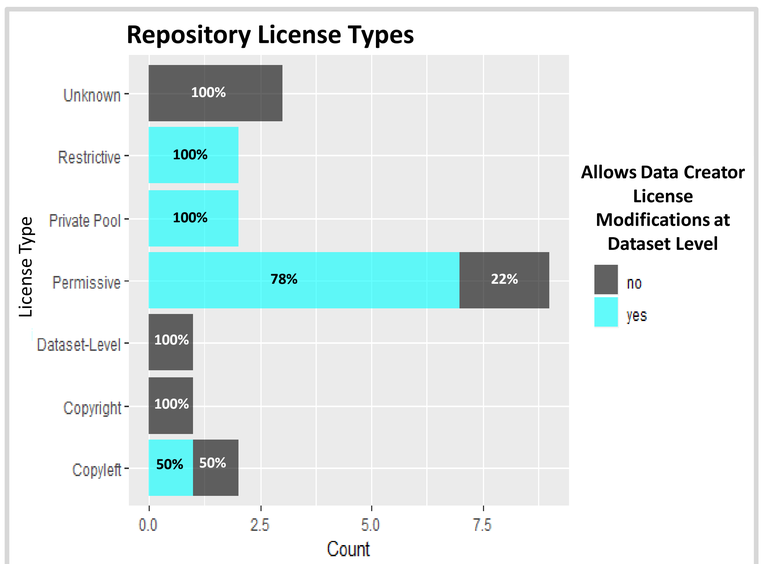


Figure 2 shows the license types that each of the sampled repositories contained, and the percentage of each license type that allowed creators to make modifications to the licenses at the dataset-level. License types were directly based on the study conducted by Carbon et al. (see QR code).

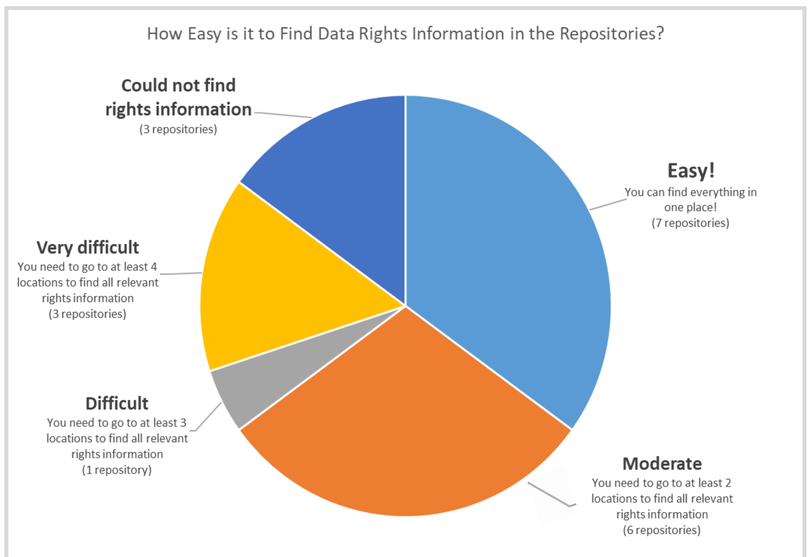


Figure 3 shows the minimum number of locations one would need to look in order to find all relevant licensing terms for the sampled repositories

Conclusions

Discussion

Work has been done to increase the legal interoperability of research data. The U.S. National Institutes of Health and the National Library of Medicine have both released strategic plans that call for making data **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (**FAIR**), with the characteristic of "reusability" requiring that all data must provide clear information on associated data licenses. The Research Data Alliance also released a set of guidelines in 2016 for the legal interoperability of research data, emphasizing the need to facilitate the lawful access and reuse of data.

However, as demonstrated by this study, implementation has flushed out unforeseen barriers to improving the legal interoperability of research data. Some of the most prevalent problems include overlapping and incompatible licenses ("licenses on top of licenses), poor discoverability of licensing terms, and variability between and within license types.

Recommendations

- **Provide use cases for the legal uses of datasets**
- **Standardize locations of data rights information**
- **Make data and data agreements FAIR; create data licensing workshops and web tutorials for data creators and submitters; and standardize agreement names and components with designated areas for customization**

This research was supported in part by an appointment to the NLM Associate Fellowship Program. This program is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine. Sources are available via the QR code provided on the right.

View PDF

Leveraging Library Expertise to Make Computable Biomedical Knowledge FAIR

[1]Conte, Marisa; Samuel, Sara; [2]Flynn, Allen; Boisvert, Peter; Stuchell, Lance; Raths, Brooke; Allan, Jack
[1]meese@umich.edu, [2]ajflynn@umich.edu

LIBRARY

Expertise in describing, organizing, disseminating, and preserving information

KGrid

Developers of technical infrastructure for managing and deploying CBK

A FAIR agenda requires deep and abiding multi-stakeholder collaboration

F — Findable → Metadata

- ✓ Persistent globally unique identifiers
- ✓ Linked data methods

UNIVERSITY OF CALIFORNIA EZID *Identifiers made easy*

JSON-LD

A — Accessible → Publishing

- ✓ Publish CBK as a primary source
- ✓ Manage repositories
- ✓ Online access

N2T RESOLVER NAMES → THINGS

I — Interoperable → Packaging

- ✓ Ingest CBK into digital libraries
- ✓ Archive CBK effectively
- ✓ Disseminate CBK in standard packages

OAIS

Open Archival Information System

R — Reusable → Policies

- ✓ Community-guided collection policies
- ✓ Clear licenses for CBK users
- ✓ Policies on CBK provenance & auditing

CC

FOR ALL — Open Science → Commitments

- ✓ Reproducible research
- ✓ Public access
- ✓ Transparent governance

CENTER FOR OPEN SCIENCE

Mission Statement

MCBK Mobilizing Computable Biomedical Knowledge

MCBK Manifesto

Related Resources:

Knowledge Grid *https://kgrid.org/*
Taubman Health Sciences Library https://www.lib.umich.edu/thl
EZID https://ezid.cdlib.org/
Name-To-Thing https://n2t.net/
Open Archival Information System https://www.oclc.org/research/publications/library/2000/lavoie-oais.html
The FAIR Data Principles https://www.force11.org/group/fairgroup/fairprinciples