Veterans Health Administration | Office of Health Informatics
Knowledge Based Systems | Standards and Interoperability – Informatics Architecture

Task Order #9 | January 29, 2019

SOLOR Support Services: Use Case #1 Manuscript Version

PRESENTED BY:

Team BZ
Program Manager: Jayme Welty

bookzurman
Transforming healthcare through technology + trust

360 Central Ave., Suite 970 | St. Petersburg, FL 33701
O 727.378.9006 | bookzurman.com

PRESENTED TO:

Dr. Keith Campbell & Stephanie Klepacki

IDIQ: VA701-16-D-0017
TO9 PO: 776-C80148
CLIN: 2005B Proposed Standards Artifacts (Medium)

A FORMATIVE EVALUATION OF THREE SOLOR EXTENSION USE CASES PROMOTING SEMANTIC INTEROPERABILTY

(CLIN 2005B_10.14, 2005B_11.14 and 2005B_12.14)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

VERSION HISTORY

DATE	VERSION	DELIVERABLE	DESCRIPTION
10/29/2018	0_0	Proposed Standards Artifacts (Medium)	Use Case
11/29/2018	0_1	Proposed Standards Artifacts (Medium)	Evaluation Plan
12/29/2018	0_2	Proposed Standards Artifacts (Medium)	Evaluation
1/29/2019	0_3	Proposed Standards Artifacts (Medium)	Manuscript Version

1. INTRODUCTION

The vision of the Department of Veterans Affairs (VA), Veterans Health Administration (VHA), Office of Informatics & Analytics (OIA), and Health Informatics (HI) is to provide timely, relevant information and data services that support improvements in Veterans' health. In meeting these goals, OIA strives to provide high quality, effective, and efficient information and data services to those responsible for providing care to the Veterans at the point-of-care as well as throughout all the points of the Veterans' health care in an effective, timely and compassionate manner. VA depends on the interoperability of information and data to meet mission goals. An essential step to achieving interoperability is the widespread adoption of clinical terminology standards, which are structured sets of codes and terms organized in hierarchies to represent and encode clinical concepts – including diagnoses, procedures, medications, administrative data, and laboratory results [1]. By 2020, adoption of certified Electronic Health Record (EHR) systems in the U.S. with data mapped to standard clinical terminologies is expected to approach 100%, due in part to EHR incentive programs created by the Health Information Technology for Economic and Clinical Health Act and Meaningful Use program [2].

Despite having such mandates, there are challenges in the application of controlled medical terminologies to clinical care that limit our ability to fully leverage EHR data to improve population health. Standard clinical terminologies are currently developed and delivered by various organizations in which the content is often created in silos, stored in different formats, represented by different models, and released with different cycles and mechanisms. As a result, end users of EHRs are taxed heavily to monitor, retrieve, implement, and analyze the ramifications of an update. This burden is compounded even further when a new set of content is required for use.

To this end, VHA's informatics architecture was created to integrate disparate knowledge sources and preserve the meaning of information for the interoperability of electronic health record data (i.e., semantic interoperability) which is critical for delivering safe veteran care and leveraging standards-based clinical decision support. The current complexity encountered by standards developers, authors, and implementers when trying to integrate disparate terminologies—and the lack of coherence between (and sometimes within) the terminologies themselves—must be overcome to build a foundation for scalable and extensible architecture. Solor, a system of logical representation, is the open source ecosystem of capabilities and services for overcoming these complexities by assimilating disparate health knowledge sources into a consistent representation based on best practices of computer science. Knowledge sources are integrated in Solor by transforming source terminologies into a common model that provides a uniform representation scheme and additional meta-data needed for semantic integration and advanced versioning. Users are able to navigate overlapping concepts, as well as the relationships between concepts. By doing this, Solor enables collaboration in health IT, unifies health terminology standards and removes ambiguity, leading to improved patient care.

1.1. Aims

The overarching objective of this body of work is to inform the development of Solor by exploring its extension as an ecosystem for integrating disparate knowledge sources and creating interoperability by making information meaningful and computable. The specific aims of this work are:

Aim 1: Develop use cases for the extension of Solor.

Aim 2: Evaluate constructs of the Solor use cases developed in previous aim.

2. BACKGROUND

To be completed as part of future deliverable.

2.1. The Solor System

To be completed as part of future deliverable.

2.2. Solor Knowledge Sources

2.2.1. Terminology Knowledge Sources

Terminology systems are increasingly critical components for achieving interoperability across applications in the healthcare domain. The role of standard terminologies in achieving interoperability for the purposes of advancing patient care is well documented [3]. Ideally, these clinical terminology standards intend to provide rules to allow for the exchange, integration, and management of electronic clinical information [4]. The federal government recognizes the benefit of standard terminologies and promotes their development and use. The *Federal Health IT Strategic Plan 2015-2020* set a strategy to encourage consistent terminology standards implementation in Electronic Health Records (EHR) and encourage use through federal payment policies [5]. A standard terminology is one that has wide industry acceptance or use.

Standards are obtained from a variety of efforts, cover different domains of clinical and nonclinical content relevant to the EHR, and serve various purposes. Currently, no one terminology or classification system contains everything that is needed for the medical record. Examples of standard terminologies include:

- Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT®): a comprehensive clinical terminology, maintained by the International Health Terminology Standards Development Organization (IHTSDO) [6] representing over 300,000 concepts including disorders (22%), procedures (17%), body structures (11%), clinical findings other than disorders (10%), and organisms (10%) [7];
- Logical Observation Identifiers, Names, and Codes (LOINC®): a terminology representing about 50,000 clinical and laboratory observations, health measurements, and documents, developed and maintained by the Regenstrief Institute [8]; and
- RxNORM: a terminology for human clinical drugs in the U.S representing drug properties such as ingredient, strength, and dose form, maintained by the National Library of Medicine (NLM) and distributed via the Unified Medical Language System (UMLS) [9].

Terminology systems typically consist of the following elements:

- Coded Concepts – the discrete units of knowledge managed within the terminology. They typically consist of numeric codes and textual preferred names, synonyms, and descriptions.
- Concept Hierarchies – the logical organization of concepts into parent-child and ancestor-descendant relationships that express the semantics of generalization and specialization. The hierarchical organization of a terminology may be explicitly expressed through stored parent-child and ancestor-descendant links, or it may be implicitly expressed through the logical definitions of individual concepts that a computer can use to infer parent-child and ancestor-descendant relationships.
- Value Sets – named lists of individual concepts that represent more abstract categories useful in decision-support logic.

New applications and new medical knowledge constantly call for expansion and enhancement of existing terminologies. However, since terminology systems are often non-static, incomplete and under specified, inconsistencies may be introduced [10].

While many of these challenges are related to terminology evolution, others may be related to the design of the standard clinical terminologies themselves. Cimino notably described the challenges of concept orientation, completeness, correctness, currency, granularity, and redundancy when designing re-usable medical terminologies [11]. Today, 20 years later, a menagerie of inconsistent and overlapping terminology models hinders efforts that try to store and analyze encoded clinical data. Several efforts aim to assist. The National Library of Medicine (NLM) integrates terms and codes from over 150 source vocabularies by concept, attribute, and meaning in the Unified Medical Language System ® (UMLS) Metathesaurus. The NLM, also, in collaboration with the Office of the National Coordinator for Health Information Technology and Centers for Medicare & Medicaid Services, hosts the Value Set Authority Center (VSAC). The VSAC aims to provide lists of values, codes, and names (i.e., value sets) from standard clinical terminologies to represent clinical concepts.

These tools, while helpful, have gaps. Raje et al. highlighted issues with completeness, correctness, and redundancy when they found gaps in the UMLS Metathesaurus' coverage of disease concepts [12]. Similarly, Winnenburg et al. highlighted duplicate value sets in the VSAC, and showed that 19% of value sets in 2011 contained invalid codes [13]. In subsequent work, they highlighted issues related to granularity by evaluating over 1,000 value sets and found that value sets varied vastly in size with some only containing one code, while other value sets included over 20,000 codes (ref). Similarly, Bahr et al. showed issues with concept orientation by analyzing medication value sets and found extraneous and missing ingredients in both the value sets and drug classes [14].

These issues related to integrating clinical content have a direct impact on patient safety and point to the need to be able to consistently represent and encode clinical data and observations. Therefore, quality assurance is an indispensable part the terminology management lifecycle. A central limitation of integrating controlled medical terminologies is that they often lack any formal model to denote the relationships among constituent data elements.

Recently, however, development teams for SNOMED CT, LOINC, and RxNorm have partnered to promote interoperability. Developers can now leverage SNOMED CT's representation model for the building blocks of LOINC, and a new drug model in SNOMED CT facilitates extensions and consistency to RxNorm [7]. Bodenreider et al. wrote about the recent collaboration: "while this evolution leads to greater compatibility and interoperability, integration of SNOMED CT, LOINC, and RxNorm still requires mappings among the three terminologies. Moreover, these three terminologies use different formalisms and tools for their representation, have their own release cycles and versioning mechanisms, which makes their seamless integration non trivial, if at all possible." [7].

2.2.2. Genome Variant Knowledge Sources

A key part of the work in the genome research domain is to identify genome variants and assign a clinical impact, if known. A genome variant knowledge source is a repository of known genome variants and associated clinical interpretations of that variant. There are many types of genome variant knowledge sources, which include (1) privately-controlled knowledge bases, such as the Human Gene Mutation

Database (HGMD) [15]; (2) open access, locus-specific knowledge bases, such as those created using the Leiden Open Variation Database (LOVD) [16]; (3) proprietary knowledge bases, typically owned and managed by genetic testing laboratories, who maintain exclusive access [17]; and (4) publicly available, centrally-managed repositories, such as ClinVar [18]. Typically, when a new variant is discovered, or new information about a known variant is made available, this information will be recorded in one or more of these knowledge bases. Furthermore, curators may monitor publications and reports in order to update a knowledge base accordingly.

ClinVar, which is a publicly available central resource managed by the National Library of Medicine, represents a model wherein genome knowledge sources can upload their expertly curated knowledge into one location [19]. Previously, genome knowledge consumers may have had to use several different genome variant knowledge bases and pay to access particular knowledge. Furthermore, with an open collaborative approach to genome variant annotation, ClinVar may become a more robust and extensive knowledge base than any single locus-specific or laboratory-managed knowledge bases. Open access, locus-specific knowledge bases tend to be curated and maintained on a volunteer basis, making the knowledge available limited. While laboratory-managed knowledge bases contain the best variant knowledge, they are also (1) limited by the number of unique variants observed by that laboratory and (2) may have tightly controlled access to the variant knowledge in order to maintain a competitive advantage over other testing laboratories [17]. Nevertheless, if ClinVar is embraced by the diagnostic laboratory community with the support of the ClinGen effort [20], the laboratory knowledge bases will likely serve as one of the most important sources of variant annotations. Additionally, several characteristics of ClinVar make it attractive for our type of work:

Format – ClinVar maintains a health data repository available via FTP download in several release formats (e.g. TSV, XML, and VCF). In particular, the tab separated values release format, which provides data in a structure similar to relational database tables, is the easiest data format to be used in the Solor transformation process.

Documentation – Robust ReadMe files within each ClinVar release, describing in detail every data point contained within the overall ClinVar release data structure. Based on these descriptions, reliable inferences can be constructed for the Solor transformation process.

Release Cycle – Within the ClinVar release data tables, there exists variations (e.g. daily, weekly, monthly, etc.) of update frequency amongst individual data entities. Variant data is updated weekly, whereas phenotypic data is updated daily. Creating a Solor transformation process around data entities that are frequently updated results in more current variant data for the Solor system.

Data Structure – Specific data entities, such as variant, gene, and disease, can be normalized, modular, and isolated from other more complex entity relationships. These aspects for such key data entities result in a less complex, more straightforward implementation of the Solor transformation process.

Variant Identifier – ClinVar utilizes the Human Genome Variation Society (HGVS) specification for naming genomic variants contained within each release. Leveraging approved standards, as part of key data elements being transformed into the Solor system, enables proper terminology concept quality assurance and classifications to be performed on all Solor health data.

2.3. Ecosystem

To be completed as part of future deliverable.

3. MATERIALS AND METHODS

3.1. Aim 1

3.1.1. Precision Medicine Use Case (CLIN 2005B_01.14)

Use Case 1 develops a Precision Medicine use case for Solor where variants which occur within genes are assessed for clinical impact using the curated genome variant knowledge base ClinVar. ClinVar, which is a publicly available central resource managed by the National Library of Medicine, represents a model wherein genome knowledge bases and laboratories can upload their expertly curated knowledge into one location [19].

Genetic data knowledge sources are not structured or maintained in a format usable for the Electronic Health Records (EHR), clinical decision support, research, or interoperability despite the fact that precision medicine has become a national priority [Ref needed]. The market cost of genetic testing continues to decrease, while at the same time, the number of known genetic variants and number of genetic tests available continue to increase. Consequently, genetic information is becoming a more common addition to an individual's health records with important implications for treatment and research.

It is critical that individual genetic information is incorporated into electronic records in a consistent way so that clinicians and computer decision support systems (CDSS) alike can realize its benefits without errors or ambiguities. Accessible and standardized genetic-based test results and data sets have the potential to help clinicians provide better patient care if integrated into the electronic health record, enable more insightful population health statistics if in a standardized format and contribute to more impactful research if interoperable.

3.1.1.1. Genome Data Acquisition and Database Storage

The ClinVar knowledge source was added to the Solor ecosystem using a transformation process which allows for ClinVar specific data representation within the Solor ecosystem. Incorporating the ClinVar knowledge source into the Solor ecosystem required a custom implemented transformation process, which focused specifically on transforming the ClinVar tab separated value data format into the Solor common model format. Below describes the three data entities and the specific data elements used in the ClinVar to Solor transformation process:

Variant Summary – Contains attribute information that further describes gene variants submitted to ClinVar. The specific name of each variant in the HGVS format and the particular National Center for Biotechnology Information (NCBI) gene ID is used in the Solor transformation process.

Gene Specific Summary – Contains attribute information to further describe individual NCBI managed table of genes, specifically focusing on both gene's identifiers, the NCBI ID and its symbol data elements.

Gene Condition Source ID – Contains all relationships between genes and correlating diseases (phenotypes) used in ClinVar. This data entity contains not only the NCBI gene ID, but also identifiers of external phenotypic terminology concepts. For example, a specific gene ID is correlated with a potential SNOMED CT concept and the associated SNOMED CT Identifier (SCTID).

All variants and genes found in ClinVar were de-duplicated and loaded into the Solor model as unique Solor concepts. Each concept contained both a fully qualified name, based on either the variant's name and or the gene's symbol, as well as String identifiers that were based off the variant's HGVS ID, or the gene's NCBI ID. In addition, parent-child (supertype-subtype) relationships between concepts for variants to concepts for genes, and concepts for genes to SNOMED CT concepts, were encapsulated as logic graph axioms, visualizing a stated (modeled) view of the concepts as well as the view after classification, and assigned to each respective Solor concept. Lastly, a comprehensive Solor taxonomy was created incorporating both ClinVar and SNOMED CT concept.

3.1.2. Medical Device Interoperability Use Case (CLIN 2005B_02.14)

To be completed as part of future deliverable.

3.1.3. HL7 FHIR Use Case (CLIN 2005B_03.14)

To be completed as part of future deliverable.

3.2. Aim 2 (CLINs 2005B_04.14, 2005B_05.14 and 2005B_06.14)

3.2.1. Evaluation Design

We will perform a formative evaluation of use case constructs – using a qualitative design. Formative studies are particularly useful for applied work, where it is more important to understand the process by which things happen in a particular situation than to measure outcomes rigorously or to compare a given situation with others [21]. Formative evaluation is a common approach for improving the quality of a program being developed by identifying weaknesses throughout the design and development efforts so that it will be as likely as possible to achieve the objectives for which it was designed [22,23]. A formative evaluation aims to help develop and improve programs from an early stage, when opportunities for influence are likely to be greatest, and to identify promising components [24]. Innovative programs provide an ideal environment for use of formative evaluation findings, with key stakeholders generally much more willing to make adjustments at an early stage than when a program is well established [25].

The goal of this formative evaluation is to collect rapid feedback from subject matter experts that would provide validation of use case constructs and context for future successive adaptations and improvement of the use case's development. Having said that, key questions for evaluating a new proof-of-concept include: Does the idea provide a new and more useful capability?; does it help developers better understand complex systems?; and does it demonstrate by its behavior that a complex assembly of components can accomplish a particular set of activities? Our formative evaluation research questions are shown in Table 4.

3.2.2. Evaluation Participants

We combined both purposeful expert sampling and snowball sampling to create an interview strategy to gather knowledge from individuals that have particular expertise[26,27]. We first identified key informants (someone knowledgeable about health informatics) to begin the process of interviewing and we then asked for the names of subject matter experts (individuals especially knowledgeable and experienced with medical terminological systems). In addition, it was also important that participants were available and willing to contribute, and able to effectively communicate their experiences.

3.2.3. Methods Used for Data Collection

This work will use as its primary data gathering method a semi-structured interview approach, as described by Steinar Kvale in Doing Interviews [28]. It's a fairly open approach where a guide is used, with questions and topics to be covered. The evaluator has some discretion with the order in which questions are asked, but the questions are standardized, and provided to ensure that the researcher covers the correct material. Unlike the structured interview where the questions are fixed and they are asked in a specific order, questions or topics can be further developed on the basis of responses from the interviewee. Semi-structured interviews allow for in-depth encounters in which focused, conversational, two-way communication is used to elicit detailed narratives and are often used by evaluators wanting to delve deeply into a topic and to thoroughly understand the answers provided.

This approach aligns with the approach for conducting semi-structured interviews described in the RAND Corporation report "Data Collection Methods: Semi-structured Interviews and Focus Groups" [29]. An overview of the important aspects of semi-structured interviews includes a number of steps. First, the main research questions need to be identified. In other words, what does the researcher hope to learn? Next, the researcher needs to consider the different participant types and determine the sampling. This study used judgment/purposeful sampling where individuals were selected based on their knowledge of medical terminologies, and because their opinion was judged to be important to the research [27].

Interviews are typically personal and intimate encounters that allow for focused, conversational, two-way communication in which open, direct, verbal questions are used to elicit detailed narratives and stories[30]. This study conducted semi-structured interviews where an interview is defined as: a method of data collection in which one person (an interviewer) asks questions of another person (a respondent) either face-to-face or by telephone[31]. Although no interview can truly be considered structured, they were relatively structured and more or less equivalent to guided conversations.

We engaged participants at a single point in time, individually, using virtual meeting software, and conducted open-ended, semi-structured interviews. Participants were contacted by email to invite them to participate and a meeting time was then set at a time and day of their convenience. The total time was allotted no more than two hours for the investigators to complete the interactions. Participation in this study was voluntary and the subject matter experts could choose not to take part in the interview. The subject matter experts could also skip any question they preferred not to answer or terminate the interview without penalty. We asked each participant four demographic questions: (1) job title, (2) number of years of experience, (3) education level and (4) previous terminology experience. All demographic data gathered about the participant were free text.

3.2.4. Methods Used for Data Analysis

Applied thematic analysis, a method for identifying and analyzing patterns of meaning in a dataset, was used to organize and describe the data collected from the interviews [32–34]. Applied thematic analysis provided a rigorous, yet inductive, set of procedures designed to identify and examine themes from textual data in a way that is transparent and credible [35]. The procedure for performing an applied thematic analysis had the following steps: (1) collect data, (2) transcribe conversations, (3) list patterns of experience, which can come from direct quotes or paraphrasing common ideas, 4) identify data that relate to already classified patterns, (5) combine and catalog related patterns into themes, and (6) formulate theme statements and develop a summary of findings.

3.2.5. Precision Medicine Use Case

Our precision medicine use case formative evaluation questions and semi-structured interview questions are shown in Table 1 and Table 2, respectively. The questions may have been modified in light of what is learned during the interview and to fit the expertise of the interviewee. See Appendix 1 for full Interview Guide.

Table 1: Precision medicine use case formative evaluation questions.

Use Case Construct	Formative Evaluation Questions
Knowledge Source(s)	What are the publicly available (domestic or international) non-proprietary sources of information for Genome Variant – Clinical Impact knowledge?
Solor System Integration	Does the integration of ClinVar into the Solor System seem to be a sound and reasonable approach for promoting genomic data set use in a clinical setting?
Relevance	Does our work contribute to advancing precision medicine and genotype-phenotype interoperability?

Table 2: Precision medicine use case semi-structured interview questions.

Use Case Construct	Semi-Structured Interview Questions
Knowledge Source(s)	• Is the ClinVar knowledge source used in our use case a valid knowledge source? • Are there any additional sources that could be utilized? • Are there any sources that should not be utilized? If so, why not?
Solor System Integration	• Do you think this approach to integrating the ClinVar knowledge source is reasonable?
Relevance	• Does this use case advance genomic interoperability? • How might this use case be extended and generalizable?

3.2.6. Medical Device Interoperability Use Case

To be completed as part of future deliverable.

3.2.7. HL7 FHIR Use Case

To be completed as part of future deliverable.

4. RESULTS

To be completed as part of future deliverable.

4.1. Precision Medicine Use Case (CLIN 2005B_07.14)

4.1.1. Participants

We interviewed three individuals with the participant characteristics described in Table 3. Participants had leadership and technical roles with 15-20 years of experience and were subject matter experts in the domain of precision medicine with knowledge of healthcare standards, terminologies, knowledge commons and genomic databases. All subjects had experience with precision medicine, ranging from 1 to 13 years, mean of 5.67 years.

Table 3 Participant Characteristics

Participant	Job Title	Professional Experience (years)	Education Level	Precision Medicine Experience (years)
1	Senior Manager	15	MS	3
2	Executive	20	PhD	13
3	Specialist Leader	15	PhD	1

Participant 1 had a wealth of knowledge related to technological health care solutions. After a career as a general nurse practitioner and public health professional, she shifted focus to Health Informatics where she has worked on electronic health record transformation as well as the development of software solutions to solve life science and health care problems. Recently, she has led National Institute of Health (NIH) health strategy and analytics projects. She did not have any specific experience with genomic data.

Participant 2 was well versed in the field of genomic data. He first started working at the NIH nearly 20 years ago on an intermittent basis but has been working full time on various NIH projects for the last 13 years. Due to his work experience with the National Cancer Institute (NCI), he has a large amount of experience specifically with genomic data. Through this work, he is familiar of the idea of using genomic data for precision and personalized medicine.

Participant 3 had a wealth of clinical research experience. She has been involved with biomedical research for over a decade during her PhD and postdoc years. She has experience at NIH as well as years of experience in the research and clinical trial arena with Military Health Systems (MHS). Furthermore, due to her background with molecular biology, she has research experience manipulating the promoter and enhancer regions of a gene with a pharmacologic perspective.

4.1.2. Semi-Structured Interviews

Between November 1st and November 30th, 2018, we performed three semi-structed interviews. The interviews were facilitated by the Use Case Development team. Virtual meetings were arranged at times convenient for all three attendees. There was an interview presentation to guide the conversation that included slides on Solor background, genomic-phenotype motivation, ClinVar knowledge source approach, and its integration with the KOMET GUI prototype. Each participant was asked the interview questions shown in Table 2. Interviews lasted approximately 30 minutes. These semi-structured interviews were based on components from the PRECEDE-PROCEED model [36] to identify key information about each expert's background, experience with ClinVar/ genomics data, and their insights about Predisposing, Reinforcing and Enabling Constructs in Educational Diagnosis and Evaluation (PRECEDE), and Policy, Regulatory, and Organizational Constructs in Educational and Environmental Development (PROCEED) – in the precision medicine environment. PRECEDE involves assessing community factors by determining the social problems and needs of a given population, the determinants of an identified problem, as well as the behavioral and environmental determinants that predispose, reinforce, and enable certain behaviors [36] PROCEED involves the identification of outcomes and implementation by assessing availability for resources, whether certain programs are reaching intended populations, and evaluating behaviors based on incidence of negative/positive behaviors [36].

4.1.3. Applied Thematic Analysis

We performed an Applied Thematic Analysis [34]. We conducted the analysis concurrently to data collection; we continually examined and analyzed the data in an attempt to identify and articulate patterns or themes noticed during the interviews. Our analysis involved a constant iteration between interview data, coded transcript extracts and the forming themes. Writing was an integral part of the analysis lifecycle, beginning with the jotting down of ideas and through the analysis process.

In the first step, we familiarized ourselves with the data. The interview audio recordings were transcribed into text document transcripts. We immersed ourselves in the data by repeatedly reading and rereading the interview transcripts, searching for meaning and patterns and becoming familiar with the breadth and depth of the content. Next, an initial list of codes was generated from the transcript of what appeared to be an interesting feature in the data, where codes refer to the most basic element of raw interview data [37]. We organized codes into validation and recommendation statements supported by participant interview excerpts, or snippets, as shown in Table 4, and patterns across the interview data began to form.

Table 4 Summary of understanding of interview data.

Construct	Validation	Recommendation	Participant Excerpts (Snippets)
Knowledge Source(s)	The presented knowledge source(s) is a good source for the precision medicine use case.	• Seek a more authoritative answer from individuals who may have more knowledge of specific databases.	• "ClinVar is a great first stop to find genomic data" • "I can connect you with some people that could probably answer .. what other data sets .. that you should be looking into."
Solor System Integration	The integration of ClinVar and clinical terminologies seems reasonable.	• The current integration effectively shows the connections between variants, genes, and disorders. • Get an early adopter for better guidance.	• "Seems like this will be extremely useful for physicians to see these relationships [between genes, variants, and disorders]" • "Get an early adopter to pick it up fairly quickly so you can get better guidance on whether it's useful and whether the user interface offers something to them that helps them make a decision."
Relevance	This type of work can move precision medicine interoperability forward.	• This is useful for preventative medicine. • Include information on the correlations between genes. • Show the relationships between genomic data and treatment plans.	• "Really useful for a physician to be able to take a look at what came back and use it to inform preventative measures or suggest lifestyle changes" • "Correlations like [those between genes on the same locus] will be extremely useful for physicians in making clinical decisions" • "my next question is… how do I use it for determining the right treatment for the patient?"

Next, patterns were organized to a broader level of summary of findings that captured something important about the data or meaning within the data set, as shown in Table 5.

Table 5 Summary of findings of interview data.

Construct	Context Related to Materials & Methods	Summary of Findings
Knowledge Source(s)	We use a publicly available knowledge source called ClinVar which is available through the National Library of Medicine. ClinVar reports the relationships between human variations and phenotypes .	ClinVar is an appropriate starting point and valid to demonstrate the value of this use case. However, more research must be done to validate the use of the ClinVar knowledge source compared to other existing genomic data sets.
Solor System Integration	We integrated a knowledge source into the Solor platform and created a common model, allowing for a ClinVar specific data representation within the Solor ecosystem.	The ClinVar knowledge source has been successfully integrated into the Solor platform to effectively demonstrate the connections between genes, variants, and disorders. However, to continue to improve the Solor tool in this use case, it is important to get an early adopter to being using this tool in a real-world setting.
Relevance	The integration of the ClinVar data source into the Solor model can be used to increase precision medicine interoperability.	The precision medicine use case of Solor has many potential improvements that will make it more clinically useful. These include, but are not limited to, treatment plan support and gene correlations.

For the final step, themes were developed that represent something important about the data in relation to the evaluation question. The following themes emerged:

- **Theme**: More research needs to be done to ensure the correct knowledge source is selected.
    - **Subtheme:** ClinVar is a good starting point, and demonstrates the potential of this precision medicine use case.
    - **Subtheme:** There are many data sets available. Seek guidance from someone who has more knowledge related to the type of knowledge sources that exist.
- **Theme**: The precision medicine use case must be expanded to think about how it can support clinical decision making.
    - **Subtheme:** We can incorporate more information into the taxonomy tree to help with this clinical decision support, including treatment plan support and gene correlations.
    - **Subtheme:** Get input from clinicians to better guide the development of this use case.

4.1.4. Findings Summary

The goal of the semi-structured interview process was to evaluate key constructs of our Solor precision medicine use case. The results from the first construct, knowledge source, were broadly positive. Interviewees could easily conceptualize how ClinVar might inform the understanding of a genotype-phenotype knowledge use and how there might be additional resources that could be leveraged to assist in this understanding.

The subject matter expert who seemed to have the greatest knowledge related to the ClinVar data set stated that it is a good source to find information mapping genotypes to phenotypes. However, he did caution that his expertise is in the research realm, while the application of the precision medicine use case is geared more directly toward clinical decision support. He could not confidently say if there were other data sets available that are more applicable to clinical decision support. Consistent across the three subject matter experts was the sentiment that with that vast amount of available data, many data sets can and should be considered. ClinVar seems to be a viable option and good starting point, however our subject matter experts were unable to confirm that it was the "best" data source for the precision medicine use case. Therefore, more research must be done before this can be taken toward a fully usable product capable of improving patient safety and clinical decision support.

Furthermore, the subject matter experts each reported some findings and opinions in terms of how this precision medicine use case can assist clinical decision support. Each subject matter expert shared the opinion that this seems to be an extremely effective way to view and analyze the connection between genes, variants, and diseases as well as the associated SNOMED CT code. Several of the subject matter experts agreed that this already lends itself to the application of preventative medicine, which aligns nicely with the recent trend of a focus on preventative medicine present throughout the medical field. However, given the ability of the precision medicine use case to demonstrate the relationships between genes, variants, and diseases, there are several improvements that can be made to facilitate a more effective and useful tool.

With the idea of clinical decision support in mind, the precision medicine use case can be taken one step further to include treatment plans. Currently, the precision medicine use case utilizes Solor to effectively provide information to a clinician regarding the various genes related to a disease, but it does not give any guidance on how treatment can be personalized based on that individual's genome. The precision medicine use case lends itself to include efficacy of treatment plans for specific gene types. This would likely include the utilization of another data source, so the subject matter expert suggested the involvement of a data scientist who could help ensure that the data is consumed properly while being imported into the Solor taxonomy tree.

Additionally, another extension to consider for our precision medicine use case that would assist a physician in clinical decision support, is the ability to connect genes that are correlated to each other. Often, a genetic mutation in one area can affect the entire gene locus, essentially causing a ripple effect and increasing the likelihood of other conditions that are associated with other genes on that same locus. An example given by one of the subject matter experts was the idea that an individual with an underbite may be more prone to developing a heart murmur, due to the genes associated with these disorders sharing a gene locus. Understanding these correlations would be extremely useful for physicians as they make clinical decisions. Once again however, this type of extension to the precision medicine use case would likely involve the integration of another data source.

Overall, an early adopter at NIH should be identified to collaborate on the precision medicine use case of Solor. The subject matter experts provided encouraging feedback about the ability for this use case to assist in improving patient safety and clinical decision support. The precision medicine use case to date has showed the ability to effectively form relationships between genotypes and phenotypes. This can immediately have an impact on certain preventative medicine measures. Additionally, it has the ability to be extended into a more robust model that can influence clinical decision-making processes by giving physicians extensive information not only about efficacy of treatment plans among genetic populations but also about gene-to-gene correlations and their effect on phenotypic likelihoods. Because it has been

demonstrated that this use case can be useful, it is paramount that an early adopter is identified to begin interacting with the Solor tool in a clinical environment. This will provide meaningful feedback from a physician's perspective, resulting in an effective and useful tool that assist clinical decision support and in turn improve patient care.

4.2. Medical Device Interoperability Use Case (CLIN 2005B_08.14)

To be completed as part of future deliverable.

4.3. HL7 FHIR Use Case (CLIN 2005B_09.14)

To be completed as part of future deliverable.

5. CONCLUSION

To be completed as part of future deliverable.

5.1. Limitations of the Work

To be completed as part of future deliverable.

- Small interview participant size.

5.2. Suggestions for Future Work

To be completed as part of future deliverable.

REFERENCES

1 Williams R, Kontopantelis E, Buchan I, *et al.* Clinical code set engineering for reusing EHR data for research: A review. *J Biomed Inform* 2017;**70**:1–13. doi:10.1016/j.jbi.2017.04.010

2 Wright A, Henkin S, Feblowitz J, *et al.* Early Results of the Meaningful Use Program for Electronic Health Records. *N Engl J Med* 2013;**368**:779–80. doi:10.1056/NEJMc1213481

3 Aspden P, Corrigan JM, Wolcott J ES, editor. *Patient safety: achieving a new standard for care.* Washington, D.C.: : National Academy Press 2003.

4 Fadly A El, Daniel C, Bousquet C, *et al.* Electronic Healthcare Record and Clinical Research in Cardiovascular Radiology. HL7 CDA and CDISC ODM Interoperability. *AMIA 2007 Symp Proc* 2007;:216–20. doi:10.1186/s12955-014-0106-3

5 The Office of the National Coordinator for Health Information Technology, Office of the Secretary USD of H and HS. FEDERAL HEALTH IT STRATEGIC PLAN 2015 – 2020. 2015.

6 IHTSDO. SNOMED CT Starter Guide. *Snomed* 2014;:1–56.

7 Bodenreider O, Cornet R, Vreeman DJ. Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. *Yearb Med Inform* 2018;**27**:129–39. doi:10.1055/s-0038-1667077

8 Regenstrief Institute I. Logical Observation Identifiers Names and Codes (LOINC®). Log. Obs. Identifiers Names Codes. 2017.http://www.loinc.org (accessed 6 Oct2017).

9 NLM. RxNorm Overview. 2013.http://www.nlm.nih.gov/research/umls/rxnorm/overview.html

10 Zhang* G-Q, Zhu* W, Sun* M, *et al.* MaPLE: A MapReduce Pipeline for Lattice-based Evaluation and Its Application to SNOMED CT. *Proc IEEE Int Conf Big Data* 2014;**29**:754–9. doi:10.1016/j.biotechadv.2011.08.021.Secreted

11 Cimino J. Desiderata for Controlled Medical Vocabularies in the Twenty- First Century. *Methods Inf Med* 1988;**37**:394–403. doi:10.2307/1213005

12 Raje S, Bodenreider O. Interoperability of Disease Concepts in Clinical and Research Ontologies: Contrasting Coverage and Structure in the Disease Ontology and SNOMED CT. *Stud Health Technol Inform* 2017;**245**:925–9.

13 Winnenburg R, Bodenreider O. Issues in creating and maintaining value sets for clinical quality measures. *AMIA Annu Symp Proc* 2012;**2012**:988–96.

14 Bahr NJ, Nelson SD, Winnenburg R, *et al.* Eliciting the Intension of Drug Value Sets - Principles and Quality Assurance Applications. *Stud Health Technol Inform* 2017;**245**:843–7.

15 Stenson PD, Mort M, Ball E V., *et al.* The human gene mutation database: 2008 update. *Genome Med* 2009;**1**:1–6. doi:10.1186/gm13

16 Fokkema IFAC, Taschner PEM, Schaafsma GCP, *et al.* LOVD v.2.0: The next generation in gene variant databases. *Hum Mutat* 2011;**32**:557–63. doi:10.1002/humu.21438

17 Cook-Deegan R, Conley JM, Evans JP, *et al.* The next controversy in genetic testing: Clinical data as trade secrets? *Eur J Hum Genet* 2013;**21**:585–8. doi:10.1038/ejhg.2012.217

18 US Natl Libr Med. National Center for Biotechnology Information. ClinVar. 2012.www.ncbi.nlm.nih.gov/clinvar/ (accessed 28 Nov2018).

19 Welch BM, Loya SR, Eilbeck K, *et al.* A proposed clinical decision support architecture capable of supporting whole genome sequence information. *J Pers Med* 2014;**4**:176–99. doi:10.3390/jpm4020176

20 Riggs ER, Wain KE, Riethmaier D, *et al.* Towards a Universal Clinical Genomics Database: The 2012 International Standards for Cytogenomic Arrays Consortium Meeting. *Hum Mutat* 2013;**34**:915–9. doi:10.1002/humu.22306

21 Bickman L, Rog D. *Handbook Of Applied Social Research Methods [e-book]*. Thousand Oaks, CA: : SAGE Publications, Inc 1998.

22 Brown KG, Brown KG, Gerhardt MW, *et al.* Formative Evaluation: an Integrative Practice Model and Case Study. *Pers Psychol* 2002;**55**:951–83.

23 Salie S, Schlechter A. A formative evaluation of a staff reward and recognition programme. *SA J Hum Resour Manag* 2012;**10**:1–11. doi:10.4102/sajhrm.v10i3.422

24 Henry GT, Smith AA, Kershaw DC, *et al.* Formative Evaluation: Estimating Preliminary Outcomes and Testing Rival Explanations. *Am J Eval* 2013;**34**:465–85. doi:10.1177/1098214013502577

25 Dehar M-A, Casswell S, Duignan P. Formative and Process Evaluation of Health Promotion and Disease Prevention Programs. *Eval Rev* 1993;**17**:204–20. doi:10.1177/0193841X9301700205

26 Lavrakas P. *Encyclopedia of Survey Research Methods*. 2455 Teller Road, Thousand Oaks California 91320 United States: : SAGE Publications, Inc. 2008. doi:10.4135/9781412963947

27 Palinkas LA, Horwitz SM, Green CA, *et al.* Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Adm Policy Ment Heal* 2015;**42**:533–544. doi:10.1007/s10488-013-0528-y.Purposeful

28 Kvale S. *Doing interviews*. London: : SAGE Publications 2007.

29 Margaret C. Harrell; Melissa A. Bradley. *Data Collection Methods Semi-Structured Interviews and Focus Groups*. 2009. doi:978-0-8330-4889-9

30 DiCicco-Bloom B, Crabtree BF. The qualitative research interview. *Med Educ* 2006;**40**:314–21. doi:10.1111/j.1365-2929.2006.02418.x

31 Whiting LS. Semi-structured interviews: guidance for novice researchers. *Nurs Stand* 2008;**22**:35–40. doi:10.7748/ns2008.02.22.23.35.c6420

32 Aronson J. A pragmatic view of thematic analysis. *Qual Rep* 1994;**2**:3. doi:10.4135/9781446214565.n17

33 Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2006;**3**:77–101. doi:10.1191/1478088706qp063oa

34 Clarke V, Brown V, Hayfield N. Thematic Analysis. Qual. Psychol. A Pract. Guid. to Res. Methods. 2015;**9760**:222–47. doi:10.1037/13620-004

35 Guest G, MacQueen K. *Applied Thematic Analysis.* Los Angeles: : Sage Publications 2012.

36 RHI HUB. PRECEDE-PROCEED Model - Rural Health Promotion and Disease Prevention Toolkit. Defining Health Promotion and Disease Prevention - RHIhub Toolkit. https://www.ruralhealthinfo.org/toolkits/health-promotion/2/program-models/precede-proceed (accessed 20 Aug2008).

37 Boyatzis RE. *Transforming qualitative information : thematic analysis and code development*. Thousand Oaks, CA : Sage Publications 1998.

Appendix 1

Introductory questions

- Can you tell me a little about yourself and your role?
- Can you tell me about the organization you work for and what it does?
- For how long have you worked with your organization?

Background with Genomic datasets

- Could you describe the level of experience you have with integrating genomic data?
- What resources do you think made working with genomics data easier/ more effective?
- Probe: What are some potential barriers that you feel present a challenge?
- Probe: What solutions have you deployed?
- Probe: Could you describe for us the successful strategies you or others have used for successful management of genomics data?

ClinVar

- To what extent have you reviewed/used ClinVar data? How familiar would you say you are?
- Does the ClinVar knowledge source used here seem like it could be useful in understanding gene variant – clinical impact?
- Are there any additional sources that could be utilized?
- Are there any sources that should not be utilized? If so, why not?

SOLOR demo

- What is understandable and what is confusing?
- What is ambiguous?
- Are there specific relationships (variant-gene or gene-disorder) that are easier/harder to interpret using SOLOR versus other data sources?
- Do you think this approach to integrating ClinVar is valid?
- Are there ClinVar data elements that we didn't use but should use?
- Are there other clinical terminology system relationships that can be used other than SNOMED CT?
- What quality assurance/control issues should be considered? (i.e., should a genomic SME perform reviews)

Ecosystem

- Overall, how do you think implementation of SOLOR could work for improving genomic data integration?
- Going forward, what things do you need to continue to effectively interpret genomic data relationships in SOLOR?
- What advice or input would you like to share with the genomics terminology community about what has worked well and what could be done differently in the interpretability of genomics data elements?
- What lessons have you learned about genomics data elements that you would want to share with others?
- What types of standards, policies, or industry changes do you think are needed to help achieve standard representations of genomics data elements?

And finally, we'd like to ask you:

- Are there any questions we did not ask that you think we should have asked?

Do you have any questions for us? That's all the questions we have for you today. Thank you for your time and for sharing your insights on these topics.