PROJECT NARRATIVE

A Introduction

For decades, we have wanted to harvest the latent knowledge and alerts in electronic health records (EHRs), connect with public health, and improve care delivery through technology. However, barriers to data normalization, standardization, and consistency preclude the efficient use of EHR data, delaying achievement of widespread data reuse.

Data liquidity ultimately requires the transformation of non-standard patient data into comparable, consistent formats. Currently, clinical data and unstructured information from many providers is non-standard. We propose research to generate a framework of open-source services that can be dynamically configured to transform idiosyncratic health information into standards-conforming, comparable information suitable for large-scale analyses, inferencing, and integration of disparate health data. We will apply these services to phenotype recognition (disease, risk factor, eligibility, or adverse event) in medical centers and population-based settings. Finally, we will examine data quality and repair strategies with real-world evaluations of their behavior in Clinical and Translational Science Awards (CTSAs), health information exchanges (HIEs), and National Health Information Network (NHIN) connections.

A.1 Vision Statement. We have assembled a federated informatics research community committed to open-source resources that can industrially scale to address barriers to the broad-based, facile, and ethical use of EHR data for secondary purposes. We will collaborate to create, evaluate, and refine informatics artifacts that advance the capacity to efficiently leverage EHR data to improve care, generate new knowledge, and address population needs. Our goal is to make these artifacts available to the community of secondary EHR data users, manifest as open-source tools, services, and scalable software. In addition, we have partnered with industry developers who can make these resources available with commercial deployment and support.

A.2 Key Challenges. We posit that the fundamental challenges affect the secondary use of EHR data: 1) Data exists in incompatible and inconsistent formats; 2) The mechanisms to harvest data for

specific purposes are not formalized or readily available; and 3) EHR data is of variable quality. To address these challenges, we propose three thematic foci, which we will address with multiple, interconnecting projects. These foci are: 1) Data Normalization; 2) Phenotype Recognition; and 3) Data Quality and Evaluation Frameworks. We have six projects, though they are strongly intertwined and mutually dependent. These projects are: 1) Semantic and Syntactic Normalization; 2) NLP; 3) Phenotype Applications; 4) Performance Optimization; 5) Data Quality Metrics; and 6) Evaluation Frameworks. The first two projects align with our Data Normalization theme, the Phenotype Applications theme and project are the same, Performance Optimization spans themes 1 and 2 (Normalization and Phenotyping), while the last two projects correspond to our third theme. This organization is summarized in Table 1.

Table 1: Themes and Projects for Secondary Use of EHR Data

Themes			Projects	Players	Benefits
Data Normalization	Phenotype Recognition	Data Quality and Evaluation Frameworks	Clinical Data Normalization	IBM, Mayo, Utah, Agilex	This is the foundation set of services on which everything below is based. A core element is semantic content (terminology). These resources will be manifest as UIMA services.
			Natural Language Processing (NLP)	Group Health, IBM, Utah ,Mayo, MIT, SUNY, i2b2, Pittsburgh, Colorado	Leverages the existing body of open-source NLP work and data normalization to extract findings from free-text notes using a scalable platform (UIMA). Innovation will emerge from the integration of relationship discovery to support more phenotyping applications.
			Phenotyping	CDISC, Centerphase, Mayo, Utah	This forms the core of applying normalization techniques to EHR data through algorithms and services that permit the identification of “in silico” cohorts, patients meeting eligibility criteria, and population surveillance.
			UMIA and Scaling Capacity	IBM, Mayo	UIMA forms the common framework for the suite of services and functions we will deliver. This project ensures that these software resources can scale to near–real-time performance or operate as standalone applications.
			Data Quality	Mayo, Utah	Having normalized data does not help if phenotyping encounters conflicting or inconsistent data. This project provides high-confidence algorithms and services to detect and optionally reconcile such data.
			Evaluation Framework	Agilex, MN HIE, Mayo, Utah	Assessing and validating this service framework by pieces and ensemble will continuously inform and refine our progress and development.
Overall benefits: This SHARP program will deliver open-source software resources and commercially supported versions that allow secondary users of EHR data to meaningfully blend data from multiple sources to achieve common tasks such as cohort identification, protocol eligibility, and surveillance.					

An obvious challenge is that these projects cannot be undertaken independently. Phenotyping relies on normalization to operate on consistent data, while the modalities of normalization are driven by phenotyping use-cases. Optimization is a balance between performance and functionality, interplaying with both normalization and phenotyping. Data quality metrics and repairs affect phenotyping and can be made better or worse by normalization methods; imputed repairs, in turn, can worsen or improve the precision and recall of phenotyping algorithms. Finally, evaluation can uncover practical shortcomings in all of the pipelines, while permutations of pipeline methods and services will change evaluation results.

We believe we have identified a parsimonious body of work on which we can focus coordinated resources. The scope of this work we believe is well-balanced to accommodate interacting factors while providing feasible boundaries. Our philosophy is to produce a suite of practical and efficient services, focused on improving the secondary use of EHR data.

B Approach.

We propose to assemble modular services and agents from existing open-source software resources to improve the utilization of EHR data for a spectrum of use-cases. We focus on three themes: Normalization, Phenotypes, and Data Quality/Evaluation. Our six projects span one or more of these themes, though together they constitute a coherent ensemble of related research and development. Finally, these services will have open-source deployments as well as commercially supported implementations. A uniform technology base for these services is the open-source UIMA developed at Watson Research Labs by IBM. The UIMA has served as the basis for Mayo's open-source NLP pipeline[1,2] and provides a common framework for the agents and pipelines we will adopt, develop, evaluate, and refine in this SHARP consortium program.

B.1 Clinical Data Normalization Services and Pipelines

B.1.1 Specific Aims

1) Build generalizable data normalization pipeline – ST
2) Semantic normalization annotators involving LexEVS – ST

3) Establish a globally available resource for health terminologies and value sets – LT

4) Establish and expand modular library of normalization algorithms – LT

B.1.2 Previous Work. Mayo Clinic and Intermountain Healthcare both have addressed the challenge of data normalization within their organizations, informed by national and international standards. Dr. Stan Huff is a past chair of HL7 and oversaw many of the transitions to model-driven message development and semantic binding. Drs. Huff and Chute were both co-chairs of the Vocabulary Technical Committee at HL7 and have contributed to the current HITSP standards specifications; both currently serve on the ONC HIT Standards Committee and contributed to authoring the Meaningful Use specifications. Dr. Chute currently chairs the ISO TC-215 Technical Committee on Health Informatics (standards) globally and also chairs the International Classification of Disease (ICD) Revision Steering Committee for ICD-11 as part of the World Health Organization (WHO). Building on these and other foundations for standards conformance, both organizations have compiled extensive algorithms and data models[3,4] that are available for this proposal. In addition, Regenstrief has made its data normalization software and algorithms open-source as part of its Health Open Source Software (HOSS) Collaborative,[5] which will further inform our efforts.

Mayo has collaborated with the clinical data standards community on the specification of Common Terminology Services (CTS), which is now an ISO and HL7 standard. The second version, CTSII, has achieved DSTU status at HL7 and will be formalized by the Object Management Group (OMG) as an information technology standard analogous to UML (Unified Modeling Language), though for terminology. Deriving from and informing the CTS process is LexEVS (Enterprise Vocabulary System), a body of open-source work started under the LexGrid Vocabulary Services for the caBIG (LexBIG) project[6], which was founded as part of caBIG development. Over time, this technology has been adopted as a strategic infrastructure for the caBIG community and the NCI Enterprise Vocabulary Services (EVS), as well as the National Center for Biomedical Ontology (NCBO)[7]. It provides programmable interfaces to allow users to access controlled terminologies supplied by the NCI EVS, NCBO, or any organization adopting the technology. It is a vocabulary server with a well-structured application programming

interface (API) capable of accessing and distributing vocabularies as commodity resources. In particular, the primary objectives of LexEVS include:

- Implementation of robust, scalable, open-source, EVS-compliant vocabulary services. Currently, the software is distributed as representational state transfer (REST)ful and simple object access protocol (SOAP)-based Web services with a core Java API.
- Providing a flexible implementation for vocabulary storage and persistence, allowing alternative mechanisms without affecting client applications or end users.
- Providing standard tooling for load and distribution of vocabulary content. This includes but is not limited to support of standardized representations (e.g., UMLS Rich Release Format, the OWL web ontology language, Open Biomedical Ontologies).

The LexEVS project applies the LexGrid vision of supporting a distributed network of lexical resources (e.g., terminologies, ontologies) via standards-based tools, storage formats, and access/update mechanisms. The LexGrid model defines how vocabularies should be formatted and represented programmatically and is intended to be flexible enough to accurately represent a wide variety of vocabularies and other lexicon-based resources. The model also defines several different server storage mechanisms (e.g., relational database, lightweight directory access protocol [LDAP] and an XML format). It provides the core representation and a common model for all data managed and retrieved through the LexEVS. The master LexGrid model[8] is maintained in an XML schema, which is then transformed into UML and other representational forms for publication. The model comprises multiple high-level objects that form the core in uniformly representing ontologies and vocabularies.

Recently, to closely align the LexGrid and LexEVS project with emerging semantic Web technologies, we have initiated preliminary investigations to develop a Resource Description Framework (RDF)-based rendering of the LexGrid model. Such an approach will enable us to use RDF's simple yet powerful metadata model for representing a multitude of vocabularies and terminologies ranging from the UMLS to OBO and OWL ontologies as Web resources that can be accessed and queried, for example, using the RDF query language SPARQL.

B.1.3 Methods. We will leverage the Mayo Clinic and Intermountain Healthcare data models into a common format (the Intermountain – Clinical Element Model [CEM])[3], work that is largely completed under an existing cooperative work program among Mayo, Intermountain, and General Electric Healthcare. These models conform to HITSP and Meaningful Use standards and comprise the target information structures for well-formed data.

We will use the Agilex NHIN Connect tooling, augmented where needed by MirthConnect[9], an open-source standards-based healthcare integration engine that includes HL7 message filtering, transformation, and transmission functions. We will wrap these elements into discrete UIMA pipeline elements, with dataflow, decision point logic, and dynamic pipeline configurations based on the data managed within UIMA. We will design specific data transformation services that will operate as pipeline elements for the denormalized use-cases we encounter: for example, systematically ill-formed V2 messages (including earlier versions – pre-V2.5.1) and semantic errors. Mayo, Intermountain, and the open-source HOSS resource at Regenstreif provide a rich starting set of transformation algorithms for data normalization.

The problem of normalizing semantic areas is one of our particular strengths. We will invoke the LexEVS[8] suite of terminology services described above, containing HL7 and HITSP value sets bound to our target models. Again, this is work that is already partially complete under the same Mayo/Intermountain/GE work program. Based on the CTS[10,11] specification, these services handle primary classification, derivative value sets and, pertinent to our proposal, arbitrary maps between and among these entities. Thus, in our proposal, we will populate and publicly make accessible a LexEVS repository that will be the reference source for our semantic normalizes. Users can access the SHARP LexEVS site (which may be part of the National Center for Biomedical Ontology[12], NCI, or a Mayo site) or choose to clone their own LexEVS server and synchronize content from the project site.

Mayo and Intermountain have already built extensive libraries of synonyms and value set maps, informed by content in the UMLS, HL7, HITSP, USHIK, WordNet and our own experience. These will be selectively updated and refined to match the data transformation challenges we encounter. LexEVS API elements will be wrapped into UIMA pipeline elements, work we have already piloted with Mayo's

NLP entity recognition annotators. Utilization of these services will require specification of the source and target value sets (e.g., local lab codes to LOINC codes) and the creation of these mappings if they do not already exist; Intermountain and GE have created a robust LexEVS editor to manage this task. The obvious long-term effort will be to compile a sufficiently large collection of mappings so that creating a use-case–specific semantic mapping should become rare.

A long-term goal will be to annotate, curate, and manage the library of UIMA transformation services that will emerge from this work. Annotation metadata will include source and target models and value sets, message transport format (CDA, HL7 V2, data warehouse SQL call), and provenance. Our ultimate goal is to avoid rework and reinvention for lack of appropriate search and access to the library.

B.1.4 Key Personnel. Christopher Chute, MD, DrPH, Cui Tao, PhD, Jyotishman Pathak, PhD, Harold Solbrig; Mayo Clinic; Stan Huff, MD, Tom Oniki, MD, Intermountain Healthcare; Kyle Marchant, Agilex Technologies; Dave Ferrucci, IBM Watson Labs.

B.2 Natural Language Processing (NLP). The overarching goal of this project is the development of enabling technologies for high-throughput phenotype extraction from clinical free text. We will also explore the ability to hybridize clinical data extracted from medical reports with the already-structured data in our data repositories to support outcomes research and the development of knowledge useful in clinical decision support. Our focus is NLP and Information Extraction (IE), defined as the transformation of unstructured free text into structured representations. We propose to research and implement modular solutions for the discovery of key components to be used in a wide variety of use cases: comparative effectiveness, clinical research, translational research, and the science of healthcare delivery. Specifically, our efforts are on methodologies for clinical event discovery and semantic relations between these events. Subsequently, the discovered entities and relations will populate templated data structures informed by conventions and standards in the biomedical informatics and general standards communities. The two specific aims reflect the goals of concept discovery and normalization (Aim 1, ST) and semantic processing of the clinical narrative for deep language understanding and extraction (Aim 2, LT).

B.2.1 Specific Aims. Specific Aim 1: Clinical event discovery from the clinical narrative consisting of (1) defining a set of clinical events and a set of attributes to be discovered, (2) identifying standards to serve as templates for attribute/value pairs, (3) creating a "gold standard" through the development of annotation schema, guidelines, and annotation flow, and evaluating the quality of the gold standard, (4) identifying relevant controlled vocabularies and ontologies for broad clinical event coverage, (5) developing and evaluating a methodology for clinical event discovery and template population, (6) extending Mayo Clinic's clinical Text Analysis and Knowledge Extraction System (cTAKES) information model[13], and implementing best-practice solutions for clinical event discovery.

Specific Aim 2: Relation discovery among the clinical events discovered in Aim 1 consisting of (1) defining a set of relevant relations, (2) identifying standards-based information models for templated normalization, (3) creating a gold standard through the development of an annotation schema, guidelines, and annotation flow, and evaluating the quality of the gold standard, (4) developing and evaluating methods for relation discovery and template population, (5) implementing high-throughput scalable phenotype extraction solutions as annotators in cTAKES and UIMA-AS, either within an institution's local network or as a cloud-based deployment integrated with the institution's virtual private network.

B.2.2 Previous Work. We describe our previous work paralleling the above aims. Our experience in *the creation of gold standards and training sets* for a variety of tasks in the biomedical domain is described in a number of publications.[14-21] In addition, Prof. Uzuner led a community-wide annotation effort for medications with 80 researchers as part of the third clinical NLP i2b2 workshop. Prof. Palmer demonstrated that the use of active learning can reduce the required training data by one-third to one-half[22]. Currently, there are no existing multi-layered, semantically annotated, community available clinical corpora for concepts, attributes, and relations, a gap that we propose to address under the current proposal. We have developed *methodologies and toolsets* for both the general and the biomedical domains. The University of Colorado's open-source CLEAR-TK[23] extracts features from the input text and then uses support vector machine (SVM) classifiers to assign semantic labels to constituents[24-27], atomic events (F-score of 75), entities, and their relations as part of the automatic content extraction

(ACE) evaluation (F-score of 85 for entity detection and 77 for relations). Colorado developed a system for recognizing temporal expressions in English and Chinese, which was one of the top-performing systems in the time expression recognition and normalization (TERN) 2004 evaluation, and the discovery of temporal relations[28]. Dr. Chapman has developed modules for determining negation, experiencer, and temporality status of clinical conditions[29-31]. Dr. Chapman and Dr. Haug have developed modules for identifying relations among modifiers and heads and for making higher-level inferences from identified concepts[29,32]. They have also evaluated diagnostic decision support systems capable of reasoning from a combination of NLP-derived and classical structured clinical data[33]. Further evidence of our team's strengths is Profs. Szolovit's and Uzuner's work on relation extraction from the clinical narrative[34-38]. *cTAKES*[2,13] is the core of the proposed NLP work to be extended with novel methodologies to support high-throughput phenotype extraction. cTAKES[2,13] is an open-source modular pipeline of annotators in the UIMA[39] for the discovery and codification of clinical concepts from clinical free text that has been in production to process more than 80 million clinical notes at Mayo Clinic.

Our experience in *high-throughput phenotype extraction* is further evidenced by cTAKES applications to a number of active investigative efforts (e.g., patient-cohort identification for a congestive heart failure study (PI Redfield, Mayo Clinic), the eMERGE GWAS study for ascertaining cardiovascular risk factors for a case-controlled study of peripheral arterial disease using the EMR (PI Chute, Kullo, Mayo Clinic), treatment classification for a PGRN pharmacogenomics of breast cancer treatment study (PI Weinshilboum, Olson, Mayo Clinic), and prospective cohort identification for the management of chronic unstable angina study (PI Roger, Mayo Clinic). The co-investigators have collaborated in the i2b2 NLP space (Uzuner, Szolovits, Savova), an NCBO grant (Chute, Chapman, Savova), an NIDCR grant (Chapman, Haug), an eMERGE GWAS study (Chute, Savova, Carrell), an ARRA-funded grant (Carrell, Chapman, Savova), and another ARRA-funded grant (Chute, Martin, Ward, Palmer, Savova). Under the latter effort, we are integrating CLEAR-TK and cTAKES. The co-investigators form a team with outstanding achievements and a world-class track record in general and biomedical NLP domains.

B.2.3 Methods. The set of clinical events and relations will be defined by the phenotyping algorithms in Project 3. Our proposed work aims to provide flexible and modular methodology for the discovery of concepts/events (e.g., disease, signs/symptoms, procedures, medications, actions, states) and relations (e.g., "treats," "causes"[35], "overlaps," "before," "after"[18]) to support the mapping to a wide range of terminologies, ontologies, and domain models specified by the end user. To determine the normalized forms for entities/events and their attributes, as well as the relations between them, we will draw on the results of Project 1. These models (e.g., HL7 Version 3 Standards[40], University of Utah) will be the normalized forms, or templates, to be populated with values discovered from the free-text clinical narrative to present the most relevant physician/patient encounter information in a structured form.

The creation of the gold standard will consist of the development of an annotation schema and annotation guidelines[17,41] and the evaluation of the annotated corpus quality using Kappa and/or F-score[42]. The initial schema will be developed by the co-investigators. The details of the guidelines' extensions and refinements will be defined through cycles of open annotations in which discussions are allowed. The gold standard will be created by a double annotation (linguistic and clinical experts) methodology for annotations for Treebank[43], Propbank[24] VerbNet[44], named entities, clinical events, and relevant relations, which are the central components of the proposed high-throughput phenotype-extraction system. The resultant de-identified, annotated corpus will include clinical free-text from all participating institutions, ensuring wide representation and system portability. Deidentification will be performed with i2b2 tools[45] and manual validation to remove per HIPAA requirements. A Kappa and F-score greater than 0.7 indicates a high-quality gold standard. Our previous experiences indicate that manual annotations can be completed at a rate of about 10 sentences an hour for treebanking, 60 predicates an hour for PropBanking, five or six entities/events with attributes per hour, and 1.6 to 3 notes per hour per relation-type annotations. We expect to annotate 400,000 tokens of clinical narrative as a gold standard.

The discovery of events and entities is the building block of a relation-extraction system for deep language understanding. We propose using as features the local and global linguistic and domain context through multi-layered linguistic annotations generated by cTAKES, CLEAR-TK, and the tools provided

by the participating co-investigators. The extracted features will be used to train a machine learner to distinguish events from non-events and to label each event. Our proposed relation-extraction methodology relies on the syntactic structure of the sentence and each constituent's semantic role in that sentence (e.g., "patient," "agent," "theme," "location") and employs machine learning. Concept graphs[46,47] will be used as the knowledge representation for the discovered relations across the sentences in the given document. Events and entities are the nodes with the edges between the nodes being derived from the semantic relations. The information in the concept graphs will be used to populate the templates for entities and relations. In addition, we will use active learning techniques[22,32] to minimize the amount of training data. Semi-supervised methods will also automatically generate more training data. Each component is a supervised machine-learning–based classifier trained on manually annotated data to add a specific type of annotation to input text. Each module will be evaluated in a standard 10-fold evaluation (9 folds for training, 1 fold for testing) for its task against the gold standard.

A full cycle of software development will be implemented for cTAKES extensions. Functional and performance test plans will be executed. A scalable, modular, high-throughput toolset will be developed with IBM Watson UIMA-AS architects and engineers (see IBM Watson letter of commitment). It will be deployed and fully tested in a cloud computing environment, work that Seattle's Group Health and Mayo have already started exploring. The resulting software will be released open-source, with dissemination activities for the community. The platform will consist of interchangeable best-of-breed annotators to be assembled in a pipeline by the end user according to their specific applications. Other deliverables are the annotation schemas, annotation guidelines, deidentified annotated corpus to be shared through a data use agreement, and manuscripts on the quality of the annotated corpus, the methods, and their evaluations.

B.2.4 Key Personnel. David Carrell, Seattle's Group Health; Wendy Chapman, University of Pittsburgh; Peter Haug, University of Utah; Jim Martin, University of Colorado; Martha Palmer, University of Colorado; Guergana Savova, Mayo Clinic; Peter Szolovits, MIT and i2b2; Ozlem Uzuner, SUNY and i2b2; and Wayne Ward, University of Colorado.

B.3 Phenotyping. The project and theme of phenotyping encompasses techniques and algorithms that operate on normalized data to identify cohorts of potentially eligible trial patients on the basis of disease, symptoms, or related findings.

B.3.1 Specific Aims. 1) Phenotyping logic specification – ST; 2) Applications of phenotype characterization – S&LT 3) Expansion of Cohort Amplification methods – LT.

B.3.2 Previous Work. Mayo Clinic has been a member of the eMERGE (Electronic Medical Records and Genomics) consortium[48], a cooperative agreement funded by NHGRI (U01-HG04599; Chute PI). This consortium focuses on generalizable algorithms to identify patient phenotypes suitable for Genome-Wide Association Studies (GWAS), constituting "in silico" cohorts. We have been able to demonstrate the reproducible application of disease algorithms that operate on EHR data across consortium members, including hypothyroidism, cataracts, diabetes, peripheral artery disease, and hypertension (manuscripts in review). These algorithms typically involve administrative data, laboratories, medications, and observations derived from NLP. Thus, we have demonstrated methods and algorithms that will form the model for *high-throughput phenotyping* from EHRs. One of our Advisory Committee members, Dr. Dan Masys, is the PI of the coordinating center for eMERGE, enhancing our experience depth.

The ICD-11 Revision, which is chaired by Dr. Chute, includes as a goal the scientific consensus on clinical phenotype. We have experimented with many expression logics for phenotype specification, informed by eMERGE work, the decision support community, and W3C ontology specifications. Thus, we have an additional body of work to contribute to phenotyping expression. Dr. Ustun, who coordinates Health Classifications at the WHO, including ICD, will add additional depth to our advisory board.

A common application of phenotyping logics is the expression of inclusion/exclusion criteria for clinical trials. Mayo has been extremely active in CDISC[49] and BRIDG[50], which are the major standards efforts contributing to HITSP and caBIG in the area of protocol-driven research. BRIDG is a consortium among CDISC, HL7, the NCI, and the FDA to achieve a consensus specification about clinical protocols, including inclusion and exclusion criteria. Dr. Chute just completed a 2-year term as the inaugural board chair of BRIDG and a term on the CDSIC Board. Mayo also contributed a full-time member to the

Technical Harmonization Committee of BRIDG. Dr. Rebecca Kush and the CDISC team will also make important contributions to phenotyping. Thus, the assembled team has deep knowledge and experience in clinical trials specification, including phenotype characterization for cohort definition.

Cohort amplification is a related associative classification model developed at the University of Utah and Intermountain Healthcare. The model is based on machine-learned patterns discovered among the nationally standardized coded EHR data elements populated as byproducts of the care process51,52 A mathematical, variable-threshold attribute reduction algorithm was employed on the basis of the frequency of occurrence of each element for both cases and controls. The patterns reflect the combined actions of multiple healthcare providers in response to the patient's set of problems. The more integrated the EHR, the more complete the picture. Data mining methods offer a solution for noisy, incomplete EHR data. By virtue of sheer volume, accurate patterns emerge. This inelegance is tempered with the requirement that machine-learned patterns be curated by domain experts. Mayo has experimented with similar machine-learning classification systems.[53]

Common data elements are related to phenotype specification. Mayo has been very active in the caBIG data element work, and Harold Solbrig served on the ISO 11179 Metadata standard committee, serving as primary architect of the semantic representations for data elements. More recently, Mayo and CDISC have jointly developed the CDISC Shared Health And Clinical Research Electronic Library (SHARE)[54], an open-source forum for data-element specification based on LexWiki[55] and supported by the NCI. This allows common data elements to have formal semantic binding via LexEVS and to be used in protocol specification, including inclusion and exclusion criteria. These data elements are harmonized with HITSP, because Dr. Kush and the CDISC organization were the primary drivers for the recognition of clinical research as a full-rank use-case for prioritization of HITSP activities and specifications.

Centerphase's technology platform intends to leverage data extracts from multiple AMCs to offer enhanced feasibility assessment services (i.e., robust quantitative and qualitative assessments). The goal is to optimize the study protocol and translate the findings into accelerated study startup and trial implementation. Operationally, Centerphase will assume administrative responsibilities for the trials,

which allows AMCs to focus on their strength (i.e., conducting the trials) and Centerphase to drive efficiencies in study design to make the studies more enrollable and executable.

Centerphase leverages a comprehensive EMR system and leading clinical researchers. The new approach will draw from AMC databases containing information on millions of patients and a network of thousands of physicians to facilitate highly effective identification of patients and physicians for clinical trials. The technology platform will enable the integration and analysis of patient deidentified information across these institutions to accelerate eligibility and trial start-up activities.

B.3.3 Methods. An important aspect of high-throughput phenotyping is the design and development of phenotype-extraction algorithms that are generalizable yet consistent and portable. This inherently obviates the need to develop a "common grammar" that can represent the formal syntax and semantics of the phenotype-extraction algorithms in the form of constraint statements with appropriate boolean and logic operations. To this end, we will investigate a generalizable "covariate grammar" that can be used for formulating queries to determine the phenotype(s) on the basis of the extracted covariate data. Such a grammar will provide a lightweight syntax for representation of phenotype extraction algorithms comprising boolean connectives, attribute-value pairs/groups, and terms and concepts for phenotype ontologies. Furthermore, "flow patterns" describing the control flow of the phenotype extraction algorithms will be studied.

In particular, we will investigate the emerging compositional grammar for SNOMED-CT Expressions in HL7 Version 3 that is currently undergoing development at the International Health Terminology Standards Development Organization (IHTSDO)[56]. This grammar, based on Augmented Backus-Naur Form (ABNF), provides a formal syntax that enables machine-interpretable representation of a concept. For example, the following expression describes an "operation to remove an ovary using a laser:" 83152002|oophorectomy|: 260686004|method|=257820006|laser excision-action|, where 83152002, 260686004, and 257820006 are SNOMED-CT concept identifiers. By extrapolation, this compositional grammar could be used to represent covariate data (and hence the name, “covariate grammar”), using similar "attribute = value" expressions binding to different phenotype ontologies. Not only would such a

generalizable covariate grammar put more rigor in representation of phenotypic traits that can in fact be validated (syntactic validation) automatically using an ABNF parser, but also promote the use of standardized phenotype ontologies for encoding such data.

In practice, a phenotype extraction algorithm typically comprises analysis of multiple covariates that follow a certain logical flow using standard boolean connectives (e.g., OR, AND, NOT) as well as different types of relational operators (in addition to equality [=], e.g., less than [<], greater than equal to [>=]). However, the current release of SNOMED-CT compositional grammar (Version 0.06) is not expressive enough in modeling such boolean connectives and operators[57] Furthermore, the grammar cannot represent control flow patterns (e.g., sequence, parallel split). To address this requirement, this study will investigate how the ABNF definition of the grammar can be expanded to include boolean connectives, relational operators, and basic control flow patterns. In particular, we will investigate how standard relational and boolean connective operators can be added to the basic ABNF syntax, as well as study "flow patterns" such as sequence, parallel split (AND-split), synchronization (AND-join), exclusive choice (XOR-split), and simple merge (XOR-join). We will eventually build a library of phenotype-extraction algorithms represented using the grammar syntax and implemented as “UIMA pipelines” that can be executed on the normalized EMR data

Applications of phenotype characterization will include applying the algorithms to traditional patient selection tasks in clinical research. Building on our eMERGE and ICD11 experience, we will develop configuration utilities that parameterize UIMA pipelines and services. These utilities will address the scripted creation of ABNF phenotype descriptions. Naïve users will have access to tutorials and wizards to facilitate this, though such specification skills are likely to require community support for new users. Some of this support can be delivered through the NCI supported Vocabulary Knowledge Center[58] managed by our group at Mayo.

We will formally evaluate the effectiveness of these scripts and services in Project 6, though our development will iteratively apply these techniques for “in silico” cohort generation (high-throughput phenotyping, typically in partnership with GWAS or other high throughput genomic methods) and

protocol eligibility criteria against gold-standard test datasets for internal validation. Eligibility scripts will follow CDISC and BRIDG motifs, informed by Mayo's participation in Ida Sim's Human Studies Database Project projects across CTSAs, Ontology-Based Integration of Human Studies Data (R01 RR026040) and the Ontology of Clinical Research (OCRe). Finally, the BRIDG eligibility model incorporates risk strata that we can incorporate to inform investigators whether a particular patient is high-risk or some other specified level.

Cohort-amplification applications will be used to iteratively refine phenotype characterization logic (PCL). Patterns are learned among a broad range of PCL exemplar's coded EHR elements to induce predictive rules to identify additional cases. Upon expert adjudication of the rules and evaluation of their predictive accuracy, the PCL can be updated. The software and process steps to accomplish cohort amplification will be documented and packaged for use among collaborators as well as for public download and local usage. Cohort amplification tools will be expanded to support (1) sequential pattern discovery and the currently implemented single- pattern discovery to better accommodate phenotypes such as adverse events or stages of a chronic disease and (2) the concept-hierarchy and "IS-A" terminology mapping feature will be harmonized with project standards-based representations. (LT)

The Cohort Amplification Toolset (CAT), currently implemented at Intermountain Healthcare, is based on an associative classification data mining model[51,52,59] that operates on nationally standardized, coded EHR data elements. The tool is based on the associative classification rule mining provided in the Waikato Environment for Knowledge Analysis (WEKA)[60] public-domain data-mining software. Our goal is to expand CAT to incorporate additional algorithms from the WEKA software, such as Generalized Sequential Pattern (GSP) algorithm for sequence association mining. As part of this update, the individual WEKA Java packages and classes used in the CAT will be extracted and implemented in a consistent and documented workflow. This organization contributes to simplification of the process and quality assurance of the update and is preparatory to the subsequent aims for the CAT. The deliverable for this aim is evaluation of a use-case for the association of sequential data, such as annual patterns over several years of patients with type 2 diabetes.

Furthermore, as illustrated earlier, the interface with the terminology services should support mapping from "legacy" terminologies to standards-based terminologies and support mapping from one terminology to another standards-based one as needed. The specific use-cases for terminology mapping for the CAT will be determined by surveying the terminologies used in collaborators' EHRs and harmonized with the Phenotype Logic Specification. The CAT terminology mappings are currently input from a simple file format containing mappings. The interface will be modified to load from the program terminology services. Use-cases will be explored to design the user specification of terminologies to load when the CAT association mining process is invoked and other user controls. The interfaces with terminology services will be designed in collaboration with members of Project 1. It will be implemented and validated using software-development quality-assurance processes. Collaborating healthcare data analysts who desire to use the cohort amplification toolset will contribute to the organization, documentation, and testing of software by use and review of initial and modified releases. Thus refined, the toolset will be deployed for public use.

B.3.4 Key Personnel. Christopher Chute, MD, DrPH, Mayo Clinic; Stan Huff, PhD, Susan Rea Welch, Intermountain Healthcare; Rebecca Kush, PhD, CDISC; Jyotishman Pathak, PhD, Mayo Clinic.

B.4 Scaling Capacity to Enable Near–Real-Time Throughputs. Building modules, resources, pipelines, and their local or network instantiations does not always scale in high-demand, high-throughput circumstances. Ensuring that these services can divide clinical streams into coherent tasks, access shared standards repositories, and leverage parallel and cloud computing infrastructures requires a separate level of software engineering.

B.4.1 Specific Aims. 1) Parallelize services to run on clusters; 2) Allow interprocess communication among multiple "clones" of servers; 3) Develop and deploy virtual machine images that can dynamically scale in cloud computing environments.

B.4.2 Previous Work. The IBM Watson UIMA team has been leading the development of the UIMA framework (now in Apache Open Source[61] for many years. The UIMA framework[62] is a standards-based[63]

integration enabler for unstructured information annotator components, supporting modularity, interoperability, multiple implementation languages, component reuse, and flexible scaleout. It has been widely adopted, both commercially and academically. IBM, for instance, includes it in many of its software products and has released tooling to support construction of annotations for these product uses.[64]

The open-source, community nature of UIMA is encouraging the widespread development of re-usable annotation components, including in the medical informatics field[65] There are independent sites that act as repositories, hosting annotators and having the ability to test them on data[66]

UIMA has been recently extended to include an advanced scaleout methodology based on JMS (Java Messaging Services standard) using the Apache ActiveMQ open-source messaging framework.[67] This scaleout enables spreading the annotation work effectively over thousands of CPU cores, both multi-core and/or networked (both wide-area and local). This framework is already being used by many groups around the world. For instance, the grand challenge research project at IBM Research to develop an open-domain question answering system[68] that could effectively compete as a contestant on the Jeopardy! TV program is being developed on top of this framework. A key aspect of this framework is the ability to achieve low-latency, near–real-time response involving the use of thousands of parallelized units of work. The framework enables detailed performance measurements to be collected and aggregated; the IBM Watson/UIMA team has experience using this data to identify bottlenecks and tune large-scale deployments for near–real-time low-latency performance.

UIMA supports modularity via its component metadata, aggregation of components capabilities, and PEAR packaging for redistribution capabilities. It allows components to be written in different languages such as Java or C++ and supports seamless inter-operation of these components. It includes support for arbitrary workflows, allowing arbitrary user-written workflows, and the ability to have particular units of work spawn other units of work, which can be processed in parallel. These units of work can subsequently be "merged" as well. This capability can be used with high scaleout to enable low-latency, near–real-time performance. UIMA supports a very flexible scaleout capability in which annotators can be redeployed to run on networks of computers with no change to the annotators; it includes facilities

needed to tolerate failures and collect information needed to diagnose and tune large-scale deployments.

UIMA pipelines are also being run inside various cloud computing frameworks, including Hadoop. IBM Watson Research has processed multi-gigabytes of corpora using this form of scaleout with extensive UIMA pipelines.

B.4.3 Methods. IBM Watson Research will participate by collaborating with others, including Mayo Clinic, in bringing their considerable experience and expertise in both the UIMA frameworks and tools as well as in applying UIMA to the unstructured information analysis problems. Components for semantic parsing and syntactic transformation will be adapted to run effectively as UIMA components. These will be combined with other UIMA components in the development of robust, locally customizable unstructured/structured information transformation for purposes of normalization. UIMA-AS (asynchronous scaleout) capability will be leveraged in the construction of systems that do near–real-time processing of high volumes of data in the presence of unreliable components. The network support capabilities of UIMA will be explored as a possible approach to supporting local data protection, running data-extraction algorithms in controlled environments on local/private data while taking advantage of more globally shared resources for additional processing. Workflow for the analytics pipeline supporting the various applications will be designed and implemented.

Our work will include all aspects of helping the team best make use of the UIMA framework and will include educating other team members in the use of UIMA, including best practices. We will collaborate on developing the application requirements for scaleout, deployment, and parallelization, and help with the scaled-out platform design. We will collaborate in areas related to workflow, including UIMA flow-controller components and CAS Multipliers. We will explore hybrid approaches that segment some of the computation to isolate sensitive data within organizational boundaries. We will participate in the scaleout deployment design, bring-up, measurements, and tuning, developing and/or applying custom tooling as needed. We will review and help improve critical annotators for normalization and phenotyping in two ways: first, to enable them to best exploit a deployment on multi-core, scaled-out parallel deployment, and second, to mitigate performance impacts of annotators that consume large resources (time and space).

We will adapt existing tooling and develop (as needed) custom approaches for managing the deployment and lifecycle of perhaps hundreds or thousands of machines that are involved in scaleout, targeting both performance and continued operation in the face of inevitable occasional component failure(s); this work will make full use of available existing tooling (open-source, where feasible). Additional integrations between UIMA and cloud computing will be developed as needed to enable the goals of delivery on internet-hosted cloud computing environments. Where needs are revealed by this project, we will work with the open-source UIMA community on improving and extending the UIMA framework itself; these contributions back to UIMA will be in open-source.

B.4.4 Key Personnel. Dave Ferrucci, PhD, IBM Watson Research Labs; Marshall Schor, IBM Watson Research Labs.

B.5 Data Quality. While our normalization and phenotype pipelines can transform and summarize data, knowing whether the data makes sense can be invaluable.

B.5.1 Specific Aims

1) Refine metrics for data consistency – ST
2) Deploy methods for missing or conflicting data resolution – ST
3) Integrate methods into UIMA pipelines - ST
4) Refine and enhance methods - LT

B.5.2 Previous Work. Mayo Clinic conducts about 4,000 record review studies each year. In addition, we conduct more than 1,000 clinical trials, including hundreds of sophisticated cancer treatment trials. Thus, we encounter the problems of incomplete, missing, or conflicting data frequently. There is a large literature base on the imputation of missing data going back at least to Rubin in 1976[69]. Classic methods include Regression methods, models for missingness and, in last 10 to 15 years, multiple imputation approaches that allow not merely the completion of the dataset but the allowance for the uncertainty caused by this process of imputation itself. Our Biostatistics Division, numbering nearly 250 people, has deep experience with these methods and their application. Dr. Kent Bailey is a senior faculty member in

biostatistics who himself has worked in the area of “informative censoring” and will bring relevant methods and techniques to this project.

As part of our phenotyping algorithm development within eMERGE, we had substantial opportunity to address the data quality problem. Simplistically, we seek patterns of laboratory or medication data that are inconsistent with diagnoses or findings. For example, if a single observation of an administrative diagnosis of diabetes mellitus is found but there is no evidence of glucose monitoring, HgA1c, or diabetic medications, then one may conclude that that diagnosis is unlikely to be valid. Conversely, if medications and laboratory studies and findings are typical of a diabetic patient, then even in the absence of an administrative diagnosis, the diagnosis may validly be inferred. We have assembled a library of such data-consistency edits and checks.

B.5.3 Methods. Overall quality metrics will be integrated into the phenotype methods by defining consistency criteria for each phenotype we (or others) create. For example, chronic diseases should appear repeatedly on administrative claims data and should have appropriate medication and laboratory confirmation. In addition, we will create measures of quality as they pertain to data transformation. While not sophisticated, these measures will provide a basis for comparing the confidence of inferences from various sources and environments.

We will develop statistical profiles of: a) malformed data (failing transformation checks), b) non-semantic data (failing vocabulary profiles), c) inconsistent data (failing phenotype specific profiles), and d) conflicting data (lab or medicine characteristics incompatible with diseases, and the presence of negation and assertion for the same elements). These profiles will include frequencies, proportions, and variance measures. While probabilities and significance could be computed (for testing whether a particular item is not spurious), we believe a more robust measure is the likelihood associated with spurious data entry, which can be used in a more flexible way. Thus, we will create statistically based confidence measures that will be reported to the UIMA pipeline, enabling users to dynamically parameterize thresholds for rejection of spurious data.

We will also develop methods for data correction or imputation. For example, if we observe a single instance of an administrative code for diabetes but no supporting medications, laboratories, or subsequent mentions, we would declare the observation unsupported and suppress it. We will pursue rules-based methods and threshold values as the most straightforward approach, but will also explore machine learning techniques using cases and controls validated among many existing Mayo studies managed by biostatistics.

All statistical code will be authored in the R open-source statistical language. Because we seek to integrate these R-based procedures and methods into UIMA pipelines, we will use the rJava[70] linkages, which bridge R functions to Java code. Thus wrapped as Java, it becomes straightforward to integrate these routines into UIMA functions.

B.5.4 Key Personnel. Kent Bailey, PhD, Mayo Clinic; CG Chute, MD DrPH, Mayo Clinic; Stan Huff, Intermountain Healthcare.

B.6 Real-World Evaluation Framework

B.6.1 **Specific Aims.** 1) Integrating normalization and phenotyping algorithms into HIE data flows and NHIN Connect linkages; 2) Cohort identification for translational science protocols; and 3) Dynamic risk factor recognition in HIE data streams.

B.6.2 Previous Work. For more than 40 years, Mayo Clinic has engaged in population-based epidemiologic study using patient records in a secondary context. This effort has been continuously funded by NIH as the Rochester Epidemiology Projects (R01-AR30582)[71,72]. While historically focused on paper records and manual abstraction, the past decade has seen substantial progress in data mining of electronic records. Furthermore, Minnesota requires patient authorization for any use of medical records in research. Mayo Clinic and its regional partners (as part of the Rochester Epidemiology Project) maintain a copy of all Minnesota Research Authorizations, which our Institutional Review Board (IRB) accepts as consent for aggregate data studies.

Both the University of Utah and Mayo Clinic are CTSA award sites and maintain patient data warehouses for translational research application[4]. These resources include comprehensive patient data from multiple EHR-type sources and form an ideal platform for normalization and phenotype pipeline evaluation. Furthermore, both organizations have a large library of carefully phenotyped cohorts and studies, which will serve as the gold-standard comparison to measure the precision and recall of our software processes.

The Minnesota Health Information Exchange is the designated HIE in Minnesota. Currently covering 4 million Minnesotans, incorporating health plans, clinical labs, providers, pharmacy benefit managers, retail pharmacies, the Minnesota Department of Health, and connections to NHIN. The NHIN provides a secure, nationwide, interoperable infrastructure to facilitate the electronic exchange of health information between providers, consumers, and others involved in supporting health and healthcare. The goal of NHIN is to enable health information to follow the consumer, be available for clinical decision making, support appropriate use of healthcare information beyond direct patient care and, ultimately, improve health. In order to foster collaboration in the health community, an open-source framework was created to provide a shared, common interface to NHIN; that gateway is called CONNECT.

For the past 22 months, Agilex Technologies, Inc. has been under contract to the Office of the National Coordinator for Health Information Technology for the development of the NHIN-CONNECT platform. The NHIN-CONNECT platform is being built as the sole gateway to connect a consortium of 26 federal health care organizations (MHS, VA, IHS, CDC, FDA, CMS, SSA, etc.) to the NHIN. The NIHN-CONNECT is an open source, service-oriented architecture platform that, since its development is being guided by the Federal Health Architecture organization, is current with all HITSP standards and guidelines and is currently in production by the Social Security Administration (SSA).

Over the past year, Agilex demonstrated the ability for the NHIN-CONNECT to securely exchange personal health information between an assortment of federal and private healthcare organizations. The use-cases demonstrated included a wounded warrior scenario in which clinical data on the patient was exchanged between the MHS, VA, HIS, and multiple health information organizations (HIOs) that were

in turn connected to regional hospitals and clinics; enabling the SSA to search and collect disability claims data from almost a dozen nationally distributed HIOs and thus reduce a typical claim data collection and processing time frame from 80 to 90 days to less than a minute; syndromic surveillance for the CDC and research data sharing for NCI. The NHIN-CONNECT, while still growing and expanding its range of services, is not simply a successful laboratory experiment. The SSA is operationally using the platform to collect actual disability claims data for beneficiaries in the state of Virginia and has plans to expand to other states over the next year. NHIN Connect allows health status to be verified in minutes instead of weeks and, consequently, the SSA has reduced processing times for disability claims from 84 to 25 days. In addition, as stated earlier, the code is open-source (free to anyone); therefore, Agilex maintains an Internet server with the code and a SDK for all interested parties. Since the code was first published in early April 2009, there have been tens of thousands of downloads.

B.6.3 Methods. Our "in situ" evaluations will work with the clinical data warehouses established at Mayo and Intermountain Healthcare. Both organizations have large portfolios of clinical cohorts and protocol based trials with humanly verified gold-standard cohorts. These resources provide opportunities to evaluation normalization and phenotyping. The Mayo Clinic Enterprise Data Trust (EDT, a data warehouse) is highly normalized[4], using methods and techniques that will contribute to Project 1; this creates a perfect opportunity to compare huge volumes of raw data from EHR and departmental source systems with the gold-standard of the EDT normalized version of the data. We will iteratively test our normalization pipelines, including NLP where appropriate, against these normalized forms, and tabulate discordance. We will sample discordant data for error analyses, establishing whether discordance is attributable to normalization pipeline failures, underlying data quality problems, EDT shortcomings, or variance in the target standard specification version or source. Each iteration will inform pipeline development activities.

Phenotyping validations will be undertaken at Mayo and Utah with established cohorts and trials. Preference will be given to trials with the CTSA framework, to ensure generalizability to this emerging community. For cohort identification efforts, we will routinely conduct two modes of identification: 1)

from our data warehouse resources, and 2) from source EHR and departmental systems. The second mode has the merit of measuring how normalization and phenotype pipeline perform in tandem, establishing whether cascade errors erode functional behavior. Given the volume of cohorts in the institutions, we will perform aggregated precision and recall metrics with a cohort as the unit of analyses. We will then be in a position to quantitate the trade-offs in annotator design and pipeline configurations.

Eligibility criteria for interventional trials are special cases of cohort identification, though they typically lacks the subtlety of identifying controls or, essentially, “not”-cases. It also defines a much broader class, because the phenotype of “being eligible” is typically much broader, conditional on established disease status, than identification as having a specified phenotype required for cohort studies.

For population-health surveillance and evaluation on HIEs, we do not presume that a practical project can be undertaken within the scope of this proposal from either a cost or ethical perspective. Using this effort, we will *demonstrate* surveillance experiments for many conditions, though early examples include a) drug-use surveillance, b) community pneumonia, and c) acute respiratory syndromes. HIE records in Minnesota and, potentially, in Utah will be normalized using our phenotype pipelines and the community infrastructure built for the Rochester Epidemiology Project. The southeast Minnesota community and the state of Utah have applied for BEACON cooperative agreements, which, if awarded, would substantially extend the framework of HIE evaluation of these tools.

For each evaluation, we will validate the identified cohort of patients identified from HIE datastreams for the specified conditions using phenotyping algorithms applied to the patients corresponding EHR and data warehouse information. The within-organization EHR data will constitute the gold standard against which detection or classification of risk factors, public health challenges, or notifiable diseases will be measured.

We will also examine how our services and tools can be integrated into NHIN Connect mechanisms to create a unified framework HIE adoption of normalization and phenotype detection pipelines.

B.6.4 Key Personnel. Christopher Chute, MD, DrPH, Kent Bailey, PhD, Mayo Clinic; Stan Huff, MD, Intermountain Healthcare; Kyle Marchant, Les Westberg, John Page; Agilex. Mike Ubl, Minnesota HIE.

C Plan for Transitioning Appropriate Research Results into Practice

We have a two-fold strategy for ensuring the translation of our efforts into practice, open-source release and encouraging the commercialization of our resources. Mayo informatics has a 20-year history of open-source resource delivery, with two high-profile products that have become highly influential in the informatics community. The first is the LexGrid suite of terminology services resources, including the LexEVS manifestation within NCI's caBIG and the National Center for Biomedical Ontology. The LexGrid suite is the de facto basis for HL7's Common Terminology Specification, and is broadly implemented by academic organizations, governments (UK Cancer Grid), and academic groups. The second major release is the cTAKEs NLP pipeline within the Open Health NLP consortium. The cTAKEs pipeline is the only fully open-source NLP effort in the health domain with broad capabilities across all ranges of clinical text sources.

Mayo has encouraged the commercialization of its open-source software. All open-source software from Mayo has been released under LGPL (Lesser General Public License) which permits the "non-viral" incorporation of its components into commercial software. Indeed, we are very proud that, within the Mayo/Intermountain/GE cooperative working group, the LexGrid suite has been commercially adopted by GE to become their core terminology service for all GE Healthcare products. To stimulate similar actions for our pipelines and services, we have engaged two small vendors, Agilex and Centerphase, and one large vendor, IBM. We encourage these organizations to commercialize our open-source development to increase adoption by the market segment that prefer commercial support options.

D Committee and Stakeholder Involvement

D.1 Program Advisory Committee. The SHARP program governance will ultimately operate under the ONC, with direct guidance and oversight of a program advisory committee (PAC) for research method review, stakeholder relationships, project collaborations, program project monitoring of set performance evaluation criteria and milestones with change control processes, implementation plans, and research outcome reports. Members of the PAC are not funded by the grant; they include:

- Suzanne Bakken, RN, DNSc, The Alumni Professor of Nursing and Professor of Biomedical Informatics. Principal Investigator, Center for Evidence-based Practice in the Underserved, Columbia University
- David Hardison, Vice President of Science Applications International Corporation.
- Issac Kohane, MD PhD, Director, Children's Hospital Informatics Program and Professor of Pediatrics and Health Sciences and Technology. Harvard Medical School (HMS). Director, HMS Countway Library of Medicine. Co-director, HMS Center for Biomedical Informatics.
- Barbara Koenig, PhD, Professor of Biomedical Ethics, Mayo Clinic College of Medicine. Professor Faculty Associate at the Center for Bioethics, University of Minnesota.
- Marty LaVenture, PhD, MPH, Director, Office of Public Health Informatics, Office for Health Information Technology and Center for Health. Minnesota Department of Health
- Dan Masys, M.D., Chair, Department of Biomedical Informatics, Vanderbilt University
- Mark A. Musen, M.D., Ph.D, Professor of Medicine (Biomedical Informatics); Division Head Stanford Center for Biomedical Informatics Research; Co-Director, Biomedical Informatics Training Program. Stanford University.
- Robert A. Rizza, M.D., Executive Dean for Research, Mayo Clinic Division of Endocrinology, Diabetes, Metabolism, and Nutrition. Earl & Annette R. McDonough Prof. of Medicine.
- Nina Schwenk, M.D., Vice President, Mayo Clinic Board of Trustees; Vice Chair, Mayo Clinic Board of Governors; Consultant General Internal Medicine, Mayo Clinic.
- Kent Spackman, M.D., Ph.D., Chief Terminologist, International Health Terminology Standards Development Organization (IHTSDO)
- Tevfik Bedirhan Üstün, M.D., Coordinator of Clinical Classifications, WHO, Geneva, Switzerland

D.2 Stakeholder Analysis & Engagement. This program entails a matrix of stakeholders; the below table illustrates each stakeholder's role or engagement and that role’s level of impact on and influence on the program. The stakeholder network engagement, communications, and coordination hub will be

managed by the SHARP project manager with support from the Biomedical Statistics and Informatics Project Management Office (BSI PMO).

The program will employ Deloitte Consulting for expertise in consortium stakeholder engagement, which may include visioning, consensus building, group facilitation, summit planning, and implementation. Each year, Deloitte would facilitate a summit for consortium members and potential external stakeholders to share ideas and have breakout groups to provide recognition and accountability for individual performance and a forum for discussion.

Stakeholder	Role or Engagement	Impact	Influence
Department of Health and Human Services		MED	HIGH
Office of the National Coordinator for Health Information Technology (ONC)	Performance evaluation	MED	HIGH
Project's Federal Steering Committee (FSC)	Performance evaluation	MED	HIGH
SHARP Program Advisory Committee (PAC)	Research direction, program monitoring, performance evaluation	HIGH	HIGH
Agilex	Transformation	MED	MED
CDISC (Clinical Data Interchange Standards Consortium)	Phenotyping, Standards	MED	MED
Centerphase	Phenotyping	MED	MED
Group Health, Seattle	NLP	MED	MED
IBM – Watson Research Labs (UIMIA)	UIMA	HIGH	MED
Intermountain Healthcare	NLP; Phenotyping	HIGH	HIGH
Mayo Clinic	Coordinating Center; all	HIGH	HIGH
MN Health Information Exchange	Evaluation, testing	MED	MED
MIT	NLP	HIGH	MED
SUNY, Albany	NLP	HIGH	MED
University of Colorado	NLP	MED	MED
University of Pittsburgh	NLP	MED	MED

E Project Management

E.1 Key Personnel. Lacey Hart, MBA, PMP®; Mayo Clinic; Deloitte Consulting

E.2 Program Project Management Approach and Coordination. Mayo Clinic's standardized best practices are based on those identified by the Project Management Institute. As defined by PMI, we will

use the tools in the *Project Management Body of Knowledge Guide*[73] to initiate, plan, execute, monitor, control, and close projects in the SHARP program. The program will be under the direction of a Mayo Clinic PMI-certified project manager, who will use best practices to coordinate daily research activities, including subcontract relations and tasks, ensuring that projects are delivered on time and within budget. The project manager will operate under the PAC and the BSI PMO, which currently oversees project management resources for a $140 million portfolio of grants, contracts, and intramural responsibilities. The BSI PMO is backed by the Mayo Clinic Enterprise Portfolio Office, leveraging the expertise of the staff supporting project portfolio best practices and enterprise-wide standardization.

E.3 Communication Management Plan. The project manager will proactively ensure effective communications for this program. This communications-management plan will set the communications framework for this program and will be updated post-award as well as throughout the program as needs evolve. The project manager will be the communications hub for program timelines and calendars, meeting organization, work group formation facilitation, committee meeting scheduling and tracking, and program documentation. The program will rely on the following communication mechanisms to ensure smooth progress, timely completion, and appropriate dissemination of work:

- Annual face-to-face summit for all members of the consortium
- Video and audio teleconferences with Internet and dial-in connections
- Wiki /Web portal tools used to house project-management documentation, track discussion, and minutes of all meetings and conference calls, provide a forum for risk- and issue-management, and function as a repository for annotated data, guidelines, and output.
- Establishing a shared code repository

This plan includes a communications matrix that maps the communication requirements of this program noted during pre-award planning.

Stakeholder(s)	Reason for Communication	Communication Method	Frequency
ONC, FSC, PAC	Formal Performance Assessment as per RFA	Formal Written Report	Months 12, 24, & 36

ONC, FSC	Project progress review	Teleconference with written documentation	Quarterly
PAC	Project progress, methods review & input opportunity	Teleconference with written documentation	Semiannually
All Stakeholders	Stakeholder engagement	Face-to-face summit	Annually
All Stakeholders	Kickoff meeting: Stakeholder engagement & project planning	Video/teleconference	Once
Deloitte	Stakeholder & program management consultation	Teleconference	Quarterly
BSI PMO	Program/project status report & cost management	PMO Status Report	Monthly
Data Normalization Project Team (s)	Status meetings: Task, risk and issue management	Teleconference	Weekly as needed; Monthly check
Phenotype Project Team (s)	Status meetings: Task, risk and issue management	Teleconference	Weekly as needed; Monthly check
Data Quality Project Team (s)	Status meetings: Task, risk and issue management	Teleconference	Weekly as needed; Monthly check

E.4 Change Management and Conflict-Resolution Plan. Our change-control approach is to provide standard terminology, clear roles and responsibilities, a detailed description of the approved change-control process, and standard templates used in that process, overseen by our project managers. It is designed to guide the program team, stakeholders, and the PAC. The objectives of the change-control process are to ensure that changes to the project have a strong research justification, monitor and control the cost and other effects of approved changes, obtain the appropriate level of approval for changes by the PAC, ensure that changes are understood and that team members do not begin work on new or unplanned tasks before change requests are approved, and maintain a record of changes.

E.5 Risk- and Issue-Management Plan. While identification and mitigation of risk is ongoing, the following risks have been identified. Risk mitigation is included with each category. The final SHARP program-management plan will refine the risk-management method as details emerge, including risk identification, risk analysis, risk response, and risk tracking and reporting. Note that the first three steps

happen in sequence, and the final step, risk tracking and reporting, occurs throughout. Risks are assessed according to probability, impact, timeline, and response activities.

Risk Description	Likelihood	Impact	Timeline	Mitigation
Project Scope Creep	Unlikely	Med	Near-Mid term	Scope initially defined in project plan, reviewed monthly to prevent undetected scope creep
Consultant Project Deliverables Unclear	Unlikely	Med	Near-Term	Included in project plan, subject to amendment
Timeline Estimates Unrealistic	Somewhat likely	Med	Mid-far term	Timeline reviewed monthly by three groups (Project Manager and Steering Committee) to prevent undetected timeline departures
Absence of Commitment Level/Attitude of PAC	Unlikely	Low	Near-term	Project manager frequently seeks feedback to ensure continued support
Project Team Availability	Somewhat likely	Med	Near-mid term	Continuous review of project momentum by all levels. Project manager to identify any impacts caused by unavailability. If necessary, increase commitment by participants to full time status
Physical Location Dispersement of Team Hinders management	Likely	Low	Near-term	Use of Intranet project website, comprehensive communications plan
Change Management Procedures undefined	Unlikely	Low	Near-term	Use of comprehensive project Management plan
Quality Management Procedures unclear	Unlikely	Low	Near-term	Use of comprehensive project-management plan

Our issue-management approach is to track current happenings that are having a negative effect on the SHARP program and require resolution. The process of identifying, analyzing, responding to, and tracking issues will be assessed weekly. The objectives of our issue-management approach are to monitor and identify new or changing issues, understand and minimize cost and other impacts of project issues, focus attention on high-priority issues, make issue-related decisions at the proper level of authority, communicate clearly about issues with the PAC and stakeholders, and maintain a record of issues and related actions over the grant project(s). These steps include issue identification, analysis, response, and tracking and reporting.

E.5 Cost-Management Plan. The cost-management plan identifies the processes and procedures used to manage costs throughout the program. A more detailed plan, to be completed post-award, will cover the

cost-management approach, expenditure tracking, variance analysis, oversight of contractor costs, and reconciliation of the budget, accounting, and project-management cost processes.

The principal investigator will approve the cost-management plan, is ultimately responsible for the allocation and expenditure of the program budget, and has the authority to make changes to bring it back within budget. The project manager manages and reviews project costs from a project-management perspective to ensure appropriate progress is being made for the funds being expended. The project manager works with the finance staff to reconcile the cost-management data to the current accounting data. Subcontractors identify funding needs and assist with tracking expenditures, including tracking staff effort and costs. Expenditure reports are generated for project use in validation and tracking of expenditures against the budget. The Mayo Clinic Accounting Office sends reports to the sponsor.

Cost reporting and metrics include: 1) Spending plan (by fiscal year) and a run chart showing the actual costs against the baseline by month for the fiscal year and the cumulative total to date for the fiscal year. 2) Labor hours by deliverables in a plan, a chart showing the effort expended towards the plan by month for the fiscal year and the total to date for the fiscal year.

F Evaluation

F.1 Success Factors. We judge the ultimate success of our efforts will be the wide-scale adoption of our pipelines and services by the academic and commercial communities engaged in the secondary use of EHR data. Interim successes include effective evaluations and demonstrations in Project 6, applied to the real world. Technical successes include seamless integration of services and pipelines, and their ability to scale in size and scope for the challenges confronted by the secondary user community. Finally, user satisfaction with the tools and parameterization capabilities must be achieved to ensure adoption.

F.2 Evaluation Processes (Quality-Management Plan). Program project quality management is conducted continually. Under the project manager, assessments will be conducted to determine whether the project's processes conform to the plans, if they are being executed as defined, and if they are being effectively implemented and maintained. Process quality assessment also includes an audit of project-

management plans. The program will identify, collect, analyze, and report on metrics throughout the program. The selection of metrics and data items will evolve to focus on specific areas. The use of metrics within a project is expected to have a beneficial effect by making quality (or lack of quality) visible. Quality metrics will include but not be limited to the following:

- Process quality
- Schedule and progress
- Resource and cost
- Process performance
- Data normalization & phenotyping quality
- Conformance to requirements
- Technology effectiveness
- Customer satisfaction

The program establishes quality-improvement strategies on the basis of the value of each improvement with respect to the objectives. Improvement strategies are determined on the basis of such measures as time to recover the cost, improvement in project performance, and the program's ability to respond to the changes. Selected quality improvements will be managed and controlled through the project's established change-control process governed by the PAC.

The responsibility for testing the deliverables is the responsibility of each subcontractor. Project staff verify and validate the acceptability of the delivered system or component, provide oversight, and participate in testing activities, including system and acceptance testing.

Financial audit(s) will be conducted annually to ensure that general grant receipts and expenditures associated with the program are properly recorded. In addition, technical review or compliance review(s) will focus on program management and determine if grant terms and conditions are being met. Significant findings will be escalated through the PAC and the ONC, with a formal response of followup and resolution processes.

F.3 Milestones

F.3.1 Milestones by term

Initial Planning Conference (IPC)	Month #1
Performance Assessment #1 & Fiscal Year Cost Reporting metrics	Month #12
Stakeholder Summit	Year 1
Performance Assessment #2 & Fiscal Year Cost Reporting metrics	Month #24
Stakeholder Summit	Year 2
Algorithms and phenotype library	Year 2 Qtr 4
Begin scaling capacity to enable near real time throughputs	Year 2 Qtr 4
Performance Assessment #3 & Fiscal Year Cost Reporting metrics	Month #36
Stakeholder Summit	Year 3
Begin HIE and NHIN deployment & evaluation	Year 3 Qtr 4
Globally available health terminology & value set resource	Year 4 Qtr 3
Modular library of normalization algorithms	Year 4 Qtr 3
NLP Gold Standards	Year 4 Qtr 3
Deploy methods for data quality & consistency and conflict resolution	Year 4 Qtr 4

G Organization and Capability Statement

G.1 Agilex Technologies. Agilex Technologies is an employee-owned application, solution, and advisory firm incorporated in March 2007 to provide “new school” and agile problem-solving approaches to challenging problem domains. Although a small company, Agilex generated $40 million in revenue after only 2 years of operation. Agilex has attracted noteworthy people such as a renowned expert in Oracle technology and founding member of Oracle’s System Performance Group; the architect and designers of the world’s largest, globally deployed electronic healthcare management system (DoD Military Health System [AHLTA]); and the former CIO for one of the world’s largest technology networks and the world’s largest intranet (USPS). Agilex is leading the architecture and development efforts for NHIN Connect and is deploying NHIN compliant adaptors for 26 government agencies, including the DoD and SSA. We are also acting in an advisory role with both the Office of the National Coordinator and the DoD for supporting the increased exchange of data between the DoD and the VA.

G.2 CDISC. The Clinical Data Interchange Standards Consortium is a global, open, multidisciplinary, non-profit organization that has established standards to support the acquisition, exchange, submission and archive of clinical research data and metadata**.** The CDISC mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. CDISC standards are vendor-neutral, platform-independent and freely available via the CDISC website. CDISC was founded in 1997 and has since grown to more than 250 organizational members around the world, with an annual budget more than $3 million. CDISC was the lead organization in the formation of the BRIDG consortium, a domain-analysis model addressing protocol driven research in partnership with HL7, FDA, and NCI. CDISC's president, Beck Kush, co-chaired the ONC and ANSI working group that established clinical research as a peer use-case on the HITSP agenda under HHS/AHIC.

G.3 Centerphase Solutions. Centerphase is a new company created jointly with Mayo Clinic as the first "independent centralized clinical trials office." The initial goal is to partner with academic medical centers to create greater efficiency in clinical trial execution. Over time, the platform will be leveraged for a variety of high-value services, including adherence/compliance programs, safety surveillance, comparison of treatment effectiveness, and health outcomes.

G.4 Group Health, Seattle. Group Health Research Institute (GHRI) is affiliated with a learning healthcare system, Group Health (GH), and conducts multisite epidemiological, health services, comparative effectiveness, and GWAS studies. Structured enrollee data begin in the 1970s; electronic pathology, radiology, and chart text in 1980, 1998, and 2003, respectively, and are extracted from an Epic EMR for processing by a UIMA/cTAKES. GHRI has experience with open-source NLP systems, including caTIES, GATE, MITRE-MIST, ONYX, and i2b2. GHRI plays key leadership roles in the HMO Research Network (HMORN) and the HMO Cancer Research Network (CRN); multisite, open-source collaboration is central to our identity and mission. Working with NLP informaticists from Mayo Clinic, Pittsburgh, and Vanderbilt since 2008, Dr. Carrell has led NLP strategic planning/adoption efforts in the

HMORN. Through an NCI-funded project to apply NLP to cancer surveillance in the CRN (Carrell, PI), GHRI is piloting a could-based (Amazon AWS) cTAKES system and negotiating with Deloitte for risk management and auditing services.

G.5 IBM Watson Research Labs. IBM Research is the largest industrial research organization in the world, with eight labs in six countries. IBM Research staff members have received extraordinary awards, including five Nobel prizes, seven national medals of technology, five national medals of science, and six Turing awards. Significant innovations have come from IBM Research, including FORTRAN, one-transistor memory cell, fractals, relational databases, speech recognition, RISC architecture, silicon germanium chips, Deep Blue, Silicon-on-Insulator, and Blue Gene/L. Our portion of the research will be carried on at the IBM Thomas J. Watson Research Center, the headquarters for IBM Research in Westchester County, New York. The facilities include thousands of networked server computing nodes, a cloud computing infrastructure, multiple Blue Gene supercomputers[74], and high-capacity networking interconnects. IBM's recent grand-challenge effort around open-domain question answering[68] are using these facilities to advance the emerging infrastructure of near–real-time, massively parallel computation. This work, in turn, provides a significant platform for advances in the parallelization and high performance of the UIMA platform being used in the proposal.

G.6 Intermountain Healthcare is a not-for-profit healthcare organization founded in 1975. Intermountain has a legacy of excellence in building and using clinical information systems. It is recognized as a pioneer in developing innovative IT applications to improve the quality of care by improving access to information, streamlining and standardizing processes, and providing timely decision support to caregivers at the point of care. The HELP hospital information system was operational in 1967 at LDS Hospital and continues to be used heavily in numerous inpatient environments. In recent years, Intermountain has built an ambulatory, Web-based EHR, HELP2, to complement this system. This environment is ideal for applied informatics research. The Department of Medical Informatics at Intermountain has designed and implemented a variety of clinical information systems for both research

and care delivery. The Homer Warner Center for Informatics Research (HWCIR) maintains a hardware and software infrastructure within Intermountain's clinical network dedicated to the support of clinical and informatics research activities. We also maintain a nationally recognized enterprise data warehouse.

G.7 Mayo Clinic. Mayo Clinic is the first and largest integrated, not-for-profit group practice in the world. Doctors from every medical specialty work together to care for patients, joined by common systems and a philosophy that the needs of the patient come first. More than 3,600 physicians and scientists and 50,000 allied staff work at Mayo, which has sites in Rochester, Minn., Jacksonville, Fla., and Scottsdale/Phoenix, Ariz. Mayo Clinic also serves more than 70 communities in the upper Midwest through the Mayo Health System. Collectively, these locations treat more than 3 million people each year.

Information Technology at Mayo is staffed by 1,200 analysts, programmers, and managers, overseeing enterprise-wide EHR systems (Cerner and GE/IDX), departmental systems. Mayo supports five major data centers in three states, about 7,600 data or application servers, and 182,000 connected devices, including 53,000 workstations, and 8Pb of online storage.

The informatics research group within the Division of Biomedical Statistics and Informatics comprises 13 doctoral faculty, 8 master level assistants, and 25 analyst programmers with deep experience in ontologies, terminology services, natural language processing, and data standards. Biostatistics support is provided by 25 PhD statisticians and 55 statisticians with master's degrees, whose activities are facilitated by 75 statisticians with bachelor's degrees ("statistical programmer analysts") and 20 clerical personnel. In addition to general consulting on more than 2,000 ongoing investigations, the statistical group provides core statistical support for a number of program projects as well as for the Rochester Epidemiology Project.

Project management within the Division is provided by five project managers assisted by four administrative assistants and deep institutional support from research services, research accounting, and enterprise project management.

G.8 Massachusetts Institute of Technology (MIT). Peter Szolovits is professor of computer science and engineering in the MIT Department of Electrical Engineering and Computer Science, professor of health sciences and technology in the Harvard/MIT Division of Health Sciences and Technology , and head of the Clinical Decision-Making Group within the MIT Computer Science and Artificial Intelligence Laboratory. His research centers on the application of artificial intelligence methods to problems of medical decision making and the design of information systems for healthcare institutions and patients.

Dr. Szolovits plans to develop novel methods and tools to support the translation of significant portions of unstructured medical texts into formalisms that encode the clinical entities that are described or mentioned in the text as well as the temporal and semantic relations stated to hold between them. In his 35-year career at MIT, he had concentrated his research on knowledge representation and inference in medicine, probabilistic and temporal reasoning about healthcare data, architectures for sharing health information, and privacy-preserving systems that require techniques for d-identifying clinical data to make them safely usable for secondary research. This last task, when applied to unstructured textual data, has re-involved him in the past decade in NLP techniques.

G.9 MN Health Information Exchange. MN HIE is a state-wide secure electronic network designed to share clinical and administrative data among providers in Minnesota and bordering states. MN HIE is a not-for-profit organization with a public/private governance model that includes providers, health plans, and the State of Minnesota. MN HIE's purpose is to improve the health of all Minnesotans through more informed decision-making by the provider and patient at the point of care. More than 4 million Minnesota residents are included in MNHIE's secure patient directory. Medication history is implemented, with patient eligibility, lab results, immunization history, and exchange of medical record information began in the fall of 2009.

G.10 SUNY and i2b2. Dr. Uzuner's lab within the Information Studies department of College of Computing and Information in University at Albany, SUNY conducts research on natural language processing applied to clinical records. This lab has teamed up with colleagues from MIT, Harvard

Medical School, Brigham and Women's Hospital, and Partners Healthcare in order to create phenotype-extraction systems that can benefit the biomedical community. Their methods improve automatic information extraction from clinical records and, in their conclusion, these methods are applied for automatic annotation of clinical records, which are then released to the research community for further research. Dr. Uzuner has affiliated appointments in both the computer science department of SUNY and at the Computer Science and Artificial Intelligence Laboratory at MIT. She has a history of supervising students in computer science departments of both SUNY and MIT. The expertise of the Clinical Decision Making Group at MIT coupled with her research direction has so far resulted in work on various aspects of information extraction from clinical records. Example projects include coreference resolution and question answering.

G.11 University of Colorado. The Center for Computational Language and EducAtion Research (CLEAR) at the University of Colorado at Boulder is dedicated to advancing human language technology and applying it to diverse application areas. With funding from NSF, NIH, the IES/Dept of Education, and DARPA, center investigators have established unique collaborations with other leaders in NLP, including machine translation (ISI, Rochester, Brandeis), question answering (Columbia, Stanford), medical (Mayo) and bioinformatics (Colorado Health Sciences), and personalized learning (Denver Public Schools). Professors Palmer and Martin codirect the center; Professor Ward is a founder and former director of the center. CLEAR conducts research and development that informs theoretical questions in NLP and leads to effective and scalable solutions. A major focus is the development of increasingly rich linguistic annotation schemes that can serve as training and evaluation data for machine learning, information extraction, and natural language understanding using semantic role labeling and coreference resolution. These data have provided a springboard for the development of the ClearTK framework, which provides infrastructure for developing UIMA-based natural language analysis engines that use statistical learning as a foundation for decision making and annotation creation.

G.12 University of Pittsburgh. The Biomedical Language Understanding (BLU) Lab in the University of Pittsburgh Department of Biomedical Informatics is led by Dr. Chapman and comprises three post-doctoral associates, a research associate, two programmers, and one PhD student. Our work focuses on extraction of information from clinical reports and has been applied to biosurveillance, clinical research, and real-time charting in the medical and dental domains. Our research includes identifying contextual properties of findings, such as negation and temporality, efficiently creating training sets for learning, developing classifiers from NLP-extracted features, and methods for generating annotated corpora. We are releasing the first publicly available corpus of d-identified clinical reports for NLP research and are collaborating with researchers from the Veteran's Administration, Mayo Clinic, the University of Utah, and other institutions in developing, evaluating, and applying information extraction from clinical reports.

G.13 University of Utah. Utah has one of the oldest departments of medical informatics in the world, having been established more than 50 years ago by Homer Warner and pioneering efforts in EHR development and the HELP system. Today, Dr. Peter Haug directs the Homer Warner Center foInformatics Research (HWCIR), which has deep expertise in decision support and NLP. There has been long-standing collaboration with Dr. Chapman at the University of Pittsburgh in the NLP domain.