Published in final edited form as:
J Biomed Inform. 2010 June ; 43(3): 451–467. doi:10.1016/j.jbi.2009.12.004.

Formal Representations of Eligibility Criteria: A Literature Review

Chunhua Weng[1], **Samson W. Tu**[2], **Ida Sim**[3], and **Rachel Richesson**[4]

[1]Department of Biomedical Informatics, Columbia University

[2]Stanford Center for Biomedical Informatics Research, Stanford University

[3]Division of General Internal Medicine, Department of Medicine, UCSF

[4]Department of Pediatrics, University of South Florida College of Medicine

Abstract

Standards-based, computable knowledge representations for eligibility criteria are increasingly needed to provide computer-based decision support for automated research participant screening, clinical evidence application, and clinical research knowledge management. We surveyed the literature and identified five aspects of eligibility criteria knowledge representations that contribute to the various research and clinical applications: the intended use of computable eligibility criteria, the classification of eligibility criteria, the expression language for representing eligibility rules, the encoding of eligibility concepts, and the modeling of patient data. We consider three of them (expression language, codification of eligibility concepts, and patient data modeling), to be essential constructs of a formal knowledge representation for eligibility criteria. The requirements for each of the three knowledge constructs vary for different use cases, which therefore should inform the development and choice of the constructs toward cost-effective knowledge representation efforts. We discuss the implications of our findings for standardization efforts toward sharable knowledge representation of eligibility criteria.

Keywords

Knowledge Representation; Clinical Research; Clinical Decision Support Systems; Electronic Health Records; Eligibility Criteria

1. Introduction

In clinical research, eligibility criteria are specifications of the clinical and other characteristics of patients for whom a research protocol might be applicable. According to the definition from ClinicalTrials.gov (http://clinicaltrials.gov/), eligibility criteria for clinical trials are “the medical or social standards determining whether a person may or may not be allowed to enter a clinical trial; they are based on such factors as age, gender, the type and stage of a disease, previous treatment history, and other medical conditions.” Similarly, condition criteria are one of the key representation primitives in computer-based clinical practice guideline models [1].

Corresponding Author: Chunhua Weng, PhD, Department of Biomedical Informatics, VC-5, 622 W 168 Street, New York, NY 10032, cw2384@columbia.edu, Phone: 212-305-3317, Tax: 212-305-3302.

Since the requirements and issues surrounding the representation of clinical research eligibility criteria are virtually the same as those encountered in modeling condition criteria for clinical practice guidelines, both are referred to as eligibility criteria from this point on in this paper.

Eligibility criteria are usually written in free text to be human-readable; therefore, they are not amenable for computational processing. As electronic health record (EHR) systems become broadly adopted, standards-based, computable knowledge representations for eligibility criteria are increasingly needed to convey unambiguous logical statements to provide decision support for various research tasks, such as matching eligible patients to clinical trials, matching patients to applicable clinical evidence, and reuse of eligibility criteria from related clinical research studies on similar patients during the design of a new study. Appropriate uses of standards and computational formalisms are the two important desiderata of computer-based eligibility criteria, where computability can be achieved at three levels: the syntactic, semantic, and knowledge level [2]. Despite advancements in computable representations of clinical guideline and electronic clinical trials management systems (CTMS), knowledge representations for eligibility criteria generally have not been robust enough to be executable by computer [3], nor have they been represented in a consistently generic manner [4]. There is no consistent practice of "knowledge representation for eligibility criteria" [5]. In fact, knowledge representation has been used to indicate computer-based classifications of eligibility criteria content, standardization of the syntactic forms of eligibility criteria, or standardization of eligibility criteria terminologies. Existing knowledge representations for eligibility criteria vary in their underlying conceptualizations of eligibility criteria, and in their uncoordinated uses of standards to represent medical concepts and uses of expression languages to model eligibility rules. To our knowledge, there is no comprehensive review that summarizes the state-of-art in this domain. Meanwhile, a number of groups are developing knowledge representations for eligibility criteria, including the Eligibility Rule Grammar and Ontology (ERGO) [6], the Clinical Research Filtered Query (CRFQ)[7], and the Agreement on Standardized Protocol Inclusion Requirements for Eligibility (ASPIRE)[8].

In this paper, we survey the literature and develop a conceptual framework of five dimensions that support a structured comparison of various knowledge representations for eligibility criteria. We also discuss the implications of these results for ongoing and future standardization efforts toward generalizable knowledge representation for clinical research eligibility criteria.

2. Methods

We used the PubMed Central and Google search engines to identify relevant publications published by August 2008. Since formal eligibility criteria representations were primarily embedded in computer-based clinical practice guidelines or clinical trial protocols, we used the following parenthesized phrases one a time to conduct the search and aggregate all relevant manuscripts: ("ontology" AND "eligibility criteria"), ("automated patient screening"), ("computer-interpretable clinical guidelines"), ("matching patients to clinical trials"), and ("knowledge representation" AND "eligibility criteria"). We also reviewed the references of each included manuscript to identify related work missed by the above queries.

3. Results

We retrieved publications for a total of 27 models or systems with computer-based eligibility criteria knowledge representations. Some models and systems were described by multiple publications over time; in those cases, we reviewed all the publications. We analyzed this body of work from five perspectives: (1) the use case of eligibility criteria knowledge representation; (2) the conceptual classification of eligibility criteria; (3) the choice of expression and query languages; (4) the encoding of medical concepts; and (5) the modeling of patient data. We also

reviewed the domain specificity and uptake status of the identified systems. Expression and query languages model the logic of the relationships between the medical concepts used in eligibility criteria, while patient information models support inference using various medical concepts in reference to typical patient information representations in EHR. The details of these 27 models are summarized in Appendix Table 1, and summarized below.

3.1 Use Cases for Eligibility Criteria Knowledge Representations

Computer-based knowledge representations of eligibility criteria have been designed to support three major use cases, described below, that support tasks in both clinical research and clinical care delivery.

Use Case 1: Eligibility determination—finding clinical trials for which a patient is eligible [7–28], or identifying a cohort that is eligible for a clinical trial. Example systems include OncoDoc [10,13], OncoLink [17], EligWriter [19], ASPIRE [8], caMatch [18], T-Helper [26], EON [29], AIDS2 [20],, OaSIS [30], PROforma [31,32], Asbru [33,34], GLIF [35], SAGE [36], GUIDE [37,38], and PRODIGY [39]. Many of these systems were developed prior to the age of EHR. In recent years, with the broad adoption of EHR systems, formal eligibility criteria representations have become a crucial component to data-driven clinical decision support solutions, and promise significant value for clinical and public health research. An example is automated EHR-based clinical research subjects prescreening. For large multi-site clinical trials, eligibility criteria preferably should be encoded once in a formal representation that is interoperable with different EHR systems at multiple sites. Also, many surveillance systems use locally hard-coded criteria to select populations and to allow the monitoring of important health measurements for selected populations.

Use Case 2: Applicability determination—determining when clinical evidence is contributory to a patient's clinical care or identifying a set of clinical practice guidelines for which a patient is eligible [26,29–40], based on the patient's clinical situation. Much of the evidence generated from clinical trials research is underused because it is not amenable for computer-based retrieval and processing. Sim has developed The TrialBank[41] to improve the accessibility and computability of clinical trials evidence modeling at fine granularity levels.

Use Case 3: Knowledge management of clinical research—providing structured representations for eligibility criteria to optimize *automatic classification, retrieval, search, and reuse* of eligibility criteria across clinical research studies. There is a compelling need to query and re-use eligibility criteria in order to accelerate the development of new clinical research protocols and related clinical research documents (e.g., case report forms, data collection forms, training materials, etc). Related effort include EligWriter [19] and Design-a-Trial [42] that supported the reuse of eligibility criteria during clinical trial protocol authoring, as well as ERGO [6] and ASPIRE [8] that supports eligibility criteria annotation.

Essentially, use cases 1 and 2 are similar in that they both match patient information to clinical research protocol or evidence, while use case 3 is focused on supporting interoperability among different clinical research studies. Use cases 1 and 2 can both operate in two modes: (a) patient-driven mode (rank or filter clinical protocols one patient a time) or (b) protocol-driven mode (rank or filter patients one protocol a time) [20]. In mode (a), individual patient data can be acquired from manual data entry or extracted from EHR systems; in mode (b), patient data are usually extracted from an EHR or Personal Health Records (PHR) system and automatically matched to computer-based eligibility criteria. The use of EHR or PHR systems requires semantic matching between representations of eligibility criteria and representations of patient data in EHR or PHR systems, which is an active research area with challenges such as the

semantic gap between eligibility criteria and patient data as first identified by Chute [43], and the need for complex aggregation and inference over both structured and free-text patient data [44].

Some knowledge representations have been designed to serve multiple use cases. For example, ASPIRE and EligWriter are both designed to support use cases 1 and 3. To date, use case 1 is the predominant (87%) use driving the formal knowledge representation of eligibility criteria. Among the systems designed to support use case1, Fink [15] and Ohno-Machado [20] generated recommendations for additional data gathering tests to minimize uncertainty related to patient eligibility status based on the descending order of associated test costs.

3.2 Classification of Eligibility Criteria

We observed great variation in approaches to characterizing eligibility criteria across different eligibility criteria knowledge representations. The space of existing classifications of eligibility criteria can largely be divided along three dimensions: *content, use in eligibility determination, and complexities of semantic patterns of eligibility criteria*.

The first dimension for classifying eligibility criteria deals with the content - or the information needed to answer eligibility queries. Along this dimension, subcategories of eligibility criteria include intent, main clinical category, and main medical topic [5]. By intent, criteria can be divided into inclusion criteria and exclusion criteria. By main clinical category, criteria can be categorized in many ways – typically by demographics, clinical findings, medical history, allergies, procedural or surgical history, behavioral characteristics, laboratory data, device data, vitals, prior or concomitant medications, and administrative and informed consent issues. By main medical topic, eligibility criteria can be separated into disease areas, such as cardiovascular diseases, diabetes, cancer, and so on. Within each medical topic, criteria can be further classified by finer clinical details specific to the topic. For instance, in the United Kingdom's CancerGrid project [45], criteria for breast cancer are be further classified by tumor size, stage, receptor status, and so on.

Content-oriented classifications of eligibility criteria have been popular and motivated by several different conceptualizations. Van Spall et al. recently examined eligibility criteria from a random sample of randomized controlled clinical trials published in high impact journals, and found variability in the content and nature of exclusion criteria across studies [46]. Their interest was to look at the clinical, scientific, or ethical justification for exclusion criteria. They characterized the nature and constructs of different eligibility criteria across the protocols that they sampled. This classification included age, sex, sex-specific conditions such as menstruation, pregnancy, or lactation, race, ethnicity, religious background, language ability, educational background, socioeconomic status, cognitive ability, physical ability or disability, chronic health condition, and condition under investigation. Sim presented a summary of 1000 eligibility rules randomly sampled from ClinicalTrials.gov [5] and found that the sampled rules fell into three high-level constructs: 46% medical histories, diagnoses, or conditions; 36% treatments; and 25% tests or procedures performed and results. Metz et al. classified eligibility criteria representations into demographics, contact information, personal medical history, cancer diagnosis, and treatments to date [17]. In the Trial Bank Project, Sim classified criteria into age-gender-rules, ethnicity-language-rules, or clinical rules [41]. The ASPIRE project differentiates "pan-disease criteria" (e.g., age, demographics, functional status, pregnancy, functional status, etc.) from disease-specific criteria (as of 2008, only in the domain of breast cancer and diabetes) [8]. The caMatch project is primarily focused on eligibility criteria representations for breast cancer [18]. Rubin et al. classified eligibility criteria by clinical states to ensure that research protocols for patients at similar clinical states would have similar eligibility criteria and to reduce the total number of criteria needed to author several clinical protocols [27]. They developed 24 categories for classifying the eligibility criteria from NCI's

Physician Data Query (PDQ) database and found great redundancy in protocols for similar clinical states [47]. Seroussi et al. also classified eligibility criteria according to patient clinical states in the OncoDoc system [10].

A second dimension for characterizing eligibility criteria relates to eligibility criteria properties that are useful for optimizing eligibility determination. Such classifications often aim to reduce uncertainty in eligibility or applicability, to minimize test costs, or to reduce risks for patients involved in eligibility screening. In the AIDS2 system, Ohno-Machado grouped criteria in three broad categories: “history”, “examination”, and “tests”. For “tests”, she classified eligibility criteria by their importance in determining eligibility status for the protocol, risks imposed on patients, and cost (including cost of objective tests and clinicians’ time to assess different criteria) [20]. In the T-Helper system, Tu viewed a participant’s eligibility for a clinical protocol as a dynamic property [48] and differentiated eligibility criteria by their objectiveness, variability, and controllability of the underlying clinical conditions. Based on this rationale, criteria were organized into five groups: (1) stable requisite; (2) variable routine; (3) controllable; (4) subjective; and (5) special. The stable-requisite criteria are preconditions that are immutable, such as history of a disease, having intolerance to certain drugs, or having received a prior treatment. The variable routine criteria are criteria that depend upon data that are relatively stable over short time periods (e.g., the results of lab tests) and are collected routinely during patient care. The controllable criteria involve patient circumstances that a physician can modify. The subjective criteria involve a physician’s judgment. Examples include the likely duration of patient survival or the Karnofsky score for patient functional capacity. Finally, the special criteria are those that depend upon the results of unusual lab tests (often costly and invasive) that are not typically performed in the context of routine care. Such tests should not be performed until a patient is identified as a likely study candidate. An advantage of Tu’s classification is that prospectively, patients who are considered ineligible only because of variable routines or controllable criteria can potentially become eligible later or when specific actions are taken. Later, Papaconstantinous divided the criteria into two classes: constant or variable, which can be seen as a simplification of Tu’s classification method [9].

The third dimension of eligibility criteria classification addresses the complexities of semantic patterns in eligibility criteria. Fink et al. classified criteria into three types of questions: the first type takes yes/no response, the second takes multiple choices, and the third requires a numeric answer [14–16]. In a more sophisticated manner, in a review study of six representative computer-based clinical guidelines (EON, Asbru, PROforma, GUIDE, GLIF, PRODIGY) [49], Peleg et al. categorized eligibility criteria into presence criteria, template-based criteria, firstorder logic criteria, temporal criteria, “if-then-else” statements, and context-dependent expressions. A recent clinical guideline model SAGE [36] extended previous work in clinical guideline modeling and similarly classified eligibility criteria into presence criterion, N-ary criterion, goal criterion, comparison criterion, temporal comparison criterion, variable comparison criterion, and adverse-reaction presence criterion. In guideline models, condition criteria can be classified by their implications for the execution states of guideline actions or plans. For example, Asbru defines six types of conditions, including filter preconditions for a guideline to be applicable, set-up preconditions to enable a plan to start, suspend-conditions that determine when an active plan instance has to be suspended, as well as abort-condition, complete-conditions, and reactivate-conditions [34].

The classification of eligibility criteria is definitely a non-trivial problem that introduces great variation in practice. We found no prior study about the limitations or uses of existing classifications for eligibility criteria. Ideally, with a formal knowledge representation for eligibility criteria, all dimensions of eligibility criteria (e.g., content such as clinical topics or medical concepts, uses, features such as uncertainty for eligibility determination, and so on)

should be represented explicitly to enable automatic and flexible indexing, classification, retrieval, and usage of eligibility criteria.

3.3 Expression Language for Eligibility Criteria

Expression language is a critical component of a knowledge representation for eligibility criteria, because it serves to formally model relationships between multiple concepts embedded within eligibility criteria statements. Examples of relationships include comparison (e.g., "*serum creatinine < 2.0*") and constraints (e.g., "*hypertension among men above 70 years old*"). Languages of varying expressiveness have been used to represent the logic of eligibility criteria, including (a) ad hoc expressions; (b) the Arden Syntax; (c) variants of logic-based languages, including the PAL language in Protégé [50], the Structured Query Language (SQL) [51], and description logic [52]; (d) object-oriented query and expression languages, such as GELLO [53]; and (e) temporal query languages, such as Asbru [34] and Chronus II [54]. The above expression languages are compared below. The "Design-a-trial" system and OncoDoc both represent eligibility criteria as text strings; therefore, they are excluded from this analysis.

(a) Ad hoc Expressions

Description: The development of ad hoc expression languages is driven by use cases instead of any theoretical basis. In contrast, the Structured Query Language (SQL) has a theoretical basis in relational calculus and GELLO has root in Object Constraint Language, which has being given a formal foundation [55]. Examples of ad hoc expression languages include EON [29], SAGE [36], and ERGO [41].

The simplest type of ad hoc expression languages defines a set of *parameters* (e.g., *Presence_of_Renal_Insufficiency*, *Sitting_Blood_Pressure*, and *Menopausal_Status*) that can take Boolean, numeric, or enumerated values. The expression language provides comparison and logical operators (e.g., "=," ">," "AND") that allows the construction of logical expressions (e.g., "*Menopausal_Status = Post-Menopaual AND Sitting_Blood_Pressure > 150*") that can be evaluated in terms of raw patient data. Ad hoc expressions also can be constructed using a rich information model of patient data. In the latter approach, the expression languages assume relational or object-oriented information models (e.g., HL7 RIM). A formal syntax for ad hoc expressions allows the definition of queries, variables, and logical statements involving comparison, arithmetic combinations, arbitrary conjunctions and disjunctions, as well as temporal constraints on data records. Templates are often developed to assist the formulation of logical expressions by objectifying patterns such as queries including a comparison between a numeric entry (e.g., serum creatinine value) and a cut-off value (e.g., 140 mmHg), and (2) Boolean combinations of multiple queries.

Generally, it is easier to convert eligibility criteria to simple parameter-value ad hoc expression representations automatically rather than to translate such criteria into representations based upon computable and adaptable clinical data models, due to the semantic and knowledge gap between eligibility criteria and clinical data. For instance, "having major surgeries" may be represented as a simple parameter with a yes/no value without having to specify 'major surgeries' and how they are documented in EHR (e.g., surgical data could be encoded by Current Procedure Terminology [56] or textual surgery procedure names in patient reports). The downside of such simple parameter/value representation, of course, is that it mandates extra work to check eligibility by querying existing EHR systems.

Expressiveness: The expressiveness of ad hoc expression languages depends on the specific use cases driving the construction of the language. Some eligibility criteria pose requirements for considerable expressivity (e.g., "*average systolic blood pressure over two consecutive encounters more than 2 months apart is greater than 120*") [36]. In general, ad hoc expression

languages have limited capability to use formal reasoning methods, such as temporal constraint algorithms or predicate logic. Thus, for example, an expression like "*presence of an authorized medication that is contraindicated by some medical condition*") involves a relationship (*contraindication*) between two variables (*authorized medication* and *medical condition*). Without generic methods for formulating relationships among variables, ad hoc expression languages typically cannot handle such complex logical expressions.

Uses: Templates for ad hoc expressions of eligibility criteria are very popular and have been used broadly in OaSIS, OncoLink, caMatch, ASPIRE, CRFQ, EON, SAGE, and ERGO and the clinical trial screen system developed by Fink et al. at the University of South Florida [14–16]. Ad hoc expressions for eligibility criteria were further translated to SQL database queries in T-Helper. PROforma and Asbru represent "presence criteria" as data items with a Boolean value ("yes" or "no").

(b) The Arden Syntax

Description: The Arden Syntax originates from the knowledge encoding schemes of the Health Evaluation through Logical Processing (HELP) system developed at LDS hospital in Salt Lake City [57] and of the CARE system developed at Regenstrief Institute in Indiana for providing alerts and reminders [58]. It is not a programming language, but a hybrid between a production ("if-then") rule system and a procedural formalism. The Arden Syntax for Medical Logic Modules (MLM) is a HL7 standard for encoding medical knowledge and capturing condition-action rules. Arden provides rich time functions and explicit links to clinical data embedded in curly brackets that can support localization of variables flexibly. In addition, Arden has the concept of a single primary time associated with data variables, which is analogous to time-stamped data queries expressed as logical expressions. Wang et al. extended Arden by using "start time" and "end time" associated with all variables to represent time interval and by adding structures that may have attributes [12]. The extended Arden Syntax enables reasoning over the relationships among clinical terms contained in eligibility criteria.

Expressiveness: In comparison to ad hoc expressions, the Arden Syntax offers enhanced expressiveness through its rich collection of operators and time functions, and procedural control structures such as iteration. Arden does not provide declarative properties or defined semantics for making temporal comparisons or for performing data abstractions (e.g., retrieving an episode of uncontrolled blood pressure) [34]. Wang's extended Arden Syntax for eligibility criteria made incremental improvements by adding "structure data types", "dot operator" (e.g., "CHF.severity"), and an "enumerated data type" so that variables can only take on a value from a pre-defined list.

Uses: The Arden Syntax is probably the best supported expression syntax. Vendors who have developed Arden-compliant decision support applications include Eclipsys Corporation, McKesson Information Solutions, Siemens Medical Solutions Health Services Corporation, and MICROMEDEX. Arden Syntax was chosen as the representation language for an early version of GLIF, which was referred to as the Guideline Expression Language (GEL) [35]. Later it was used by Wang et al. [12], Ohno-Machado et al. [11], and Lonsdale et al. [21] for representing clinical trial eligibility criteria. Wang tested it on the PDQ database and estimated that 90% of criteria could be encoded using this syntax.

(c) Logic-Based Languages

Description: Major computer-based guideline models, including PROforma and GUIDE, use logic-based to represent decision criteria that contain more complex logic than ad hoc criteria, as thoroughly analyzed in Peleg's review paper [49]. These languages vary in their expressiveness. Here we briefly analyze three examples that have been used to represent

eligibility criteria: Protégé's constraint language (PAL) [50], The Structured Query Language (SQL), and description logic (DL).

PAL specifies a subset of first-order predicate logic written in the Knowledge Interchange Format (KIF) syntax [59]. It supports functions that test argument frames, as well as type-coercion functions and a few arithmetic functions. A constraint or query expressed in PAL consists of a set of variable range definitions and a logical statement that must hold on those variables.

SQL uses a set of operators, such as SELECT, DELETE, and UPDATE, to interact with relational databases. Eligibility criteria are conditions about patient data in clinical databases and can be directly encoded in a "where" clause in a SQL query. The syntax for the where clause is simple: "where column operator value." Column specifies variables that represent patient data; operators can be =, <, >, <>, <=, >=, between, like, and IN; value can be a single value, a set of values, or a range. Multiple where-clauses can be used to join several tables. A SQL query can be directly issued on a clinical database, but is hardly transferrable to other clinical databases with different schemas. It is expressive but has poor support for knowledge reuse or inference.

Description Logic (DL) is a special subset of first-order predicate logic that provides tractable classification reasoning over concept definitions. Description logic can be used to formally represent each criterion as a concept that has a necessary and sufficient formal definition to support computer-based reasoning over eligibility rules. Using an example provided by Patel et al. [24], '*presence of hematocrit test*' can be represented as a concept defined as "∃ assesses-sample.BloodTest Specimen ∧ ∃ entity-measure.Hematocrit", which says that 'presence of Hematocrit Test' is true if there is some 'blood test specimen' sample and the entity measured is 'Hematocrit'. The strength of DL is that there are pre-built reasoners that can classify patients into categories described using the description-logic expressions.

Expressiveness: The three types of logic-based languages encountered in our review each have pros and cons in expressiveness, SQL and PAL assumes negations as failures in a closed world view while DL makes open-world assumption, where something may exist even if it is not known to the reasoning system. The consequence is that DL expressions cannot have aggregation operators like maximum, most recent, or average. DL enables automated concept-based classification reasoning and query expansion when converting eligibility criteria to database queries. Recent versions of DL, especially Web Ontology Language (OWL) 2.0, allow specifications of transitive roles and numeric comparisons. DL cannot be used to formulate expressions such as "AST no greater than 5 times ULN" if the ULN (upper limit of normal) and measured AST value are part of the same patient record. SQL does not support as many first-order predicate logic functions as DL, but it is optimized for efficient processing of large data sets. PAL enables reasoning over the relationships between clinical terms and between classes and instances in knowledge bases.

Uses: EON uses PAL to define decision criteria [60]. Butt et al. used SQL to represent eligibility criteria and implemented a real-time patient recruitment system for clinical trials studies [22]. GUIDE also used SQL to represent decision criteria. Patel et al. implemented a description logic-based clinical eligibility screening system [24], where eligibility criteria were represented in OWL 1.0 as semantic queries over the New York Presbyterian clinical data warehouse. This system does not support reasoning over temporal constraints and cannot query numeric comparison criteria such as "SBP > 130 mm/Hg".

(d) Object-Oriented Expression and Query Language

Description: At present, there is only one object-oriented language for eligibility criteria: GELLO [53]. GELLO expressions operate over objects, and operators of the language can be extended by user-specified methods defined for classes of objects. GELLO is based on the Object Constraint Language (OCL), a query and expression language designed for writing constraints in a UML model. GELLO uses an object-oriented information model, such as a virtual medical record model based on HL7 Reference Information Model [35].

Expressiveness: In GELLO, users construct expressions using typed data values, variables, queries, functions, collections, and user-defined classes. The expression language is strongly typed. In order to facilitate the process of encoding and evaluation of expressions and more importantly, to maximize the ability to share such queries and expressions, GELLO includes basic built-in data types while providing the necessary mechanisms to access an underlying data model with all its associated classes and methods. This is especially important in enabling decision rules and guidelines to successfully work with different data models, inasmuch as classes and relationships specified can vary from one data model to another [61].

Uses: Initially developed as part of the GLIF representation language [36], GELLO has been adopted as an HL7 and ISO standard. It has gained support in the commercial sector where vendors, such as Medical Objects in Australia [62] and InferMed in United Kingdom [63], have implementations that are driving GELLO's further development. The GELLO query language has been designed within the context of a guideline execution model proposed in the HL7 CDSTC. This model proposes the use of a VMR (Virtual Medical Record) that provides a standard interface to heterogeneous medical record systems to construct decision criteria by building up expressions with which to reason about particular data features/values. These criteria can be used to provide alerts and reminders, guidelines, or other decision rules [61].

(e) Temporal Query Languages

Description: The representation of time within eligibility criteria for clinical research protocols is an important challenge [64]. While most implementations of SQL have date-time data types with special operators, none of the logic-based languages support sophisticated temporal reasoning. Requirements for temporal knowledge representation for clinical trial protocols include the representation of (i) relative time information (e.g., events are relative to protocol time points such as “Baseline” and “Day 1”), (ii) indeterminacy (e.g., “+/− 1 day”), (iii) cyclical event patterns (e.g., “*every 3 weeks*”), (iv) both time points and time intervals (e.g., “*pre-study*” and “*treatment period*”), and coalescing of temporal intervals that satisfy some condition (e.g., duration of period when the dose of drug is > N). Also, single-point time intervals, such as “follow-up” events where the start-point or start date is known but the interval is ambiguous, are very common in clinical research protocols. Many eligibility criteria contain temporal constraints and require abstraction or reasoning of temporal patient conditions. Some representations for eligibility criteria do not explicitly support temporal queries [14,15,20], while many others only support query over time-stamped data [12,30,40,65].

Numerous groups have proposed temporal query languages for managing clinical data in the past. Here, we reviewed two representative examples that have been explicitly used for encoding and evaluating eligibility criteria: Asbru [34], a constraint-based language, and Chronus [66] and its successor Chronus II [54] that adapted query languages developed by the temporal database community. Asbru's temporal expression language, with its syntax specified using Backus-Naur form (BNF) [34], supports the specification of temporal constraints on the beginning point, ending point, duration, and repeating patterns in parameter/value-type conditions which need to hold at a plan step to induce a particular state transition in the plan instance. Chronus II adapts the TSQL2 temporal query language [67] to extend the standard

relational model and the SQL query language to support temporal queries that include temporal projection, joins, granularity conversion, and coalescing. It provides an expressive general-purpose temporal query language that is tuned to the querying requirements of clinical decision support systems.

Expressiveness: Asbru excels in expressing temporal constraints among events. It is XML-based criterion language that allows specification of value set or ranges, context, and temporal extent of parameters. However, Asbru does not use any standard clinical terminology or patient information model; therefore, its strength is limited to only temporal aspects. Chronus II adopted features, such as temporal coalescing, that were developed in the temporal database community that are not implemented in any other expression language reviewed in this paper.

Uses: Asbru was used in the Asgaard project [33]. Chronus II was used in the ATHENA decision support system and some data-mining projects at Stanford [68].

(f) Others—In addition to the above expression languages, Ohno-Machado et al. represented eligibility criteria using the Bayesian Belief Net mechanism and probabilistic methods to address the frequent “missing data” challenge in eligibility matching by representing complex relationships among different variables in the AIDS2 system [20], as did Cooper et al. in another screening system connected to a clinical data repository [69]. Eligibility criteria were represented as criteria nodes with probabilities on possible values. Every node in their belief network either represented a clinical parameter (e.g., “Hemoglobin”) that was used in a criterion (e.g., “Hemoglobin > 11”), or represented a clinical data element (e.g., "Anemia"), which differed from a clinical parameter in that it influenced other nodes. These influences were expressed as probabilistic dependencies (e.g., probability that hemoglobin > 11 given that the patient has clinical signs of anemia). The values of nodes in the network could be either unknown or set to a particular predefined state. If the state of a node was known, the probability of that state was 1.0, whereas the probabilities of all other states were 0.0. If the node value was unknown, the network was used to compute a posterior distribution conditioned on all nodes that have a relationship with the node being observed. AIDS2 determined the patient eligibility status based on each criterion by examining the value of the corresponding belief-network criterion node. The patient eligibility for each criterion was represented by a probability, which was the addition of the probabilities of all mutually exclusive states of the criterion node that are considered to be eligible.

(g) Summary of Expression Languages—In summary, expression languages employed to represent eligibility logic include ad hoc expressions (with or without the use of templates), the Arden Syntax, logic-based languages (i.e., PAL, SQL, and DL), object-oriented languages (i.e., GELLO), and temporal query languages (e.g., Asbru and Chronus II). The next generation of clinical research systems that depend upon knowledge representations for clinical research eligibility, whether they be protocol authoring, clinical research results databases, study metadata archives, patients screening, or public health (health services delivery research), will require more robust expressive languages. Ad hoc formalisms were functional and innovative for the first generation of systems processing eligibility criteria, and have paved the way for our understanding of this complex and vast area. SQL-based queries on a clinical database are more expressive but not extensible for knowledge reuse or inference. These mechanisms all suffer from the lack of scalability. We observed that occasionally multiple query languages were used for different types of logic within the same model or system. For instance, EON used three languages to represent eligibility criteria of different complexities [60], including (1) using ad hoc templates to encode common but relatively simple criteria by filling in forms for presence criteria, comparison criteria, and Boolean combinations of multiple criteria; (2) using PAL constraint language to implement criteria that require reasoning over relationships

of medical concepts; and (3) using Chronus-II temporal query language to encode complex criteria that require complex reasoning over overlapping intervals for two events and coalescing of time intervals.

3.4 Encoding of Eligibility Concepts

The expression language defines the syntax for specifying eligibility criteria statements, and the eligibility concepts provide the semantics. Olasov and Sim summarize the challenge of a representation for computable eligibility criteria as having two major components: (1) mapping terms within individual eligibility rules to concepts in a controlled clinical vocabulary and (2) capturing intended relationships between concepts and their modifiers [4]. A sample of eligibility criteria from active studies in the Rare Disease Clinical Research Network [70] was presented at the 2007 AMIA symposium [5]. Of 452 eligibility criteria from 19 protocols (largely observational studies) on 22 diseases, the majority of criteria (44%) represented clinical findings, which included clinical diagnostic criteria, diseases, and symptoms. Almost half of the 452 eligibility items surveyed contained multiple clinical concepts, e.g., "e*vidence of significant chronic or acute inflammation outside the lung such as connective tissue diseases, panniculitis or acute infection.*" Non-specific concepts are often included in eligibility criteria, such as "*Neurological illness*" or "*Uncontrolled seizure disorders*". Laboratory measures might be expressed in terms of their interpretation (e.g., "*elevated sodium*") or an institution-specific reference range (e.g., *ULN for upper limit of normal*). The use of vague terms in clinical research protocols is common and has been observed by Ohno-Machado [20]. Therefore, an important component of eligibility criteria representation is codified terminologies [71]. Computable eligibility criteria representations should support reasoning over the relationships among different concepts – particularly the determination of equivalence and subsumption between different terms.

Some clinical trial recruitment systems use locally developed medical concept classifications [20,21]. We observed that most systems prior to year 1999, including ONCOCIN, T-Helper, AIDS2, OaSIS, and some other unnamed systems [9,27], did not employ any standard clinical terminology to encode medical concepts in eligibility criteria, likely because clinical terminologies and supporting tools such as the Unified Medical Language System (UMLS) [72] were still being developed and unavailable until the late 1990s. Since 1999, the importance of using clinical terminologies has been recognized in the literature as a critical practice to support information interoperability, although no widespread agreement on standards exists. Wang et al. noted that the choice of a clinical vocabulary was tightly linked to the implementation of a practical data-query and data-modeling scheme [12], and used UMLS to extend the Arden Syntax for representing clinical trial eligibility criteria. With the increasing adoption of EHR and PHR since early 2000s, UMLS has been a popular choice for encoding medical concepts in eligibility criteria because of its interoperability with other medical terminologies and notable natural language processing software such as MedLEE [73]. The Guideline Interchange Format (GLIF v3.0) [35] used the UMLS to support clinical concept representation as well. Similarly, Sim allows an option to map the clinical-rules and the longest phrases in RuleEd to UMLS concepts [4]. LOINC [74] was suggested as a candidate terminology for representing concepts in lab results [75]. Although systems (e.g., Trialx [76]) have been developed to match clinical trials to PHR, to date, there is no consumer health vocabulary available or in use for encoding medical concepts in eligibility criteria to serve the growing needs of clinical trial search initiated by health consumers. The UMLS remains the popular choice for medical concept encoding.

Another trend in encoding medical concepts in eligibility criteria is using Common Data Elements (CDEs), which started with Gennari's use of the NCI's CDEs to serve as a medical terminology for oncology clinical trial protocols [19]. Later, ASPIRE and caMatch also

collaborate with CDISC in developing CDEs for encoding or annotating medical concepts in eligibility criteria. Different from UMLS, CDEs are standards for content and do not define formal relationships among concepts, and hence do not provide inferential capacity.

Perhaps because UMLS is too broad in scope for the focused domain of eligibility criteria, the Systematized Nomenclature in Medicine – Clinical Terms (SNOMED CT) [77] has been preferred as the encoding terminology for clinical concepts by researchers working on GUIDE, SAGE, ERGO, and others in recent years [4,23–25,28,36,38]. One advantage of SNOMED CT is that it allows for the creation and logical definition of new concepts using pre-coordinated terms that already exist in SNOMED CT. The READ codes [78] were another comprehensive clinical terminology developed and used in the United Kingdom, and later merged with SNOMED RT to produce SNOMED CT. The READ codes were used in the PRODIGY clinical guideline model as the encoding vocabulary [39].

Several systems used more than one terminology for different data. For example, the current version of RuleEd (http://rctbank.ucsf.edu:9002/BaT/RuleEd.html) can map extracted clinical phrases to either Medical Subject Headings (MeSH) [79] or SNOMED CT. The SAGE model uses a suite of terminologies, including SNOMED CT for clinical terms, LOINC for lab tests, and the National Drug File –Reference Terminology (NDF-RT) for drugs and related class information [80], and delineates three levels of use (pre-coordination, post-coordination, and Boolean combinations) for standard terminologies for encoding and executing clinical practice guideline knowledge bases.

Mapping from concepts (e.g., "*patients with high blood pressure*") to clinical data manifestations (e.g., "*SBP > 140 mm Hg*") is often not straightforward. Terminologies by themselves are insufficient for helping us achieve automated matching between computable eligibility criteria and EHR data for two primary reasons. First, concepts embedded within eligibility criteria can be underspecified. For instance, there is no straightforward mapping for "*chronic diseases*" without manual and subjective selection of relevant common chronic diseases such as diabetes, hypertension, and so on. Second, there is a knowledge gap between concepts in eligibility criteria and EHR data captured in specific clinical contexts. For instance, an eligibility criterion may specify the concept "*renal failure*", while EHR data must be pieced together to identify, for example, "*an 80-year old white female with serum creatinine = 1.0 mg/dl*". There is no way to map "*renal failure*" to serum creatinine value without knowing the patient's age, gender, ethnicity, and their relationship to renal failure. Therefore, terminologies need to be used in combination with a patient information model and relevant medical knowledge in order to facilitate automatic matching between eligibility criteria and EHR data. Some systems, such as OncoDoc, compensated this knowledge gap by replacing underspecified terms. OncoDoc is a decision support system that assists physicians in deciding patient eligibility for clinical trials. In OncoDoc, terms from controlled vocabularies (e.g., UMLS, SNOMED CT, etc.) have not been used because they were considered either too general or incomplete to take into account patient's preferences and to support daily medical oncology practice [13]. Instead, explicit definitions were created to compensate for the ambiguity or incompleteness in these concepts. For example, "cardiac function" in OncoDoc is expanded to be "*good cardiac function (fractional shortening > 35% or ejection fraction > 50%)*" and "*bad cardiac function (fractional shortening < 35% or ejection fraction < 50%)*".

3.5 Patient Data Modeling

Another important component of formal representations of eligibility criteria – one as important as the use of controlled clinical terminologies – is a patient information model that supports the inference of medical concepts in reference to corresponding patient data [71]. In order to serve the decision support use cases identified earlier and to enable translation from eligibility criteria to EHR-based patient data queries without knowing individual EHR implementation

details, representations of eligibility criteria need to support standard-based modeling of patient data, which is often through standard patient information models [1].

Systems developed prior to the year 2000 rarely used a patient information model, but tended to define patient data as pairs of parameters and values, where parameters represented the attributes of patients, drugs, tests, and so on, and each value had an associated time stamp that denoted when that value was observed or a temporal interval when the value held true. Examples include AIDS2, OncoLink, Fink's system, the Arden Syntax, and OaSiS. Systems developed after the year 2000 largely adopted some form of Virtual Medical Records (VMR) [81] based on the HL7 Reference Information Model (RIM) [82], which provides an abstraction layer on top of a real EHR. Among the 27 knowledge representations for eligibility criteria that we surveyed, nearly half adopted a VMR, including GUIDE, GLIF3, SAGE, ERGO, CRFQ, as well as Patel's [24] and Lonsdale's [21] systems, with varying degrees of adoption. (For instance, only one "observations" class from a HL7 VMR model was used in Lonsdale's system.) Although there is no consensus in the medical informatics community regarding a standard patient information model, the development of a VMR based on the HL7 RIM shows promise to mitigate the classic site-specific data mapping problem (again, the "curly bracket problem").

3.6 Domain Specificity and Uptake Status of Existing Systems

Many systems or models we surveyed are generic or domain independent representations for condition criteria, though several (ONCOCIN, OaSIS, OncoDoc, OncoLink, and caMatch), were specifically designed for the cancer domain. ASPIRE was initially disease-specific with a focus on breast cancer, but has been expanding to other disease areas and includes domain independent (pan disease) data elements.

The adoption status of most of the models sampled in our review is unavailable, primarily because they represent academic research prototypes, which did not lead to real-world or widespread adoption. Though uptake is hard to assess, it is clear that most systems we surveyed were implemented in single organization settings. To our knowledge, EON, ERGO, PROforma, Asbru, OncoLink, caMatch, ASPIRE, SAGE, and Arden Syntax are actively used in ongoing projects. Among those, caMatch, ASPIRE, and OncoLink are used in the following web-based patient recruitment systems respectively: https://www.breastcancertrials.gor/bct, http://clinicaltrials.cop.org/, and http://www.oncolink.com. OncoDoc has been implemented at the Institute Gustave Roussy (IGR), known as the first European cancer research center, and routinely used at the point of care during a 4-month period [10]. GELLO has been adopted as an HL7 and ISO standard. Commercial vendors, such as Medical Objects in Australia [62] and InferMed in United Kingdom [63], have active implementations that are driving GELLO's further development.

4. Discussion

In this study, we used a set of keywords to search the literature for existing knowledge representations for eligibility criteria. We identified five aspects of eligibility criteria knowledge representations that contribute to the existing heterogeneous approaches: *the intended use of computable eligibility criteria, the classification of eligibility criteria, the expression language for representing eligibility rules, the encoding of eligibility concepts, and the strategy for modeling patient data*. Each of these aspects has a spectrum of options. We also consider three of these aspects - *expression language, codification of eligibility concepts, and underlying model of patient data* - to be essential constructs of a formal knowledge representation for eligibility criteria. Requirements for these three knowledge constructs vary for different use cases and eligibility criteria statements of different complexities. It is feasible to combine multiple expression languages and multiple terminologies to achieve expressive

and interoperable eligibility criteria knowledge representations. Next, we discuss the implications for related standardization efforts in this area.

4.1 A Conceptual Framework for Organizing Eligibility Criteria Representations

Using the five knowledge constructs for eligibility criteria identified in this paper, uses of eligibility criteria knowledge representations can be categorized with respect to the expressiveness of eligibility rule expression languages, the range of terminologies for eligibility concept modeling, and the inclusion of a patient model. Some use cases such as knowledge management do not require patient data modeling as well as eligibility rules expression and inference. For example, reuse of eligibility criteria during protocol-authoring does not expressive expression language or patient data modeling. In contrast, use cases such as eligibility determination or clinical evidence applicability determination require interoperability with patient data. The knowledge representation requirements for eligibility criteria in these contexts are more stringent, including highly expressive language(s) to achieve executable eligibility rules, a patient information model, and an appropriate clinical terminology to facilitate mapping from eligibility concepts to patient data. Applications that support indexing, classification, and annotations of clinical studies (e.g., ASPIRE and caMatch) require medium expressiveness for representation languages - less than that required for applications designed for applicability determination but more expressivity than is required for protocol authoring. They do not need a computable representation for a patient data model.

Appendix Table 1 shows that an expressive language is very important for clinical decision support uses of formal representations of eligibility criteria (e.g., applicability and eligibility determination); a patient model is less important for uses such as knowledge reuse in protocol authoring; and a controlled terminology is indispensible for all uses but the choice of a specific terminology greatly depends on the use. Eligibility criteria representations designed to support systems for applicability determination may better use a literature-oriented terminology such as MeSH to encode medical concepts, while representations designed to support eligibility determination may better use a clinically-oriented terminology such as SNOMED CT. Because terminologies have coverage of different domains and variable structures, they are individually suited for particular uses and should not be hard-coded in knowledge representations. Natural Language Processing (NLP) of eligibility criteria can be used to extract key eligibility concepts and support flexible mappings to a range of terminologies (e.g., MeSH and SNOMED CT); RuleEd already supports this feature. Therefore, a formal knowledge representation of eligibility criteria may better support a component-based design to allow flexible “plug-and-play” of options for each construct suitable for different uses. Moreover, Appendix Table 1 implies some dependencies among the choices and/or needs for terminologies, patient data models, and expression languages. For example, representations that need patient data modeling tend to need expressive query languages and a clinically oriented terminology with coverage of patient data (e.g., SNOMED CT).

We can cluster various eligibility representations by use cases. The clusters that indicate similar representation efforts should be harmonized. For example, there is a higher degree of similarity between caMatch and ASPIRE than between ERGO and ASPIRE; caMatch and ASPIRE share the same level of expressiveness in their query languages, while ERGO has a more expressive query language than ASPIRE. Therefore, it is easier to convert instances from ERGO to ASPIRE, but conversion in the opposite direction is harder, because ASPIRE instances do not specify computable rules for eligibility criteria while ERGO uses a more expressive query language to represent eligibility rules. Moreover, at present, the activities in the clinical research cycle do not share a unified underlying eligibility criteria model. The clinical research cycle includes multiple sequential steps, including literature review, new research question identification, protocol authoring, subject recruitment, data collection (i.e., study conduct), and

results publishing. Each step can be mapped to an intended use of a formal knowledge representation for eligibility criteria. At present, the uses of research protocol authoring and management systems are the primary driving efforts for formal representations of eligibility criteria. Practically, it may not be cost-effective to use one single, complicated representation to support all possible uses of computable eligibility criteria. On the other hand, a use-driven knowledge representation design principle (that allows for multiple representations) may create discontinuity and barriers for a holistic and streamlined discovery process because a user may have to use different knowledge representations for various uses, such as searching literature, authoring protocols, linking protocols to patient databases, and publishing research results. The clinical research informatics community may want to address this research challenge together and develop formal knowledge representations that support translational research through the whole research life cycle.

4.2 Standardization of Eligibility Criteria Representations: Challenges and Opportunities

Standards-based knowledge representation for eligibility criteria is an active research area [83]. With the availability of rich standards for each construct for formal knowledge representation of eligibility criteria, many researchers have been using multiple standards in one representation. For example, EON used three expressional languages to represent eligibility rules. Concurrent uses of multiple standards may be a pragmatic solution, but also contributes to the challenges of standardization and harmonization in this domain. Fortunately, several high-profile efforts, including ERGO, ASPIRE, and CDISC, have been working collaboratively to achieve community-wide agreements on related standards. We believe that representations within a given cluster could be harmonized to reach community consensus, while representations in different clusters will need interoperability support between each other. In the past decade, shared clinical terminologies, standard patient information models, and standard expression and query languages have been increasingly recognized as important tools for achieving interoperability across health organizations. Although these goals are gaining popular support, clearly there are significant barriers to achieving them.

The challenges for standards-based encoding of eligibility concepts are multifaceted. The complexity in clinical statements is a key factor that causes the variant needs for a broad range of expression languages. One single terminology often cannot cover all the concepts embedded in eligibility criteria so that multiple controlled terminologies for clinical findings, test results, labs, medications, or medications often need to be used together. Additionally, it has been shown that there are certain structural features and information facets of eligibility criteria that are not fully represented by some terminology models [84]. The Consolidated Health Informatics (CHI) standard [85] for standardized Patient Assessment items and subsequent Department of Health and Human Services (DHHS) standards recommendations are based upon the premise that one standard (LOINC) is required to represent structural and question administration features (e.g., unit, method, subject, period of observation), and that additional clinical vocabulary is required to represent the clinical concepts contained in the questions [86]. To date, one of the few terminology-encoded eligibility criteria knowledge representations is ERGO (Eligibility Rule and Grammar Ontology) model, which supports using vocabularies (SNOMED CT or MeSH) in the context of an information model (HL7), and points to the best practices for dealing with terminology [6].

In recent years, the latest standardization focus has been on common data elements (CDE), which serve as standard metadata [87]. The National Cancer Institute (NCI) and the Clinical Data Interchange Standards Consortium (CDISC) are developing CDE repositories in support of standards-based clinical research activities. The CDEs are structured data reporting elements, consisting of precisely defined questions and answers, which represent eligibility criteria the same as any other research data element. The uses of metadata and vocabulary

standards for indexing eligibility criteria, in applications such as the NCI's caDSR [88,89], could drive better-authored eligibility rules (when investigators understand how the questions will be indexed) and thereby improve the reuse of existing eligibility criteria. Since such repositories (also called metadata repositories or item banks) have only emerged within the past few years, only a few recent formal representation efforts of clinical eligibility criteria, e.g., including caMatch and ASPIRE, have adopted and extended these CDEs. The caMatch project includes collaborations with HL7, CDISC, and OMG Vocabulary-driven data entry to use CDEs from NCI's caDSR repository. However, there is a significant difference between the ASPIRE approach and the caMatch approach. ASPIRE uses CDE as metadata to index and annotate eligibility criteria, instead of trying to capture the precise clinical statements represented by the criteria. In contrast, the caMatch approach uses standard CDE terms to define eligibility criteria constructs. Both approaches have advantages and disadvantages. In the ASPIRE approach, comprehensive domain-specific data elements are closely connected to the retrieval accuracy of annotated eligibility criteria.

It is advantageous in that a criterion can be flexibly annotated with multiple CDEs anytime. The CDEs then can be used to enable flexible and dynamic multidimensional categorizations of eligibility criteria, and consequently support their storage and re-use. However, this approach does not define computable expressions of eligibility criteria. The caMatch approach that uses CDEs together with expression and query languages to represent eligibility criteria can be expressive, but the expressiveness is contingent on the coverage of the clinical terminologies and patient information models being used. A range of clinical terminologies are needed to collectively represent a variety of clinical statements. Multiple clinical terminologies will be needed to support representation for different data sources, including lab tests, medication, diagnosis, and free text reports. While the NCI caDSR is beginning to formally relate CDEs to the standardized terminologies hosted by the NCI, there has been criticism on the completeness and validity of these relationships [87]. We consider CDE and expression languages to be complementary for the development of formal representation of eligibility criteria. Expression languages can be used to organize CDEs to construct computable eligibility criteria statements so that the CDEs can be evaluated against EHR data. To leverage their complementary strengths, a useful implementation of the CDE idea would be a library of executable rules expressed in computable eligibility criteria languages. Regardless, it is foreseeable that there will be criteria that either cannot be formalized as computable expressions or will not have EHR data to support automated evaluation of CDEs.

During the natural evolution of methods and conceptualization of formal representations for eligibility criteria, when no standards existed, none were used; as standards begin to emerge and multiply, standards start to be used in various ways. Therefore, a recent trend is also the standardization of the uses of standards, or *standards best practice*. We envision the harmonization of existing standards for expression languages, patient information models, and supporting terminologies will inform best practices for the authoring of eligibility criteria and their formal representation. Robust and harmonized representations for eligibility criteria can have immediate impact on the speed of clinical research and improving human health.

4.3 Future Work

Future work includes addressing the emerging needs from public health informatics, consumer health informatics, and clinical research informatics, improving interoperability between computable eligibility criteria and clinical data in EHR and PHR by bridging the semantic and knowledge gaps between both, and developing terminology standards that can cover a broad range users, especially health consumers who may use lay language terms to describe their medical situations and search for related clinical evidence or clinical research opportunities. It will not be a trivial undertaking to make the next generation of eligibility criteria representations

fully standards-based and amenable to automatic retrieval, agile classifications, indexing and reuse. Similarly, bridging the semantic gap and using such representations in heterogeneous EHR and PHR systems will require time, resources, and intellectual input from a broad group of stakeholders. As a rule, the ideal choices of standards should suit intended use cases. For the same use case, related clinical research standards should be harmonized within the research community; for different uses, representations should have interoperability with each other. Efforts should be made to achieve a comprehensive, standards-based knowledge representation for eligibility criteria that supports the full cycle of translational research, from literature review, to protocol authoring, to trial recruitment, and to study publishing. Therefore, knowledge representation for eligibility criteria should not be narrowly focused on one of the above use cases but is expressive or flexible enough to support multiple use cases.

Based on the new emphasis from the DHHS on translational research, interoperability between data systems and standards between health delivery and research will drive the requirements for formalisms for eligibility criteria. Programmatic aims for increasing the efficiency of clinical research (i.e., The NIH Roadmap Initiative for Re-engineering the Clinical Research Enterprise) intrinsically include requirements for scalability of systems – which will depend upon standardized representations for eligibility. Harmonization – both within clinical research communities, and across healthcare and public health communities, will be fruitful. The quality and uptake of standards will require participation and support of many stakeholders – researchers and systems developers, vendors and academia. The topic of formalized and standardized knowledge representation for eligibility criteria should therefore be at the top of discussion agendas for clinical research informatics stakeholders. We believe that these discussions should involve all stakeholders and advocate for continuous communications in this very important area.

4.4 Limitations

Our keywords list for the literature search may not have been exhaustive and we might have neglected to consider some relevant representations, especially those embedded in clinical decision support systems without explicit references to eligibility criteria representations. Moreover, some of the articles that we retrieved did not have sufficient details for condition criteria representations, and hence we could not review such systems' representations as thoroughly as we would have liked. Further, our depiction of the relationships between features of eligibility representation and uses in turn derived from our limited sample, and is subject to the same omissions and bias. Our characterization of "essential" representation primitives for eligibility criteria, and our characterization of intended purposes of various prototype systems, derived from the literature sample itself, and therefore might not be exhaustive. We did not control for the use, effectiveness, evaluation, or any other measure of fitness or success of any of the systems which we have described in this review. Despite that we did choose to speculate on future directions for eligibility criteria representation formalisms, standards, and systems use, we drew upon our expertise to support these speculations, but our ideas have not yet been formally vetted with clinical research informatics specialists, trialists, or systems developers.

5. Conclusion

We reviewed the diversity of the existing knowledge representations for eligibility criteria across three representation primitives, which are expression language, codification of medical concepts, and modeling of patient data, as well as the variations in their intended uses and content classification. This review demonstrates the complexity in eligibility criteria statements, which entails the need for the combinational uses of multiple standards. We also hope this conceptual framework can serve as an evaluation matrix for future users or developers

of computable eligibility criteria to select relevant standards and to identify compatible representation efforts toward collaborative standards development in this area.

Acknowledgments

This research was funded under NLM grant 1R01 LM009886-01A1 and CTSA award UL1 RR024156. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH.

References

1. Wang D, Peleg M, Tu SW, Boxwala AA, Greenes RA, et al. Representation primitives, process models and patient data in computer-interpretable clinical practice guidelines: A literature review of guideline representation models. International Journal of Medical Informatics 2002;68:59–70. [PubMed: 12467791]
2. Newell A. The Knowledge Level. AI Magazine 1981;2:1–33.
3. Klausner, RD.; Silva, JS.; Ball, MJ.; Chute, CG.; Douglas, JV., et al. Cancer Informatics. 1st edition. Springer; 2002.
4. Olasav, B.; Sim, I. American Medical Informatics Annual Symposium. Washington, D.C: 2006 Nov 11–15. RuleEd, a Web-based Semantic Network Interface for Constructing and Revising Computable Eligibility Rules; p. 10512006
5. Niland, J.; Dorr, D.; El Saadawi, G.; Embi, P.; Richesson, RL., et al. American Medical Informatics Association Annual Symposium. Chicago: 2007. Knowledge Representation of Eligibility Criteria in Clinical Trials.
6. ERGO. A Template-Based Expression Language for Encoding Eligibility Criteria. Available at http://128.218.179.58:8080/homepage/ERGO_Technical_Documentation.pdf.
7. Clinical Research Filtered Query. Available at http://hssp-cohort.wikispaces.com/space/showimage/SFM_CRFQ_v1.0.doc.
8. Niland J. ASPIRE: Agreement on Standardized Protocol Inclusion Requirements for Eligibility. an unpublished web resource. 2007
9. Papaconstantinou C, Theocharous G, Mahadevan S. An expert system for assigning patients into clinical trials based on Bayesian networks. J Med Syst 1998;22:189–202. [PubMed: 9604785]
10. Séroussi JB, Bouaud J, Antoine E-C, Zelek L, Spielmann M. Using OncoDoc as a computer-based eligibility screening system to improve accrual onto breast cancer clinical trials. Artificial Intelligence in Medicine 2001;29:153–167.
11. Ohno-Machado L, Wang S, Mar P, Boxwala A. Decision support for clinical trial eligibility determination in breast cancer. Proc AMIA Symp 1999:340–344. [PubMed: 10566377]
12. Wang SJ, Ohno-Machado L, Mar P, Boxwala AA, Greenes RA. Enhancing Arden Syntax for Clinical Trial Eligibility Criteria. Proc AMIA Symp 1999:1188.
13. Séroussi B, Bouaud J, Antoine É-C. OncoDoc: a successful experiment of computer-supported guideline development and implementation in the treatment of breast cancer. Artificial Intelligence in Medicine 2001;22:43–64. [PubMed: 11259883]
14. Fink E, Hall LO, Goldgof DB, Goswami BD, Boonstra M, et al. Experiments on the automated selection of patients for clinical trials. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics 2003:4541–4545.
15. Fink E, Kokku PK, Nikiforou S, Hall LO, Goldgof DB, et al. Selection of patients for clinical trials: an interactive web-based system. Artificial Intelligence in Medicine 2004;31:241–254. [PubMed: 15302090]
16. Nikiforou S, Fink E, Hall LO, Goldgof DB, Krischer JP. Knowledge acquisition for clinical-trial selection. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics 2002:66–71.
17. Metz MJ, Coyle C, Hudson C, Hampshire M. An Internet-Based Cancer Clinical Trials Matching Resource. Journal of Medical Internet Research 2005;7:e24. [PubMed: 15998615]
18. Cohen, Eea. caMATCH: A Patient Matching Tool for Clinical Trials. caBIG Annual Meeting. 2005

19. Gennari J, Sklar D, Silva J. Cross-tool communication: From protocol authoring to eligibility determination. Proc AMIA Symp 2001:199–203. [PubMed: 11825180]
20. Ohno-Machado L, Parra E, Henry S, Tu S, Musen M. AIDS2: a decision-support tool for decreasing physicians' uncertainty regarding patient eligibility for HIV treatment protocols. Proc Annu Symp Comput Appl Med Care 1993:429–433. [PubMed: 8130510]
21. Lonsdale DW, Tustison C, Parker CG, Embley DW. Assessing clinical trial eligibility with logic expression queries. Data & Knowledge Engineering 2007;66:3–17.
22. Butte A, Weinstein D, Kohane I. Enrolling patients into clinical trials faster using RealTime Recuiting. Proc AMIA Symp 2000:111–115. [PubMed: 11079855]
23. Patel C, Cimino J, Dolby J, Fokoue A, Kalyanpur A, et al. Matching Patient Records to Clinical Trials Using Ontologies. IBM Thomas J Watson Research Center. 2007
24. Patel C, Cimino J, Dolby J, Fokoue A, Kalyanpur A, et al. Matching Patient Records to Clinical Trials Using Ontologies. Proc. of International Semantic Web Conference 2007:816–829. 2007.
25. Patel C, Cimino J. Semantic Query Generation from Eligibility Criteria in Clinical Trials. AMIA Annu Symp Proc 2007:1070. [PubMed: 18694168]
26. Musen M, Carlson R, Fagan L, Deresinski S, Shortliffe E. T-HELPER: automated support for community-based clinical research. Proc Annu Symp Comput Appl Med Care 1992:719–723. [PubMed: 1482965]
27. Rubin D, Gennari J, Srinivas S, Yuen A, Kaizer H, et al. Tool support for authoring eligibility criteria for cancer trials. Proc AMIA Symp 1999:369–373. [PubMed: 10566383]
28. Patel, C.; Cimino, J. Proc of AMIA'06. Washington DC: 2006. Using Ontology Reasoning to Match Electronic Patient Records to Clinical Trials.
29. Musen MA, Tu SW, Das AK, Shahar Y. EON: a component-based approach to automation of protocol-directed therapy. JAMIA 1996;3:367–388. [PubMed: 8930854]
30. Hammond P, Sergot M. Computer support for protocol-based treatment of cancer. The Journal of Logic Programming 1996;26:93–111.
31. Sutton DR, Fox J. The Syntax and Semantics of the PROforma Guideline Modeling Language. JAMIA 2003;10:433–443. [PubMed: 12807812]
32. Fox J, Johns N, Lyons C, Rahmanzadeh A, Thomson R, et al. PROforma: a general technology for clinical decision support systems. Comput Methods Programs Biomed 1997;54:59–67. [PubMed: 9290920]
33. Shahar Y, Miksch S, Johnson P. The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. Artificial Intelligence in Medicine 1998;14:29–51. [PubMed: 9779882]
34. Miksch, S.; Shahar, Y.; Johnson, P. Proc. of the 7th Workshop on Knowledge Engineering Methods and Languages (KEML-97). Milton Keynes, UK: 1997. Asbru: A task-specific, intention-based, and time-oriented language for representing skeletal plans; p. 1-20.
35. Boxwala A. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. Journal of Biomedical Informatics 2004;37:147–161. [PubMed: 15196480]
36. Tu SW, Campbell JR, Glasgow J, Nyman MA, McClure R, et al. The SAGE Guideline Model: Achievements and Overview. JAMIA 2007;14:589–598. [PubMed: 17600098]
37. Quaglini S, Stefanelli M, Cavallini A, Micieli G, Fassino C, et al. Guideline-based careflow systems. Artificial Intelligence in Medicine 2000;20:5–22. [PubMed: 11185420]
38. Quaglini S, Stefanelli M, Lanzola G, Caporusso V, Panzarasa S. Flexible guideline-based patient careflow systems. Artificial Intelligence in Medicine 2001;22:65–80. [PubMed: 11259884]
39. Johnson P, Tu S, Booth N, Sugden B, Purves I. Using scenarios in chronic disease management guidelines for primary care. Proc AMIA Symp 2000:389–393. [PubMed: 11079911]
40. Shortliffe, EH.; Scott, AC.; Bischoff, MB.; Campbell, AB.; Melle, Wv, et al. ONCOCIN: An expert system for oncology protocol management. Seventh International Joint Conference on Artificial Intelligence; Vancouver, B.C. 1981.
41. Sim I, Olasov B, Carini S. An ontology of randomized controlled trials for evidence-based practice: content specification and evaluation using the competency decomposition method. Journal of Biomedical Informatics 2004;37:108–119. [PubMed: 15120657]

42. Nammuni K, Pickering C, Modgil S, Montgomery A, Hammond P, et al. Design-a-trial: a rule-based decision support system for clinical trial design. Knowledge-Based Systems 2004;17:121–129.
43. Chute, C. Medical Informatics: Knowledge Management and Data Mining in Biomedicine. Vol. Chapter 6. Springer US: 2005. Medical Concept Representation.
44. Li L, Chase H, Patel C, Friedman C, Weng C. Comparing ICD9-Encoded Diagnoses and NLP-Processed Discharge Summaries for Clinical Trials Pre-Screening: A Case Study. Proc of 2008 AMIA Fall Symp 2008:404–408.
45. CancerGrid. Available at http://www.cancergrid.org/.
46. Van Spall HGC, Toren A, Kiss A, Fowler RA. Review. Eligibility Criteria of Randomized Controlled Trials Published in High-Impact General Medical Journals. A Systematic Sampling Review. JAMA 2007;297:1233–1240. [PubMed: 17374817]
47. PDQ. [Accessed on August 4, 2008]. Available at http://www.nci.nih.gov/search/clinical_trials/
48. Tu SW. A Methodology for Determining Patients' Eligibility for Clinical Trials. Methods of Information in Medicine 1993;32:317–325. [PubMed: 8412828]
49. Peleg M, Tu S, Bury J, Ciccarese P, Fox J, et al. Comparing Computer-interpretable Guideline Models: A Case-study Approach. JAMIA 2003;10:52–68. [PubMed: 12509357]
50. Protege. 2000 Available at http://protege.stanford.edu/.
51. Chamberlin DD, Boyce RF. SEQUEL: A Structured English Query Language. Proceedings of the 1974 ACM SIGFIDET Workshop on Data Description, Access and Control: Association for Computing Machinery 1974:249–264.
52. Baader, F.; Calvanese, D.; McGuinness, DL.; Nardi, D.; Patel-Schneider, PF. Theory, Implementation, Applications. Cambridge, UK: Cambridge University Press; 2003. The Description Logic Handbook. ISBN 0-521-78176-0
53. Sordo M, Boxwala A, Ogunyemi O, Greenes R. Description and status update on GELLO: a proposed standardized object-oriented expression language for clinical decision support. Stud Health Technol Inform 2004;107:164–168. [PubMed: 15360796]
54. O'Connor, MJ.; Tu, SW.; Musen, MA. AMIA Annual Symposium. San Antonio, TX: 2002. The Chronus II Temporal Database Mediator.
55. Richters M. A Precise Approach to Validating UML Models and OCL Constraints. Logos Verlag Berline. 2001
56. Current Procedural Terminology. Available at http://www.amaassn.org/ama/pub/category/3657.html.
57. Hulse R, Clark S, Jackson J, Warner H, Gardner R. Computerized medication monitoring system. Am J Hosp Pharm 1976;33:1061–1064. [PubMed: 973633]
58. McDonald CJ, Murray R, Jeris D, Bhargava B, Seeger J, et al. A computer-based record and clinical monitoring system for ambulatory care. Am J Public Health 1977;67:240–245. [PubMed: 842761]
59. The Knowledge Interchange Format (KIF). Available at http://wwwksl.stanford.edu/knowledge-sharing/kif/#manual.
60. Tu S, Musen M. Modeling data and knowledge in the EON guideline architecture. MedInfo 2001:280–284.
61. GELLO at OpenClinical. Available at http://www.openclinical.org/gmm_gello.html.
62. Medical Objects. Available at http://www.medical-objects.com.au/.
63. InferMed. Available at http://www.infermed.com/.
64. Weng, C.; Kahn, M.; Gennari, J. American Medical Informatics Association Annual Symposium. San Antonio; 2002. Temporal Knowledge Representation for Scheduling Tasks in Clinical Trial Protocols; p. 879-883.
65. Carlson R, Tu S, Lane N, Lai T, Kemper C, et al. Computer-based screening of patients with HIV/AIDS for clinical-trial eligibility. Online J Curr Clin Trials. 1995
66. Das AK, Musen MA. A temporal query system for protocol-directed decision support. Methods Inf Med 1994;33:358–370. [PubMed: 7799812]
67. Ahn, I.; Ariav, G.; Batory, D.; Clifford, J.; Dyreson, C., et al. The temporal query language TSQL2. Snodgrass, RTE., editor. Health telematics for clinical guidelines and protocols, Kluwer Academic; 1995. p. 3-15.

68. Lin, RS.; Rhee, S.; Shafer, RW.; Das, AK. A combined data mining approach for infrequent events: analyzing HIV mutation changes based on treatment history. Stanford: Computational Systems Bioinformatics; 2006.
69. Cooper, G.; Buchanan, B.; Kayaalp, M.; Saul, M.; Vries, J. Using computer modeling to help identify patient subgroups in clinical data repositories. Proc AMIA Symp; 1998. p. 180-184.
70. Hampton T. Rare Disease Research Gets Boost. JAMA 2006;295:2836–2838. [PubMed: 16804140]
71. SAGE Guideline Model Technical Specification. Available at http://sage.wherever.org/references/references.html.
72. Lindberg D, Humphreys B, McCray A. The Unified Medical Language System. Methods Inf Med 1993;32:281–291. [PubMed: 8412823]
73. Friedman, C. Towards a comprehensive medical language processing system: methods and issues. Proc AMIA Annual Fall Symposium; 1997. p. 595-599.
74. Huff SM, Rocha RA, McDonald CJ, De Moor GJE, Fiers T, et al. Development of the Logical Observation Identifier Names and Codes (LOINC) Vocabulary. Journal of Americal Medical Informatics Association 1998;5:276–292.
75. Jenders, RA.; Corman, R.; Dasgupta, B. Making the standard more standard: a data and query model for knowledge representation in the Arden syntax. AMIA Annu Symp Proc; 2003. p. 323-330.
76. Trialx. Available at http://www.trialx.com.
77. SNOMED-CT. Available at http://www.snomed.org.
78. Chisholm J. The Read clinical classification. BMJ 1990;300
79. Medical Subject Headings (MeSH). [Accessed on May 26, 2009]. Available at http://www.nlm.nih.gov/mesh/meshhome.html
80. Carter, J.; Brown, S.; Erlbaum, M.; Gregg, W.; Elkin, P., et al. Initializing the VA medication reference terminology using UMLS metathesaurus co-occurrences. Proc AMIA Symp; 2002. p. 116-120.
81. Johnson, P.; Tu, S.; Musen, M. A virtual medical record for guideline-based decision support. Proc AMIA Annu Fall Symp; 2001. p. 294-298.
82. Jenders, R.; Sujansky, W.; Broverman, C.; Chadwick, M. Towards improved knowledge sharing: assessment of the HL7 Reference Information Model to support medical logic module queries. Proc AMIA Annu Fall Symp; 1997. p. 308-312.
83. Richesson RL, Krischer J. Data Standards in Clinical Research: Gaps, Overlaps, Challenges and Future Directions. JAMIA 2007;14:687–696. [PubMed: 17712081]
84. Richesson R, Andrews J, Krischer J. Use of SNOMED CT to Represent Clinical Research Data: A Semantic Characterization of Data Items on Case Report Forms in Vasculitis Research. Journal of the American Medical Informatics Association 2006;13:536–546. [PubMed: 16799121]
85. CHI. Consolidated Health Informatics. Standards Adoption Recommendation. Functioning and Disability: Consolidated Health Informatics. 2006. 12/12/2006. Report No.: Disability Public Full.doc
86. Making the "Minimum Data Set" Compliant with Health Information Technology Standards. [Accessed on November 24, 2009]. Available at http://aspe.hhs.gov/daltcp/reports/2006/mds-hit.htm
87. Nadkarni P, Brandt C. The Common Data Elements for cancer research: remarks on functions and structure. Methods Inf Med 2006;45:594–601. [PubMed: 17149500]
88. Covitz PA, et al. caCORE: a common infrastructure for cancer informatics. Bioinformatics 2003;19:2404–2412. [PubMed: 14668224]
89. Warzel, DB. Common data element (CDE) management and deployment in clinical trials. Proc of AMIA Fall Symp; 2003. p. 1048

REFERENCE

1. Shortliffe, EH.; Scott, AC.; Bischoff, MB.; Campbell, AB.; Melle, Wv; Jacobs, CD. ONCOCIN: An expert system for oncology protocol management. Vancouver, B.C. Seventh International Joint Conference on Artificial Intelligence; 1981.

2. Kahn MG, Fagan LM, Tu SW. Extensions to the Time-Oriented Database Model to Support Temporal Reasoning in Medical Expert Systems. Methods of Information in Medicine 1989;30:4–14. [PubMed: 2005832]
3. Ohno-Machado, L.; Parra, E.; Henry, S.; Tu, S.; Musen, M. AIDS2: a decision-support tool for decreasing physicians' uncertainty regarding patient eligibility for HIV treatment protocols. Proc Annu Symp Comput Appl Med Care; 1993. p. 429-433.
4. Carlson R, Tu S, Lane N, Lai T, Kemper C, Musen M, Shortliffe E. Computer-based screening of patients with HIV/AIDS for clinical-trial eligibility. Online J Curr Clin Trials. 1995 Mar 28; Doc No 179.
5. Tu SW. A Methodology for Determining Patients' Eligibility for Clinical Trials. Methods of Information in Medicine 1993;32:317–325. [PubMed: 8412828]
6. Musen, M.; Carlson, R.; Fagan, L.; Deresinski, S.; Shortliffe, E. T-HELPER: automated support for community-based clinical research. Proc Annu Symp Comput Appl Med Care; 1992. p. 719-723.
7. Hammond P, Sergot M. Computer support for protocol-based treatment of cancer. The Journal of Logic Programming 1996;26:93–111.
8. Papaconstantinou C, Theocharous G, Mahadevan S. An expert system for assigning patients into clinical trials based on Bayesian networks. J Med Syst 1998;22:189–202. [PubMed: 9604785]
9. Rubin, D.; Gennari, J.; Srinivas, S.; Yuen, A.; Kaizer, H.; Musen, M.; Silva, J. Tool support for authoring eligibility criteria for cancer trials. Proc AMIA Symp; 1999. p. 369-373.
10. Wang, S., et al. Enhancing Arden Syntax for Clinical Trial Eligibility Criteria. Proc AMIA Symp; 1999. p. 1188
11. Ohno-Machado, L.; Wang, S.; Mar, P.; Boxwala, A. Decision support for clinical trial eligibility determination in breast cancer. Proc AMIA Symp; 1999. p. 340-344.
12. Butte, A.; Weinstein, D.; Kohane, I. Enrolling patients into clinical trials faster using RealTime Recuiting. Proc AMIA Symp; 2000. p. 111-115.
13. Séroussi B, Bouaud J, Antoine É-C. OncoDoc: a successful experiment of computer-supported guideline development and implementation in the treatment of breast cancer. Artificial Intelligence in Medicine 2001;22:43–64. [PubMed: 11259883]
14. Séroussi JB, Bouaud J, Antoine E-C, Zelek L, Spielmann M. Using OncoDoc as a computer-based eligibility screening system to improve accrual onto breast cancer clinical trials. Artificial Intelligence in Medicine 2001;29:153–167.
15. Gennari, J.; Sklar, D.; Silva, J. Cross-tool communication: From protocol authoring to eligibility determination. Proc AMIA Symp; 2001. p. 199-203.
16. Modgil S, Hammond P. Decision support tools for clinical trial design. Artificial Intelligence in Medicine 2003;27:181–200. [PubMed: 12636978]
17. Nammuni K, Pickering C, Modgil S, Montgomery A, Hammond P, Wyatt JC, Altman DG, Dunlop R, Potts HWW. Design-a-trial: a rule-based decision support system for clinical trial design. Knowledge-Based Systems 2004;17:121–129.
18. Nikiforou, S.; Fink, E.; Hall, LO.; Goldgof, DB.; Krischer, JP. Knowledge acquisition for clinical-trial selection. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics; 2002.
19. Fink E, Kokku PK, Nikiforou S, Hall LO, Goldgof DB, Krischer JP. Selection of patients for clinical trials: an interactive web-based system. Artificial Intelligence in Medicine 2004;31:241–254. [PubMed: 15302090]
20. Fink, E.; Hall, LO.; Goldgof, DB.; Goswami, BD.; Boonstra, M.; Krischer, JP. Experiments on the automated selection of patients for clinical trials. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics; 2003. p. 4541-4545.
21. Sim I, Olasov B, Carini S. An ontology of randomized controlled trials for evidence-based practice: content specification and evaluation using the competency decomposition method. Journal of Biomedical Informatics 2004;37:108–119. [PubMed: 15120657]
22. Metz MJ, Coyle C, Hudson C, Hampshire M. An Internet-Based Cancer Clinical Trials Matching Resource. Journal of Medical Internet Research 2005;7(3):e24. [PubMed: 15998615]
23. Cohen, Eea. caMATCH: A Patient Matching Tool for Clinical Trials. caBIG Annual Meeting; 2005.

24. Niland J. ASPIRE: Agreement on Standardized Protocol Inclusion Requirements for Eligibility. 2007 an unpublished web resource.
25. Sutton DR, Fox J. The Syntax and Semantics of the PROforma Guideline Modeling Language. JAMIA 2003;10:433–443. [PubMed: 12807812]
26. Fox J, Johns N, Lyons C, Rahmanzadeh A, Thomson R, Wilson P. PROforma: a general technology for clinical decision support systems. Comput Methods Programs Biomed 1997;54:59–67. [PubMed: 9290920]
27. Shahar Y, Miksch S, Johnson P. The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. Artificial Intelligence in Medicine 1998;14:29–51. [PubMed: 9779882]
28. Tu, S.; Musen, M. A flexible approach to guideline modeling. Proc AMIA Symp; 1999. p. 420-424.
29. Tu S, Musen M. Modeling data and knowledge in the EON guideline architecture. MedInfo 2001:280–284.
30. Johnson, P.; Tu, S.; Booth, N.; Sugden, B.; Purves, I. Using scenarios in chronic disease management guidelines for primary care. Proc AMIA Symp; 2000. p. 389-393.
31. Quaglini S, Stefanelli M, Lanzola G, Caporusso V, Panzarasa S. Flexible guideline-based patient careflow systems. Artificial Intelligence in Medicine 2001;22:65–80. [PubMed: 11259884]
32. Quaglini S, Stefanelli M, Cavallini A, Micieli G, Fassino C, Mossa C. Guideline-based careflow systems. Artificial Intelligence in Medicine 2000;20:5–22. [PubMed: 11185420]
33. Boxwala A. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. Journal of Biomedical Informatics 2004;37:147–161. [PubMed: 15196480]
34. Tu SW, Campbell JR, Glasgow J, Nyman MA, McClure R, McClay J, Parker C, Hrabak KM, Berg D, Weida T, Mansfield JG, Musen MA, Abarbanel RM. The SAGE Guideline Model: Achievements and Overview. JAMIA 2007;14:589–598. [PubMed: 17600098]
35. Patel, C.; Cimino, J. Using Ontology Reasoning to March Electronic Patient Records to Clinical Trials. Washington DC. Proc of AMIA'06; 2006.
36. Patel, C.; Cimino, J. Semantic Query Generation from Eligibility Criteria in Clinical Trials. AMIA Annu Symp Proc; 2007.
37. Patel C, Cimino J, Dolby J, Fokoue A, Kalyanpur A, Kershenbaum A, Ma L, Schonberg E, Scrinivas K. Matching Patient Records to Clinical Trials Using Ontologies. IBM Thomas J Watson Research Center. 2007
38. Patel, C.; Cimino, J.; Dolby, J.; Fokoue, A.; Kalyanpur, A.; Kershenbaum, A.; Ma, L.; Schonberg, E.; Srinivas, K. Matching Patient Records to Clinical Trials Using Ontologies. Proc. of International Semantic Web Conference; 2007.
39. Olasav, B.; Sim, I. RuleEd, a Web-based Semantic Network Interface for Constructing and Revising Computable Eligibility Rules American Medical Informatics Annual Symposium; November 11–15, 2006; Washington, D.C. 2006. p. 1051
40. Lonsdale DW, Tustison C, Parker CG, Embley DW. Assessing clinical trial eligibility with logic expression queries. Data & Knowledge Engineering 2007;66(1):3–17.
41. Jenders, R.; Sujansky, W.; Broverman, C.; Chadwick, M. Towards improved knowledge sharing: assessment of the HL7 Reference Information Model to support medical logic module queries. Proc AMIA Annu Fall Symp; 1997. p. 308-312.
42. Clinical Research Filtered Query. Available at http://hssp-cohort.wikispaces.com/space/showimage/SFM_CRFQ_v1.0.doc

Appendix

J Biomed Inform. Author manuscript; available in PMC 2011 June 1.

Appendix Table 1

Knowledge Representation Primitives for Clinical Research Eligibility Criteria (Patient Information Model, Clinical Terminologies, Expression Language, and Their Standardization Status)

Cita	Brand	Categorizations of criteria	Common Data Elements	Expression Language OR Query Syntax	Representations of Patient Data	Representations of Medical concepts	Use Cases	Domain
[1, 2] 1981	ONCOCIN	No	No	Parameters-Value Representation: parameters represent the attributes of patients, drugs, tests, etc. Implemented in **Interlisp**, which provides symbol manipulation capabilities. There two versions of ONCOCIN, a main frame and a workstation versions. The workstation version has a temporal data representation and query language. [2]	The original patient database in mainframe-ONCOCIN used the time-oriented databank (TOD) model, where each value has an associated time stamp that denotes when that value was observed. The limitation of this method is that it does not appreciate meaningful groupings of data. The new version's ETNET provides date-free, context-sensitive data storage and retrieval.	No	Protocol-based patient management	Cancer
[3] 1993	AIDS2	Criteria are classified into three categories: history, examination, and tests based on importance in determining eligibility status for the protocol, on risk to the patient, and on cost, which includes the cost of the procedure and the clinician's time.	No	**Three-levels of KR** medical concept level to represent classes, probabilistic belief network level to represent uncertainty about criteria, and control level to represent procedural knowledge.	No	No	Identifying eligible patients for trials	HIV
[4–6] 1995	T-Helper	Criteria are classified into five groups by objectiveness, variability, and controllability of the underlying clinical conditions: stable requisite, variable routine, controllable, subjective, and special. (pg 4 in Samson's paper)	No	Eligibility criteria were represented in as instances of structured templates whose syntax allowing simple comparison, arithmetic combinations, arbitrary conjunctions and disjunctions; Each type of criteria has templates	A simple model that defines time-stamped parameter values and interval-based events that has attributes and values	No	Identify eligible patients for trials	Generic

Cita	Brand	Categorizations of criteria	Common Data Elements	Expression Language OR Query Syntax	Representations of Patient Data	Representations of Medical concepts	Use Cases	Domain
				used for translating criteria to database queries.				
[7] 1996	OaSiS	No	No	Prolog-based representation For parameters-Values Predicates. (Rule-based) Support for time-point-based temporal arguments, including both relative and absolute descriptions of time points.	No	No	Protocol-based patient management	Cancer/ Oncology
[8] 1997	Unnamed	Criteria were divided into two classes: one is static criteria (diagnosis, age, sex), and the other is temporal criteria (lab test, subjective evaluation, symptoms).	No	Bayesian Network	No	No	For filtering protocols based on patient characteristics	Generic
[9] 1999	Unnamed	Criteria are categorized by “clinical states” of significance in cancer domain	Not sure	Criteria are represented in frame-based systems, where “Enumerated” Data Types are supported so that variables can only take a value from a predefined list.	No	No external standards, but internally used standard naming.	, Criteria representation for clinical trial authoring	Cancer
[10], [11] 1999	Unnamed	No	No	Enhanced Arden Syntax (1) “Enumerated” Data Types, where variables can take on a value from a predefined list (2) “Struct” data type and a corresponding “dot operator” that allow us to assign specific attributes to each variable or concept. (3) primary time stamp associated with data variables; start and end time to indicate intervals (4)clinical vocabulary terms were selected from UMLS and synonyms were mapped	No	UMLS	For representing eligibility criteria with a general syntax; for filtering protocols based on patient characteristics.	Breast Cancer

Cita	Brand	Categorizations of criteria	Common Data Elements	Expression Language OR Query Syntax	Representations of Patient Data	Representations of Medical concepts	Use Cases	Domain
[12] 2000	Unnamed	No	No	Database query of selected lab values	No	No	Recruiting patients to trials	Hypoglycemia
[13, 14] 2001	OncoDoc	No	No	Decision tree, where decision parameters aim to describe patient states.	No	No	Identify the best trial for a patient	Cancer
[15] 2001	EligWriter	Clinical states, similar to Rubin's work	NCI's CDE	Ad hoc logical expressions	NO	NO	Eligibility determination for protocols	Cancer
[16], [17] 2003	Design-a-Trial	No	Not sure, no specification of criteria in papers	Textual representation	No	No	Representation of XML-based clinical trials, but not computer-executable clinical trials protocols	Generic
[18], [19], [20] 2002 – 2004	Unnamed	The system supports three types of questions: the first type takes yes/no response, the second is multiple choice, and the third requires a numeric answer.	No	The knowledge base contains questions, tests, and logical expressions that represent eligibility for each trial. Logical expressions that can represent equalities, inequalities, set-element relations, conjunctions, and disjunctions. Variables are defined to represent patient data. The description of a medical test includes its dollar cost and list of questions that can be answered based on test results. Tests are ordered by their cost and among of information they provided.	No; data will be entered by users	No	To reduce the cost of eligibility screening by optimizing test ordering; To filter trials or protocols for patients through interactive user interface.	Cancer
[21]	Trial Bank	Base-rule, Recursive-rule	No	A rule-based representation that supports logical reasoning over AND, OR, and recursion.	No	No	Formal methods for representing clinical trials protocols	Generic

Cita	Brand	Categorizations of criteria	Common Data Elements	Expression Language OR Query Syntax	Representations of Patient Data	Representations of Medical concepts	Use Cases	Domain
[22] 2005	OncoLink	Questions are categorized as demographics, contact information, personal medical history, cancer diagnosis, and treatments o date.	Demographics, contact information, personal medical history, cancer diagnosis, and treatments to date.	Web-based questionnaire "Enumerated" Data Types are supported	No	No	Interactive Trials search services (filtering protocols); no automatic patient matching	Cancer
[23] 2007	caMatch	Criteria are categorized by diseases; Vocabulary-driven data entry (CDE's from caDSR)	work with HL7, CDISC, and OMG	Provides a structured form compliant with HL7 structured protocol representation effort. For patient to enter eligibility information; Representation takes parameter-value form: for each data element, there are a list of acceptable values	Personal health record compliant with HL7;	No	Interactive Trials search services (filtering protocols); no automatic patient matching	Disease specific; data collection and matching rules are customized to breast cancer; other cancer in the future
[24] 2007	ASPIRE	Criteria are categorized by diseases	Will use CDISC elements when defined; For breast cancer criteria, using NCI elements from caDSR	Representation takes parameter-value form for each data element, for which there are a list of acceptable values. Uses HL7 Clinical Research Functional Query	No	Will use CDISC terminology when defined	Interactive Trials search services (filtering protocols); no automatic patient matching	Disease-specific: pan-disease
[25, 26] 1996	PROforma	1. presence criteria Peleg 03 JAMIA page 61 2. templated-based criteria that look for qualitative and quantitative observations, medications, and other types of EMR entities. Pg 61: can	No.	First-order logic language	No	No	Generic Computer-based Clinical Guideline Representations	Generic
[27] 1998	Asbru			XML-based criterion language that	Parameter + value pairs, with	No		

Cita	Brand	Categorizations of criteria	Common Data Elements	Expression Language OR Query Syntax	Representations of Patient Data	Representations of Medical concepts	Use Cases	Domain
		declaratively express simple temporal constraints on these entities of the form within an interval of a time point. 3. first-order logic criteria Pg 61 4. temporal criteria Difference in TC: Asbru and EON supports temporal abstraction, while GLIF supports temporal operators of the Arden Syntax logic grammar		allows specification of value set or ranges, context, and temporal extent of parameters.	context and time annotation			
[28, 29] 2001	EON			EON provides **three expression** languages: templates including objects with certain attributes, first-order logical language, and temporal query language	VMR	No		
[30] 2000	PRODIGY			PRODIGY: templates	VMR	READ code		
[31, 32] 2001	GUIDE			SQL that supports some first-order logic criteria	An embedded health record model	UMLS & SNOMED		
[33] 2004	GLIF3		No	GELLO: an expression language that defines the syntax and semantics of the decision criteria. It is defined by HL7 and CDISC.	HL7 RIM (Reference Information Model)	UMLS		
[34] 2007	SAGE	A rich set of criteria templates, specified in page 46. Examples include N_ary_criterion, goal_criterion, comparison_criterion, temporal_comparison_criterion, variable_comparison_criterion, adverse_reaction_presence_criterion, presence_criterion,	No	Structured templates (see Categorization of criteria) that can be translated into GELLO	VMR based on HL7 RIM	SNOMEDCT, LOINC, NDF-RT		
[35–38] 2007	Unnamed	N/A	N/A	Web Ontology Language (OWL)	VMR	SNOMED-CT	Automated mass-screening of patients for selected protocols through automatic query translation and expansion	Generic

Cita	Brand	Categorizations of criteria	Common Data Elements	Expression Language OR Query Syntax	Representations of Patient Data	Representations of Medical concepts	Use Cases	Domain
[39] 2007	ERGO (Eligibility Rule Grammar and Ontology)	Categorized by rules about participant properties, interventions done on participants, and participant behaviors	No	Structured templates that model noun phrases, expressions, and criteria; with temporal and other modifications, and semantic relations, can be recursively composed	VMR	UMLS SNOMED MeSH	Representation of computable criteria for eligibility matching	Generic
[40] [41] 2007	Unnamed	No	HL7 Virtual Medical Records (VMR)	Arden Syntax Medical Logical Module (MLM); each database query is a VMR query (earlier it was Data Access Modules (DAM) and MED. Not the most expressive language in our set. Primarily concept mapping; no consideration for temporal logic.	VMR, specifically attributes in class “Observations”	Intermountain Healthcare’s Healthcare Data Dictionary	(1) Identifying patients for selected protocols; (2) automatic query formulation	Generic; Identify eligible patients;
[42] 2007	CRFQ	Standardizing the Parameters using Semantic Signifiers With these parameter types. 1. Demographic Data 2. Patient Disease Historical Data 3.Disease MetaData 4. Disease Data 5. Protol Listings 6. I_E Criteria 7. Patient Preference Data	ASPIRE as the source for core data elements	Structured representation by HL7	HL7 V3 Data Types		(1) interoperability of all applications using computer-based criteria: (2) filtering protocols or (3) identify eligible patients	Disease-specific