Poster PP31

Automating Study Eligibility Determination using the Resource Description Framework (RDF)

Armando Oliva, M.D., Semantica LLC, Fort Lauderdale, USA

ABSTRACT

A clinical study is fundamentally a collection of activities that are performed according to protocol-specified rules. Managing this workflow is a formidable challenge and it is not surprising that errors occur. One of the most important study activities is the determination of Eligibility for further continuation in the study.

A previous paper presented at the 2017 PhuSE European Conference in Edinburgh documents the use of the Resource Description Framework (RDF) in automating a study workflow. That paper describes how protocol information can be expressed in the RDF in a computable way, such that information systems can identify which activities have been performed and determine which activities should be performed next. This follow-up paper explores how RDF can similarly be used to automate Eligibility determination, which is, for many clinical trials, the protocol specified start rule of the Randomization activity. The use of this approach has the potential to improve how studies are conducted, resulting in better compliance and better data.

INTRODUCTION

A clinical study is fundamentally a collection of activities that are performed according to specific rules as described in the protocol. These rules describe not only when activities are conducted, but also how they are grouped in logical study components such as Visits and Elements and Epochs. It is not unusual for a single subject to undergo hundreds of study related activities. Managing this workflow is a formidable challenge. The investigator must ensure that all activities are conducted at the right time and in the correct sequence, sometimes skipping activities that logically need not be done, and ensure that the data for each activity are properly documented. It is not surprising that errors occur, which may result in protocol violations that risk the interpretability of the study results. Two of the most important activities in a study is eligibility determination and randomization. For most studies, randomization occurs only after it is demonstrated that eligibility criteria have been met. It is not unusual for subjects who fail to meet eligibility criteria to undergo randomization inappropriately. The converse may also be true, i.e. subjects who meet eligibility criteria might not be randomized.

The Resource Description Framework (RDF) is a standard developed by the World Wide Web Consortium (W3C) to represent data and metadata along with their meaning in a rich Linked Data environment. RDF has three distinct advantages over other standards, which make it desirable for testing as a new format for representing study data:

1. The RDF model, once designed, is immediately implementable. The model is computable "out of the box"
2. RDF supports linking multiple standards together in an automated fashion, significantly simplifying implementation
3. The model can be updated incrementally as new requirements emerge yet can maintain a high degree of backwards compatibility with previous versions.

Representing a study workflow in a computable format has many advantages, not the least of which is the ability to automate an analysis of how well a study was conducted, how well the protocol was followed, and the automated identification of protocol violators.

This paper explores the use of the RDF and related standards to automate the determination of eligibility to continue in the study, which for many studies means undergoing randomization and treatment with the medical intervention under study. One of the RDF-related standards is SPIN (SPARQL Inference Notation), which allows rules, such as activity start rules, to be expressed computationally so that commercially available off the shelf (COTS) software such as RDF inferencing engines can evaluate the rules and determine which activities go next and which can logically be skipped.

This proof of concept is a spin-off of an ongoing project within PhUSE called Going Translational with Linked Data (GoTWLD) [1]. In that project, the project team has created a mini study ontology in OWL [2] (see Figure 1) that is capable of representing a portion of the subject level data from a pilot study in RDF [3], and from which highly standards-conforming SDTM [4] datasets can automatically be generated. The study ontology from that project is leveraged in this work, which can be obtained via GitHub. [5]

REFERENCES

[1]http://www.phusewiki.org/wiki/index.php?title=Clinical_Trials_Data_as_RDF

[2]Web Ontology Language, another standard related to RDF. See https://www.w3.org/OWL/ .

[3] See https://github.com/phuse-org/CTDasRDF. The mini study ontology is available at https://raw.githubusercontent.com/phuse-org/CTDasRDF/master/data/rdf/study.ttl

[4] Study Data Tabulation Model, see http://www.cdisc.org/sdtm

[5] See https://github.com/phuse-org/CTDasRDF

[6] http://www.semanticallc.com/files/share/Oliva_SI08_StudyWorkflowInRDF.pdf

STRATEGY

We refer the reader to the previous paper "Managing Study Workflow Using the Resource Description Framework" for a detailed discussion on modeling study activity start rules. A brief summary is provided here, along with some important modeling improvements that have evolved since that paper was presented in 2017.

Study activities are quite varied, but their start rules are remarkably similar. A typical start rule Activity(t) (the "target" activity) may say:

Begin Activity(t) when Activity(p) (the "prerequisite" activity) is complete and has a specific prerequisite outcome (PO).

A rule with this basic structure is considered a "simple start rule." The rule may also define other constraints, such as a maximum time interval following completion of P past which T should not be done, or a delay interval that must elapse following the completion of P before T can begin. As an example, a screening activity may be an RPR Test for syphilis. The rule may say:

Begin RPR Test (T) when Informed Consent (P) is complete and the outcome (PO) is Informed Consent-Granted

The rule can be expressed in the RDF using SPIN (SPARQL Inference Notation). An RDF inferencing engine can evaluate the rule as TRUE (rule condition met) or FALSE (not met), or NOT-APPLICABLE (i.e. the activity can be skipped without violating the protocol). If the rule outcome is TRUE then T becomes an enabled activity, which makes it eligible for execution.

Start Rules can be joined using Boolean operators to achieve composite start rules. In this case, T may begin if {StartRule(p1) AND StartRule(p2) AND StartRule(p3)} are all true (or not applicable). The composite start rule p4 therefore consists of 3 simple start rules, all of which must be true (or not-applicable) for p4 to be true, and allow p4 to begin.

Study Eligibility Determination can be described and automated using this strategy if one recognizes that the Eligibility Determination Activity is in fact the start rule for the Randomization activity. In studies where there is no randomization, it becomes the Start Rule for continuation in the study (e.g. starting an open-label treatment epoch).

This paper also proposes a new modeling strategy for addressing study activities that can be logically skipped in some subjects, but not others, without triggering a protocol violation. The classic example is a serum pregnancy test that is performed if the subject is female, but may be logically skipped if the subject is male. The start rule for the Pregnancy Test activity says:

Begin Pregnancy Test (T) when Informed Consent (P) is complete and the outcome (PO) is Informed Consent-Granted

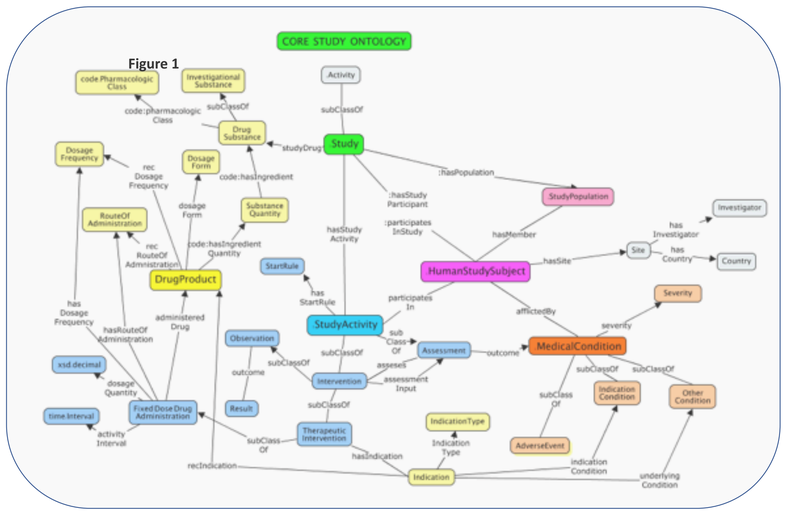
This is the same approach used previously for the RPR Test activity start rule. How do we express that the activity may be logically skipped if sex = male? We propose that, because Start Rules themselves are study activities (i.e. a type of Analysis), then a start rule can have a start rule!! The plain English representation would be something like this:

PregnancyTestStartRule = Begin Pregnancy Test (T) if Informed Consent is Granted

StartRule for (PregnancyTestStartRule) = Begin PregnancyTestStartRule if SexDocumentation Outcome = Female

In this example, the system won't even determine whether a Pregnancy Test can begin if the sex = male, which is precisely the desired approach. If the outcome of the "Start Rule of the Start Rule" is 0 (do not begin), then the system can logically assign the outcome of the Start Rule as "not-applicable", which in turn can apply the logically skipped activity flag to the PregnancyTest itself.

Figure 1



METHOD

STUDY DESIGN

To test the strategy, a simple notional study was designed. The mini study ontology was extended to accommodate activity rules and their processing. Notional study data were generated in RDF.

The notional study has a Screening Epoch consisting of these activities:

- Obtain Informed Consent
- Document Demographic (i.e. Sex) Information
- Obtain Pregnancy Test on Female Subjects Only
- Perform RPR Test for Syphilis on all Subjects

And a Randomization Epoch consisting of one activity:

- Randomization

Obtain Informed Consent and Randomization activities are each modeled as an Administrative Activity. The remaining three activities are each modeled as an Observation (Sex Determination, Pregnancy Test, RPR Test).

RULES

Any Activity that takes as input information about other activities and generates as outcome a result is considered an Analysis in the ontology; a subclass of Activity. Rules fall in this category. A Rule is therefore modeled as a subclass of Analysis. For example:

Rule: Sex must be Female
RuleOutcome: 0 (False, Sex is not Female)
1 (True, Sex is Female)

This rule looks at the outcome of the "Document Demographic (Sex) Information" activity, the prerequisite activity, and returns a value of TRUE if the prerequisite activity outcome is "F" for Female.

Rules may be simple or composite. A simple rule looks at information about one prerequisite activity. A composite rule looks at information about two or more prerequisite activities.

Some simple rules are met if the prerequisite activity is simply completed and a certain outcome is reported. Other simple rules are met if the prerequisite activity is simply completed, regardless of the outcome. Yet other rules are met if the prerequisite activity is merely started. Finally, there is the default rule which automatically has a value of TRUE. Any activity with a start rule of the default rule can automatically begin.

Composite Rules contain two or more simple rules as subrules. The algorithm for resolving a composite rule is described in detail in the 2017 paper [6]. There is a special case of a composite start rule, and that is one that determines when the Randomization Activity can begin. This start rule is in fact the Eligibility Determination Test.

For the pilot study, one can associate each activity with the following start rules.

Obtain Informed Consent	Default Start Rule
Document Demographic (Sex) Information	Rule_Informed Consent Granted
Obtain Pregnancy Test	Rule_Informed Consent Granted
PregnancyTestStartRule	Rule SexDocumentation = F
Obtain RPR Test	Rule_Informed Consent Grand
Eligibility Determination	CompositeRule ScreeningObservationsComplete
Randomization Activity	CompositeRule ScreeningObservationOutcomes are as expected.

The rules are expressed using SPIN so that inferencing can automatically assign the rule outcome.

To understand how composite rules are evaluated, we refer the reader to the 2017 Paper previously described [6].

CONTACT INFORMATION

Armando Oliva, M.D.
Semantica LLC
aoliva@semanticallc.com
www.linkedin.com/in/aolivamd
+1 954-489-8645
+1 202-330-4554 (fax)
www.semanticallc.com
aolivamd.blogspot.com

ELIGIBILITY DETERMINATION

With this ontology in place, one can now populate a knowledgebase with instance data. One can envision a notional Electronic Data Capture (EDC) study workflow management application using an RDF inferencing engine that can continuously monitor the subject data as they come in and evaluates the start rule for each activity. When the start rule evaluates as TRUE, a separate SPIN rule identifies the target activity as an Enabled Activity, which means it can be scheduled for execution, according to any predefined lag interval described by the rule.

The Eligibility Determination Activity is the Start Rule for the Randomization Activity in our Study, but before it can be executed, it has a composite start rule that determines if the correct Screening Tests have been completed. Once this is complete, then Eligibility Determination can begin. The Eligibility Determination Activity checks to make sure that all relevant eligibility criteria have been met and the outcome is TRUE and Randomization can begin, or not met and the outcome is FALSE (Randomization cannot begin)

Table 1: Study Raw Data

Subject	Informed Consent	Sex	RPR	Pregnancy
Person_1	GRANTED	F	NEG	NEG
Person_2	GRANTED	F	NEG	POS
Person_3	GRANTED	F	POS	NEG
Person_4	GRANTED	F	NEG	<null>
Person_5	GRANTED	M	<null>	N/A
Person_6	GRANTED	M	<null>	N/A
Person_7	GRANTED	M	POS	N/A
Person_8	GRANTED	M	NEG	N/A

RESULTS

A COTS inferencing engine asserted the following:

Enabled Activities: PregnancyTest (#4), RPR Test (#5,6), Randomization (#1, #8) i.e. these passed eligibility Testing

Logically Skipped Activities: PregnancyTest (#5-8, all male)

The StartRule for the Eligibility Determination Activity correctly identified 5 subjects that were able to undergo Eligibility Determination. (#1-3 and 7-8)

CONCLUSION

This paper describes how the RDF can be used to automate Eligibility Determination activities. The notional study, containing just a handful of activities, successfully demonstrates the validity of this approach when the study data and rules are expressed in RDF. Additional work is needed to fully understand the benefits of this approach as described as next steps below.

Eligibility criteria can be considered a composite start rule for the randomization activity (T). The Randomization Start Rule (which is the same as the Eligibility Determination Test), examines all the screening activities (P) and their outcomes (PO) to determine if the subject can proceed to the Randomization activity. The approach can also be used for adaptive trial designs, where the Enabled Activity may differ depending on the outcome of a previous activity, and studies where treatment is delayed until a "trigger" (e.g. headache observation) occurs.

This paradigm unlocks the information in a protocol document in a way that makes it computable so information systems can "understand" a study workflow. It has the potential to improve how studies are conducted, resulting in better compliance, fewer protocol violations, and better data.

NEXT STEPS

It would be useful to apply these techniques to an actual study to see if automated Eligibility Determination really works. The major challenge is the way most protocols are currently written, the start rules are often embedded in unstructured text, or have to be inferred based on unclear or ambiguous language.