Paper RE01

The role of FHIR Resources in testing and validating an EHR to Sponsor data pipeline.

Patrick Genyn, fhir4pharma, Pennsylvania, USA
Andy Richardson, fhir4pharma, Hungerford, UK

ABSTRACT

Collecting study data directly from EHRs into clinical trial databases is becoming more and more promising as sites can offer access to subject records using FHIR methodologies. In a data pipeline from EHR to a clinical database (e.g., eDC) the determination of what specific data is required in support of the clinical trial protocol is automated and best programming practices require the automation to be thoroughly tested and validated. This requirement splits up into two main parts: 1. How to represent the trial protocol as FHIR resources and 2. How to collect the right and accurate data from the eHR as FHIR resources. The operational challenges in testing a data pipeline in a timely and scalable manner will be discussed.

INTRODUCTION

Direct data capture is designed to reduce the burden of transcribing EHR health records to EDC systems. The replacement of a qualified person (site staff) in the transfer process removes them from confirming at every point that the data to be provided to a study complies with the protocol requirements. FHIR is now the common interoperability standard for retrieving data from EHRs and supports the set-up of a data pipeline for direct data capture in three steps, 1. Study implementation, representing the study as FHIR resources (e.g., Schedule of Activities, windowing, …), 2. Site configuration, representing the research subjects, encounters, observations, etc. as FHIR resources and 3. Integration and mapping, creating the actual transfer between site and sponsor.

MODEL AND IMPLEMENTATION

THE DIRECT DATA CAPTURE MODEL

FHIR Resources can be used to define study requirements. The figure below shows how the FHIR standards can be used to both define study requirements 'Outgoing FHIR' [from a sponsor's perspective] as well as 'Incoming FHIR' the EHR data. If the 'Incoming FHIR' data are consistent with the 'Outgoing FHIR' specifications automated transfer can proceed without intervention. If not the difference in meaning – the semantics – must be resolved first.

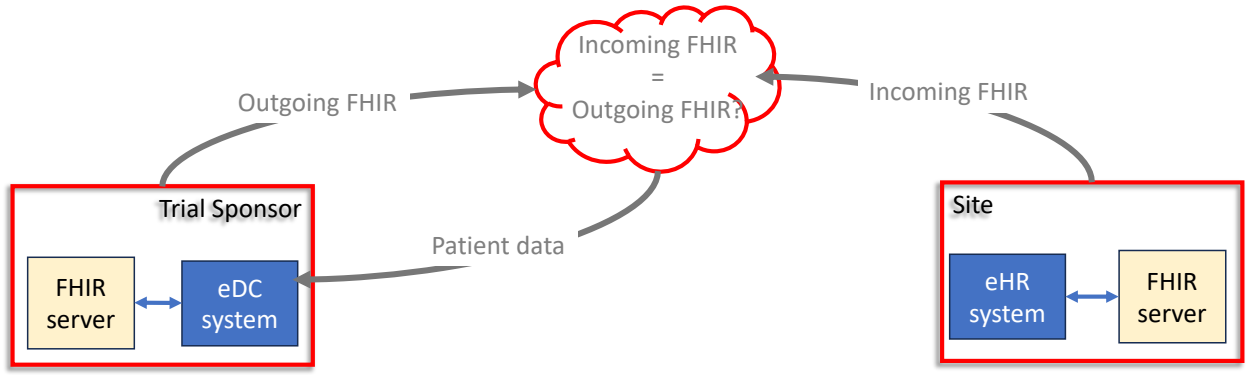

Figure 1. The direct data capture model

DIRECT DATA CAPTURE IMPLEMENTATION

Implementation of this model typically requires a third party, i.e. a data broker with specific responsibilities such as security, data privacy implementation, audit trails and provenance etc … The figure below shows the three steps involved in implementing a data pipeline supporting the direct data capture model. In the first step, the protocol is represented as a set of FHIR resources. In particular, resources such as PlanDefinition, ActivityDefinition, ObservationDefinition, SpecimenDefinition etc … are used to represent the schedule of activities and related

information from the protocol. In the second step, the resources made available by the research site are listed. Resources such as ResearchSubject (Patient), Appointment, Observation, Specimen and MedicationAdministration are reviewed with specific attention to the used standardized systems for identifying and exchanging health information (LOINC, SNOMED, …) and the version of the FHIR Resources (R2, R3, R4, R4B, R5). In the last step, the protocol specifications as FHIR resources (outgoing FHIR) and the research subject information as FHIR resources (Incoming FHIR) are integrated and mapped. If the 'Incoming FHIR' data are consistent with the 'Outgoing FHIR' specifications automated transfer can proceed without intervention. If not the difference in meaning – the semantics – must be resolved. In practice this is usually, but not always, an issue of coding equivalence.

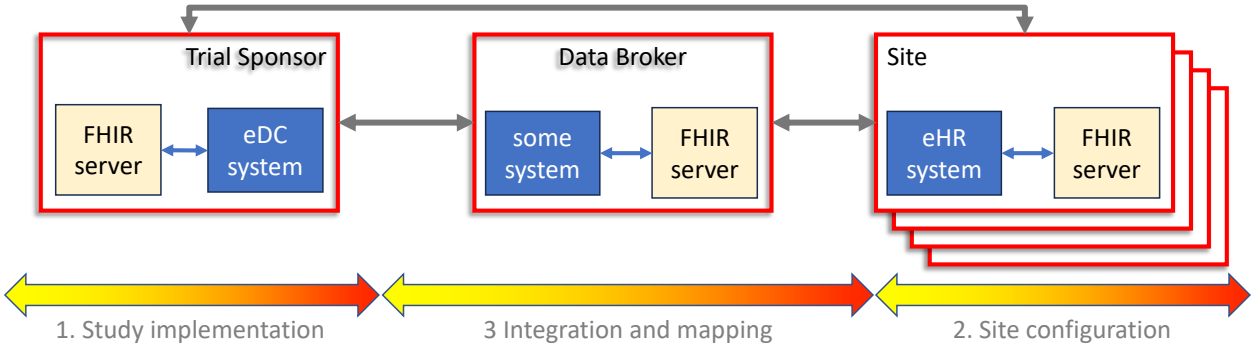

Figure 2. Direct data capture implementation

TECHNICAL IMPLEMENTATION

At FHIR4Pharma,the direct data capture model is simulated using the following tools. FHIR servers are simulated using the Meld Sandbox and Logica Sandbox. Various tools are developed in python to implement the direct data capture pipeline. These tools are maintained in python notebooks and Visual Studio Code. NetworkX python package is used for the creation and manipulation of schedules of activities as graphs. yEd graph editor from yWorks and Mermaid Gitgraph Diagrams are used to visualize the NetworkX graphs.

THE CHALLENGES AND SOLUTIONS

Electronic Data Capture (eDC), the current standard to collect clinical research data, is very effective. An electronic case report form (eCRF) is developed, validated, and implemented for all sites in a clinical trial. However, with Direct Data Capture (DDC) additional dimensions are introduced. The replacement of qualified staff at the research site who interpret the source data before entering the required data in the eCRF, complicates the automation of direct data capture. To collect the data, meeting the protocol requirements, from the research site requires the implementation of a semantic equivalence tool in the pipeline. On top, research sites can expose data using different resources and standards. Research site can implement different FHIR versions. FHIR version 5.0.0 (R5) contains more than 150 resources but only 11 resources have reached the normative state.

SEMANTIC EQUIVALENCE (REF. 2)

The schedule of activities and related information of the protocol is specified using FHIR resources. Using the example of an ObservationDefinition, a resource that defines the requirements an observation from the research site needs to adhere to. The figure below shows a part of the ObservationDefinition for diastolic blood pressure. The study specification (outgoing FHIR) expects the LOINC code "8462-4" with the unit "mm[Hg]"

Observation Definition (study spec)

* title = "Diastolic blood pressure"
* status = "active"
* code.coding.version = "2.76"
* code.coding = $loinc#8462-4
* permittedDataType = #Quantity
* permittedUnit.system = "https://ucum.org"
* permittedUnit.code = "mm[Hg]"
* permittedUnit.display = "mm[Hg]"

Figure 3. Diastolic Blood Pressure, initial specification

Take the following scenario: During the site configuration step of one or more research sites, the diastolic blood pressure is expressed as an observation with a different LOINC code “8453-3” (Diastolic blood pressure—sitting). After review, this code can be accepted as well in the trial at hand and needs to be added to the ObservationDefintion. Another example is the unit. One or more sites express the diastolic blood pressure using a “cm[Hg]”. Again, after review, this unit can be accepted as well. During the integration and mapping step, the unit can be easily converted to “mm[Hg]”. Eventually the ObservationDefinition is completed and is shown in the figure below:

Observation Definition (study spec)

* title = "Diastolic blood pressure"
* status = "active"
* code.coding[0].version = "2.76"
* code.coding[=] = $loinc#8462-4
* code.coding[+].version = "2.76"
* code.coding[=] = $loinc#8453-3
* permittedDataType = #Quantity
* permittedUnit[0].system = "https://ucum.org"
* permittedUnit[=].code = "mm[Hg]"
* permittedUnit[=].display = "mm[Hg]"
* permittedUnit[+].system = "https://ucum.org"
* permittedUnit[=].code = "cm[Hg]"
* permittedUnit[=].display = "cm[Hg]"

Figure 4, Diastolic Blood Pressure, final specification

RESEARCH SUBJECT PATHS THROUGH A TRIAL

The schedule of activities with other relevant information from the protocol defines the path of a research subject through a trial. Taking screening failures, early termination and unscheduled visits into account, the number of paths grows very quickly. In the figure below, all the paths through a trial with 7 visits are depicted.

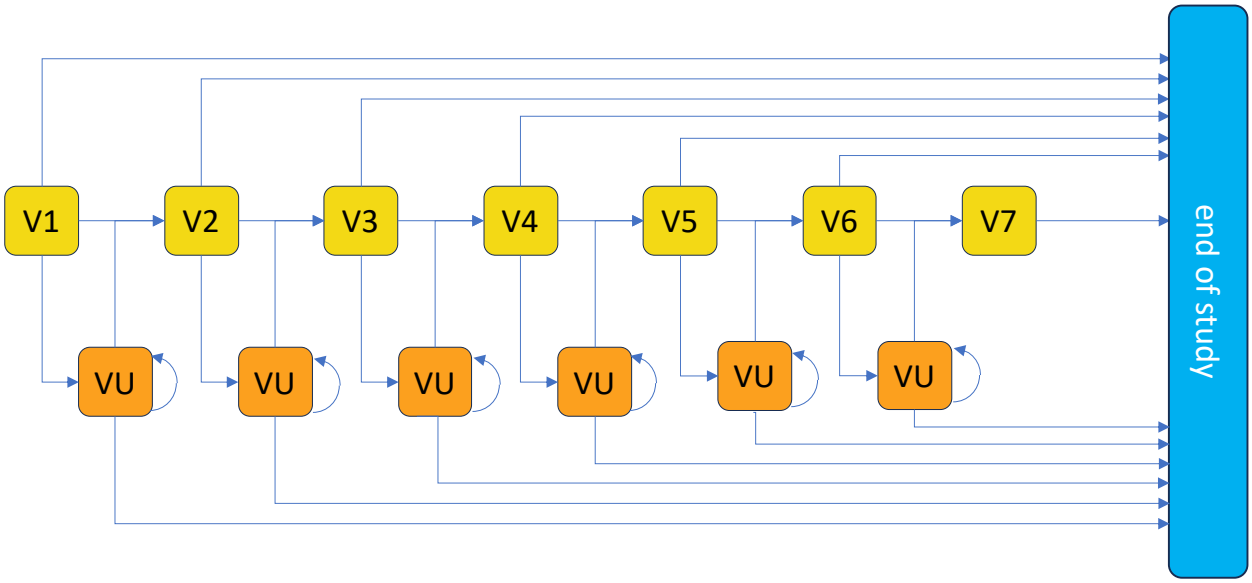

Figure 5. Paths through a trial (VU is an unscheduled visit).

The critical question is “what paths should be included in the testing of the direct data capture pipeline?”. Mathematically, there are infinite paths through the trial because a research subject can be stuck in a perpetuum unscheduled visit which of course is impossible in the real world. However, using a standard algorithm to predict the most probable paths, preferably using datasets such as trial visits (TV), Subject Visits (SV), disposition (DS) from previous similar studies, can identify the probabilities of each path. Accordingly, the research subject schedules can be generated with ActivityDefinition FHIR resources with unscheduled visits for some. Next the the related Appointment, Encounter, Observation, MedicationAdministration, … FHIR resources with test data can be generated.

WINDOWING

The trial visit timings describe the planned schedule to meet protocols research objectives. However, practical considerations at the research site or for the research subject require some flexibility to accommodate visits taking place on different times than the planned visits. Availability of services at the research site or visits falling on weekends or holidays are examples for needing flexibility.

The picture below depicts the principle behind the windowing algorithm. The algorithm for calculating the windows can be quite complex. The Vulcan project Schedule of Activities (ref. 3) addresses this point with a lot of detail.

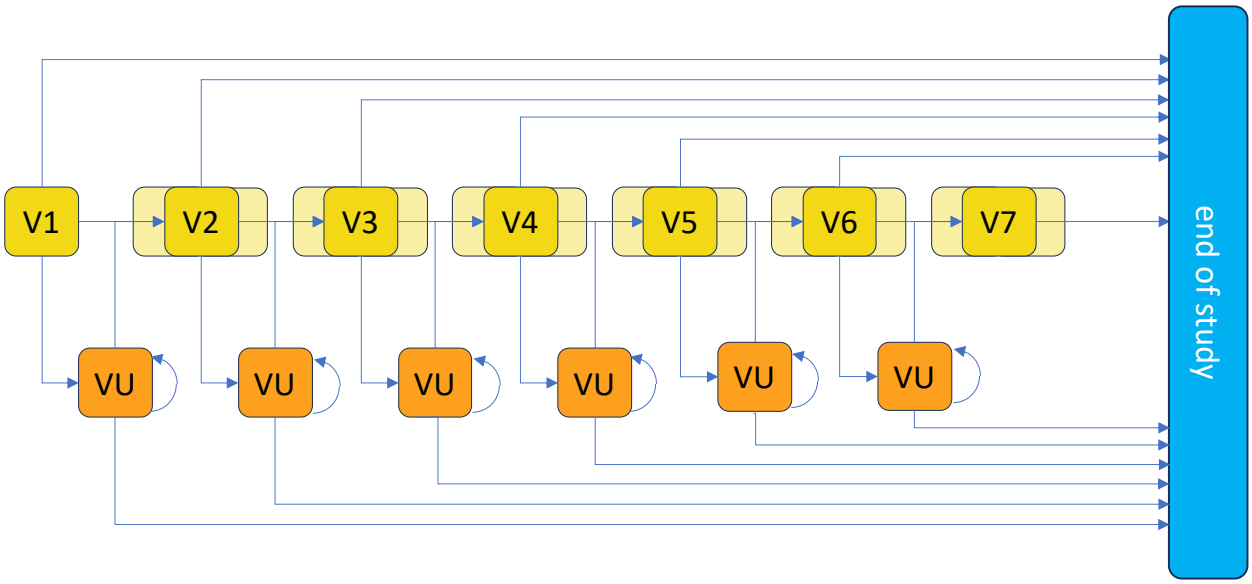

Figure 6. Windowing principle

The Encounter, Observation, MedicationAdministration, … FHIR resources with test data should contain dates that represent the windowing flexibility. During the integration and mapping step, the encounter should be mapped to the correct visit, based on the windowing algorithm. When the encounter falls outside the visit windows the encounter is mapping to an unscheduled visit.

TIMING

Clinical trials run over multiple weeks, months, or years. Hence, there's a need to reduce the real timeline of a trial into a manageable period for testing all predicted subject research paths and patient status. The figure below depicts the principle behind the time reduction for testing. Using the same 7 visit trial where visits take place every 7 days and a windowing of 1 day, the time reduction results in a visit every 7 minutes with a window of 1 minute.

Testing the direct data capture requires multiple research subject schedules for 1. candidate research subjects, 2. screening failures, 3. research subjects enrolled and active, 4. research subjects who completed the trial per protocol and 5. research subjects who completed the trial early. The related encounter and related FHIR resources implement the time reduction in the dates of the resource.

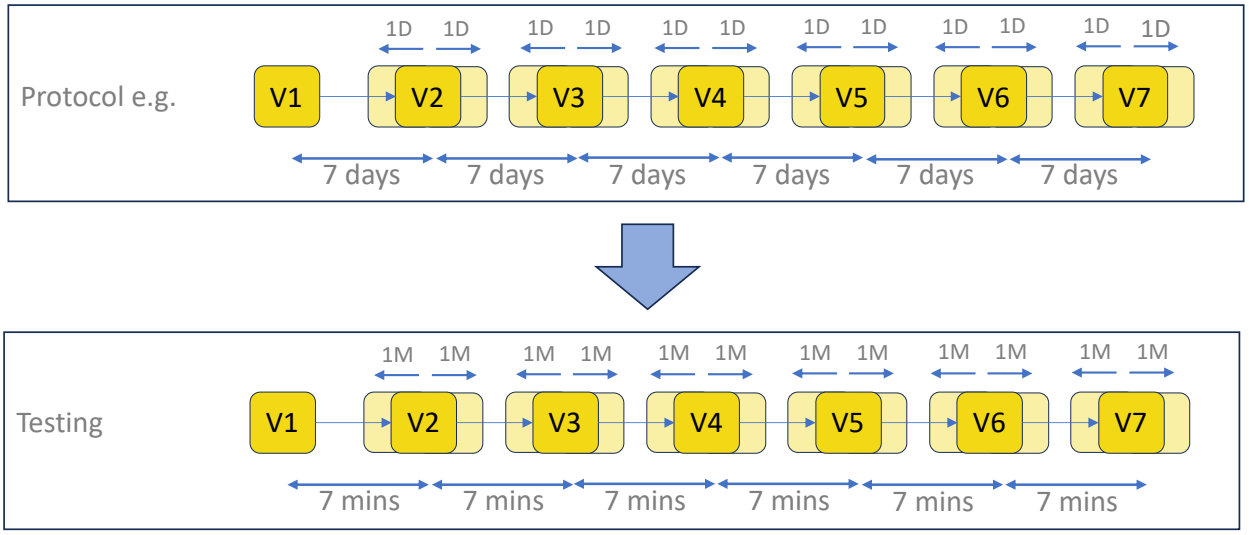

Figure 7. Time reduction

LIBRARIES AND AUTOMATION

Testing and validating a Direct Data Capture Pipeline to demonstrate and guarantee that the right data (semantically correct) at the right time and the right format (syntactically correct) is a complex undertaking. However, scalability can be achieved through libraires and automation. The figure below shows the implementation using and observation library containing Observation and ObservationDefinition FHIR resources.

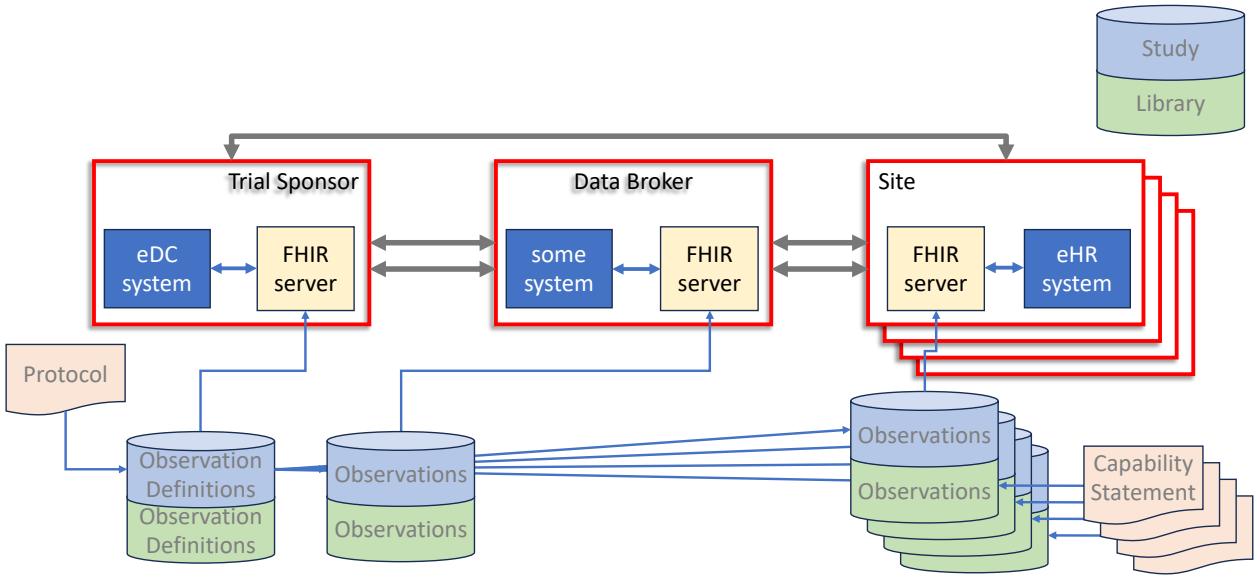

Figure 8. Libraries

A library at the trial sponsor level supports the study specification process (outgoing FHIR). When a protocol is implemented, the related ObservationDefinitions and observations are pulled from the library. However, when an ObservationDefinition and related Observation needs to be created or updated based on specific protocol requirements, it's added to the library for reuse in later trials.

A library at the research site level supports the testing and validation of the site specific configuration process (incoming FHIR). The library at site level contains the observations similar to the observations exposed through the site's FHIR server. This library takes into account the specific FHIR version and the standards terminologies and related versions used. When a protocol is implemented, the related observations are pulled from the library for testing based on the observation definitions in the study specification. These observations can be updated with generated test data, including

dates to match the study schedule. Equally, when a new observation needs to be created, it's added the site library for later reuse.
Next to the libraries, automation is the other critical component to achieve thorough testing and validation and to achieve scalability. Below are examples of tools that implement the automation:

- Creating and maintaining sponsor- and research site specific FHIR resource libraries
- Creating and maintaining schedule of activities as graphs including visualization using yEd or Mermaid
- Prediction of most probable paths through a trial
- Generating subject research schedules based on most probable paths with windowing.
- Translating protocol schedules in FHIR resources such as patients, appointments, and encounters.
- Generating descriptive statistics and best fit distributions of quantitative observations for creating test data
- Generating FHIR resources such as Observations, Medications, … with test data meeting the protocol.
- Reducing the trial timeline for testing purposes of the entire trial.

CONCLUSION

Direct data capture can be implemented today particularly for domains that are well structured in EHRs such as vital signs, laboratory results, medications etc …. This is resulting in efficiencies by eliminating labor intensive processes such as interpretation and transcription from source to EDC but also allowing for near real time data capture. Thus, as soon as a visit (an encounter) has taken place at the research site and the EHR updated, the data can be pulled from the EHR allowing, data managers, clinical research associates and clinical scientists the opportunity to immediately query the research site's FHIR server in support of their data and medical review processes. To be successful this requires both research sites to be DDC enabled *and* each trial's requirements to be implemented and confirmed.

The work here is designed to support the timely and accurate establishment and validation of any trial's specific DCC data pipeline that uses the FHIR interoperability standard. By using efficient SoA specification methods generating FHIR resources compliant with published FHIR methodologies (e.g. the HL7 FHIR Vulcan Accelerator IGs (ref 4)) enables the benefit of DDC to be realized without any additional burden on trial setup timelines and with confidence that the data collected thereafter meets protocol requirements accurately.

REFERENCES

1. Fast Healthcare Interoperability Resources (FHIR): https://hl7.org/fhir/index.html
2. The role of FHIR Resources in ensuring Semantic Equivalence in EHR2EDC direct data capture, Andy Richardson and Patrick Genyn, fhir4pharma: Connect/EU/Birmingham/PRE_RE08.pdf, Connect/EU/Birmingham/PAP_RE08.pdf
3. Vulcan Schedule of Activities (SoA) Project: https://build.fhir.org/ig/HL7/Vulcan-schedule-ig/index.html
4. Vulcan Accelerator Projects: https://confluence.hl7.org/display/VA/Vulcan+Projects.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the authors at:
Patrick Genyn – patrick.genyn@fhir4pharma.com
Andy Richardson – andy.richardson@fhir4pharma.com