# Ballot comments on FHIR Genomics STU 3

Written by Amnon Shabo (Shvo)[1], consolidating comments made along the past years in previous ballots, conf. calls, WGMs, and the HL7 Clinical Genomics mailing list.

September 2016

## Contents

Summary ..... 2
Should Sequence be a resource? ..... 3
Observation and DiagnosticReport Profiles for genetics ..... 4
Proposed changes ..... 5
Genetic family history profile ..... 6
Future work ..... 7

[1] Co-chair, HL7 Clinical Genomics and co-editor, HL7 CDA R2, CCD, GTR and Pedigree specs.

# Summary

The following comments refer to the Sequence resource and a number of genomics-related profiles mainly over Observation and DiagnosticReport, as presented in the FHIR STU3 Implementation Guidance.

Overall, this ballot proposal is an attempt to introduce a FHIR base/core resource that represents a sequence, and then profile base resources like Observation and DiagnosticReport in a way that also reference the Sequence resource, e.g., an observation of a variant points to Sequence that represents observed and reference sequences, along with pertinent data. Much effort has been put into this proposal and thus it is highly appreciated!

These ballot comments question whether Sequence is qualified to be a base resource. This issue should be examined by two main criterion sets:

1. Domain requirements and respective information modeling
2. FHIR requirements for creating a base resource

The proposed Sequence resource doesn't meet the above criteria because domain-wise Sequence is not necessarily a natural base to profile other types of omics data that are not sequence-oriented (e.g., cytogenetics, mass spec data for proteomics, etc.), assuming Sequence is the only base resource allowed in FHIR for omics. If however future base resources will be introduced – it is likely to create divergence and inconsistencies across those base resources. Also, FHIR criteria for new resource requires the resource to be naturally identifiable, while the proposed Sequence resource is a mixed-bag of sequences, variants, structure variants and more. If the variants in the Sequence resource are meant to represent the sequence itself, this is an attempt to create yet another bioinformatics format for sequences and that is, in my view, out of scope for the Clinical Genomics specifications. These reasons are further detailed in the following sections of this document.

As for the profiles, the Observation profile for genetics focuses on representing variants. This overlaps with the variants in the proposed Sequence resource, and this overlap can create confusion among implementers. In addition, this profile consists of the interpretation of the variant, which should be separated from the variant observation, to make the 'genotype-phenotype' association more explicit. While the interpretation attribute is a reference to another observation, which technically creates the aforementioned separation, it could have been better to move the interpretation to the DiagnosticReport profile for genetics, as this structure is meant to report on results and their interpretations.

Finally, the FHIR genomics specification doesn't refer to previous specifications developed and published by the HL7 Clinical Genomics Work Group in the past 13 years, which span the whole spectrum of HL7 standard families, from v2 to v3 and CDA. Lack of 'continuity of standards', at least at the underlying design principles, causes each project team to start over from scratch. It does not build on previous results; even if such results are considered unsuccessful, still, as in any scientific publication, it is crucial to cite previous work, criticize it as necessary, and point to what is done better in the current work. All of this has not been done and it is unfortunate, not only for Clinical Genomics, but for the entire HL7 community.

# Should Sequence be a resource?

The following are the FHIR guidelines for proposing a new base resource (http://wiki.hl7.org/index.php?title=Template:FHIR_Resource_Proposal ):

> ***Resource appropriateness***
>
> *Does the resource meet the following characteristics?*
>
> *Must*
>
> - *Represents a well understood, "important" concept in the business of healthcare*
> - *Represents a concept expected to be tracked with distinct, reliable, unique ids*
> - *Reasonable for the resource to be independently created, queried and maintained*
>
> *Should*
>
> - *Declared interest in need for standardization of data exchange*
> - *Resource is expected to contain an appropriate number of "core" (non-extension) data elements (in most cases, somewhere in the range of 20-50)*
> - *Have the characteristics of high cohesion & low coupling – need to explore whether coupling is good some places, not elsewhere – layers from Bo's document*

Based on the above criteria and on correspondence with Lloyd McKenzie who emphasized that a resource should be "naturally identifiable", I believe that the proposed Sequence resource does not meet these criteria.

The FHIR genomics Implementation Guidance states in section 10.9 (Overview) that "*This resource will be used to hold clinically relevant sequence data in a manner that is both efficient and versatile integrating new and as yet undefined types of genomic and other omics data that will soon be commonly entered into health records for clinical use.*" … *"This is consistent with how all FHIR resources are designed and used."*

What other FHIR resources target "undefined types of" data? The only one I can think of is the Observation resource that is a simple code-value pair (with metadata), carrying any value defined by the code and/or by external terminologies. If the Observation resource is the 'model' for creating a flexible structure, then in the case of omics we should rethink the Sequence proposal and possibly focus on something more basic (e.g., a genetic locus) that can be a better basis for existing & new types of omics data. Then, data sets such as this Sequence proposal or other types of omics data could profile the base resource, if it is truly generic as the Observation resource is.

In addition, the Sequence resource proposal doesn't meet the FHIR new resource criterion for "high cohesion". It's a mixed-bag of raw sequence data (ACGT...) along with reference sequences, variants, structure variants, quality data and more. Also, being "naturally identifiable" is certainly not a characteristic that can be attributed to variants, since their identification/detection is rather complex and uncertain at times.

Sequence should hold merely sequence data. Similarly to medical images where regions of interests are not bundled with the pixels of the image, a sequence should also not contain any information that is the result of downstream analysis. A sequence resource could include metadata about the sequence, e.g., quality, provenance, pointer to repository holding the full sequence and / or - if the sequence is not larger than the limits posed by FHIR) – also an encapsulated (inline) sequence represented in its native format (i.e., any bioinformatics format commonly used by the industry to represent sequences).

I believe we shouldn't attempt to remodel bioinformatics in HL7, but we could harmonize various metadata describing a sequence in HL7, and such harmonization could be already a significant contribution that is needed in the clinical genomics field, and is within the scope and mission of the HL7 Clinical Genomics Work Group.

It is evident that current sequence formats have their own limitations, which probably give rise to the development of new and specialized formats, e.g., HML (HLA Markup Language) that NMDP developed for HLA. However, I still think we shouldn't add yet another format. If we want to converge around a single format, we could identify one of the existing bioinformatics formats as most common and expressive, then interact with the developers of that format to get it improved based on use cases we describe in the Clinical Genomics DAM. If that's not possible, we could extend/refine this most promising format and point to it as the preferred format to use, when key chunks of raw sequence data are encapsulated. Refining/extending existing formats is better than creating a new bioinformatics format, especially when healthcare structures like FHIR are used to create a new formats, because, after all, these specifications are not meant for representing raw data, rather they target mainly clinical, administrative and financial data that are not voluminous in nature.

## Observation and DiagnosticReport Profiles for genetics

**Observation-genetics (Profile)**

The Observation-genetics profile currently represents genetic findings such as a variant, holding the variant id, along with other variant properties (e.g., location, length, etc.) represented in a number of Observation components or extensions. The latter helps disambiguating the various representations of a variant, e.g., dbSNP, HGVS, etc. by post-coordinating the primitive components.

It is proposed that this profile is where all information about a variant should be consolidated, pointing to the sequences used to detect this variant (e.g., one sequence resource could represent the observed sequence and another one could be the reference sequence used to compare the observed one in order to identify this variation.

As for the clinical interpretation of this variant, a change made to this profile in STU3 now enables representating the interpretation in a referenced Observation, where time, method, performers, etc. of the interpretation could be specified separately from those of the variation observation. This new structure is now shaping up towards a 'clinical genomics statement' that we've introduced in the CDA GTR (and with a broader scope – earlier in the v3 models). Nevertheless, best is if the interpretation will not reside at all in the observation because the reference to the 'interpretation observation' cannot hold

any semantics. That reference is the 'genotype-phenotype' association, and is the gist of any clinical genomics endeavor in my mind. Following these considerations, it is proposed to move the interpretation attribute extension from this profile to the DiagnosticReport-genetics profile (see more details below). This destination is not ideal to hold phenotypic data either, however, that is what seems to be available in FHIR.

**DiagnosticReport-genetics (Profile)**

As aforementioned, this profile over the FHIR DiagnosticReport resource already describes the interpretation of one or more genetic observations (referenced by the result attribute that points to Observation-genetics profile). For example, if the test is GJB2 Full Gene Test, there could be a number of variants found (e.g., V37I & V27I) with the interpretation set to pathogenic, using the "conclusion" attribute as well as the "codedDiagnosis" attribute, to represent the interpretation in both narrative and structured formats respectively.

Thus, it is proposed that this profile will be the only placeholder for interpretations of the referenced genetic observations. These references could then be further extended to hold a more fine-grain semantics of the genotype-phenotype association. A better modeling could be to have the association itself represented in a separate resource.

Note that in case of a study that consists of a number of different genetic tests (e.g., in hearing loss genetic testing, the GJB2 is one of a number of tests), then the 'overall interpretation' of all those tests could be best held by a document format such as the CDA GTR, using it summary section (see more details on this proposal in 'future work' below).

Minor question - What are the right profile names:

- ObservationForGenetics or Observation-genetics?
- ReportForGenetics or DiagnosticReport-genetics?

## Proposed changes

Following the comments described in previous sections, best is if the proposed Sequence Resource becomes a profile over a more common omics resource, as mentioned above. However, if this change is not accepted, then it is proposed as an intermediate approach, to at least make the following changes, in order to make a cleaner separation between the various artifacts proposed in this ballot, and make each of them better serving its scope.

This change proposal below does not describe all the details needed to implement it in FHIR, as it is focused on the 'backbone' of the FHIR genomics proposal.

**Sequence resource:**

Remove:

Move the following elements (including their nesting elements/attributes) from the Sequence resource to the Observation for genetics profile:

- referenceSeq
- variant
- allelicFrequency
- structureVariant

Change:

- Change the attribute name observedSeq to sequence

Constrain:

- Sequence (name changed from observedSeq)
    - This attribute is currently of type string, but it should be constrained to a common bioinformatics format for sequences as described above
    - A number of common formats could be allowed
    - A bioinformatics format could be constrained in its usage within this attribute

Add:

- Add a category attribute to define if a Sequence instance is an observed sequence or a reference sequence
- Alternatively, this addition can be avoided, by looking at the attributes ‘patient’ or ‘specimen’ – if they are populated then this is an observed sequence, otherwise it's a reference sequence of some kind (determined by other attributes)

## Genetic family history profile

The FHIR base resource FamilyMemberHistory has been profiled to create FamilyMemberHistory-Genetic (that roughly follows the v3 Pedigree specification) and is part of this ballot. However, I find an anomaly within this ballot: a whole exome sequencing (WES) for example is used in rare and complex disorders that are hard diagnose. By the FHIR principle of 80:20 (which, by the way, I believe is impeding HIT from having an impact on healthcare since it perpetuates non-meaningful use of HIT), WES sequences are in the 20% of cases using genetic data. Certainly if you compare it to genetic family history, see for example the paper "Someday it will be the norm: physician perspectives on the utility of genome sequencing for patient care" (Personalized Medicine, 2015, 12(1)), where physicians rated family history information to have higher utility than WES/WGS; however, this ballot introduces genetic family history to be merely a profile - aren't we creating an anomaly where sequence is a base resource and genetic family history is a profile?

# Future work

**Document structure**

As aforementioned in the comments on the DiagnosticReport-genetics profile, the CDA Implementation Guide for Genetic Testing Report (GTR) has the placeholder for the 'overall interpretation' (similarly to the v2 guide as well). This CDA specification could be ported to FHIR using the 'CDA on FHIR' resources and profiles (e.g., Composition and ClinicalDocument). If that turns out to be feasible, then each genetic testing section in the GTR will be represented in a FHIR DiagnosticReport-genetics profile, and the summary section could possibly hold an overall interpretation for the whole report (aligned with the indication/reason/assessed issue being the trigger for ordering this report). Documents also provide a more robust contextual structure with proper attestation. Documents should also the preferred way of exchanging information in interoperability scenarios.

**Phenotypes**

I believe that the Clinical Genomics Work Group should be focusing on building bridges between clinical and genomics as our group name implies. This FHIR effort has been focused on the genetic side of the main bridge we have to build, and to a lesser extent on the semantics of the association as well as the representation of phenotypic data.

The challenges around phenotypic data are pretty much a green field, while sequence formats have been developed by the bioinformatics communities and available for us to use. While phenotypic data could be seen as the entire health record of the patient, this is not that useful due to limitations of the latter, but on the other hand phenotype is not that simple as a single disease or medication. If you think of hearing loss possibly associated with the combination of presence of genetic mutations, taking certain antibiotics at certain age range, this becomes challenging to model, but without it, the whole effort of genetic representation is not that meaningful.

**Conceptual modeling**

The HL7 Clinical Genomics subgroup for Information Modeling started working on conceptual (standards-independent) models for the clinical genomics domain. Various specifications developed by the CG group should be aligned through that conceptual modeling, in a way that portions of the conceptual models are translated to logical models, corresponding with actual HL7 standard families, mainly v2, CDA and FHIR.