A Recommendation for a Minimum Data Set for Family Health History

GA4GH Clin/Pheno Pedigree subgroup - DRAFT

The collection and use of family health histories span medical activities from genetic research to heritable risk assessment in patient care. For all the stakeholders in this process, the goal must be data that is accurate and coded for effective analysis, and transferable between systems. To achieve this, a globally accepted and universally implemented family health history (FHH) data set should be established as a benchmark.

The recommendation in this document does not suggest specifications for a data model or technical standard. The companion work on a data model has been done separately (see the Pedigree Conceptual Model paragraph after the data element tables). Such a model may first exist outside of HL7 FHIR, then will be implemented within FHIR with input from the HL7 Clinical Genomics workgroup.

Background

This work was inspired by the efforts of the Personalized Health Care Workgroup of the American Health Information Community, which first released its recommendation on a core family health history (FHH) minimum data set on October 25, 2007. A peer-reviewed paper was published in December 2008.

The purpose of this document is to create an updated minimum data set that can be used not only in both research and clinical settings, but to eliminate the gap between the two disciplines. This recommendation should also guide the development of research, clinical, and patient-facing FHH data and information collection tools, applications, and data repositories. Our initial design was that the minimum data set should support the GA4GH use cases submitted here - with future support as new use cases emerge. This document should only be used as informative, with data representations (the data model) and alignment with existing standards (like PED and FHIR) to follow.

Many GA4GH driver projects and other interested parties in GA4GH have expressed interest in a pedigree standard that supports their wide-ranging use cases. These stakeholders, along with many others in the healthcare industry, want to see streamlined approaches to easier and more robust data collection, analysis, and data sharing. They want to build better applications, like the ability to change the pedigree focus and proband within the same data set. Currently, data collection processes produce limited and insufficient data that is locked in data silos. By building more robust FHH data sets, stakeholders will advance their capabilities in new knowledge generation.

Goals for the minimum data set

The operative purpose of a minimum data set is to standardize methods for FHH data collection, but then allow for other data elements to be collected based on the use case. We recommend that pedigrees support 1st-degree relatives (i.e., parents, siblings, children) and 2nd-degree (i.e., grandparents, aunts/uncles, nieces/nephews, half-siblings), with 3rd-degree (i.e., cousins) a stretch goal. It should be noted that the minimum data set has both required and optional data elements. Where appropriate and available, we list important terminologies and answer lists for data elements that require them.

Besides the larger goal of this recommendation in supporting a functional pedigree, we hope to promote the collection of both biological and non-biological family members, the ability for proband change and re-analysis, and adding genetic variant data, all for the advanced use of risk algorithms and clinical decision support - leading to when machine learning and AI tools are applied. Having data interoperability, easy data sharing, with privacy maintained, will provide benefits to all.

Long-term goal

As we build a common framework of standards and harmonized approaches for FHH data sharing, we hope to see the community of GA4GH - including healthcare, research, patient advocacy, life science, and information technology - expand their collection, study, use, and especially sharing, of FHH information.

Further opportunities of this effort should include the addition of genetic variant data and risk analysis information as part of the data set. This will lead to the development of more advanced FHH risk algorithms. A shareable repository of these algorithms should be established.

And finally, we see this document as 'living', and plan to maintain and update it as we receive use cases, implementation experience, and other evidence to do so.

Data element tables

The data element tables on the following pages are divided into 1) the data elements that describe the pedigree record as a whole, and 2) the data elements that add detail to each individual person in the pedigree. As a reference to existing formats, the data elements are mapped to both the PED file format and the HL7 FHIR FamilyMemberHistory resource. The *Notes* column includes more information about the use of the data element. The text in red is to highlight where gaps or remaining questions still exist. The required (R) elements are sorted at the top of each table. These data elements are necessary to build the structure of the pedigree or family tree. The optional (O) elements then continue the recommended "minimum data set" for those applications, databases, and implementations which seek to be a more robust risk assessment tool. The 'Optional' elements are intended to be a list from which to choose those data elements that fit your specific use case, yet maintain data interoperability.

Appendix sections

Additional information supporting the minimum data set appears in the Appendix sections. In these sections we cover the methods used in developing this document, the differing systems used to collect FHH data, and a longer discussion on Race, Ethnicity, and Ancestry (REA) issues.

We also list other data elements to consider that are outside of a minimum data set, such as data sharing status and consent status. Larger sections on capturing genetic observations and pedigree analysis results are included. And finally, we highlight the fact that the GA4GH Phenopacket Schema also includes family health history data.

Data Elements: Pedigree

Element Name	Required Optional	PED column	FHIR name	Notes
Family ID (Pedigree ID)	R	Col. 1 Family ID	FamilyMemberHistory. identifier	This is used when distinct family records are defined then shared. Also can be thought of as the Pedigree ID.
Proband ID and type	O / R		Patient <id value = "proband" Or New extension to include a Proband type	This is a repeating field. Only required when the pedigree is used to focus on heritable risk for a specific person in the pedigree. For other use cases such as research, a Proband type may be needed. Types may include *Consultant* and *First Tested Positive*. The FHIR resource ResearchSubject.identifier includes the following status choices: follow-up, ineligible, not-registered, off-study, on-study, on-study-intervention, on-study-observation, pending-on-study, potential-candidate, screening, and withdrawn.
Pedigree Source metadata	O		FamilyMemberHistory. Meta Or FamilyMemberHistory. reasonCode	Possible uses are, but not limited to, where did the pedigree come from, tool or method used, clinical or patient entered, research identifier.
Status	O		FamilyMemberHistory. status	From FamilyHistoryStatus choices are: partial, completed, entered-in-error, and health-unknown.
Language	O		FamilyMemberHistory. language	The codes SHOULD be taken from Common Languages.
Narrative	O		FamilyMemberHistory. text	A human-readable narrative that contains a summary of the resource and can be used to represent the content of the resource to a human. Resource definitions may define what content should be represented in the narrative to ensure clinical safety.
Date of FHH data collection	O		FamilyMemberHistory. date	Date should be the latest date of when any data is updated.

Data Elements: Individual

Element Name	Required Optional	PED column	FHIR name	Notes
Individual ID	R	Col. 2 Individual ID	Need to add to FamilyMemberHistory	An 'Individual type' (not type of ID) may be needed. This could include the individual's electronic medical record identifier, genealogical record identifier, could be the person is only represented via their lab sample identifier, or person information is missing or unknown (just have a relationship), or person information is private/hidden based on a consent choice, or the same individual and/or Individual ID is found in other pedigrees.
Father ID	R	Col. 3 Paternal ID	FamilyMemberHistory.extension:Parent	Must be an Individual ID
Mother ID	R	Col. 4 Maternal ID	FamilyMemberHistory.extension:Parent	Must be an Individual ID
Biological (Phenotypic) Sex	R	Col. 5 Sex	FamilyMemberHistory.sex	This element should ideally reflect whether the individual is genetically male or female (chromosomal sex). See AdministrativeGender and Patient Gender and Sex.
Person Gender	O		Person.gender	The gender might not match the biological sex as determined by genetics, but the individual's preferred identification.
Human Name Given	O		FamilyMemberHistory.name	Datatype is HumanName, includes Given name, and subsequent Given names for middle names.
Human Name Family	O		FamilyMemberHistory.name	Datatype is HumanName, includes Family name.
Relative Relationship	O		FamilyMemberHistory.Relationship	For the Pedigree conceptual model, GA4GH has created KIN, a new family relations ontology. Also see HL7 v3 Value Set FamilyMember and CodeSystem: RoleCode.
Relative first name	O		FamilyMemberHistory.name	This will either be a name or a description; e.g. "Aunt Susan", "my cousin with the red hair". This can also follow the datatype of HumanName.
Relative last name	O		FamilyMemberHistory.name	This will either be a name or a description; e.g. "Aunt Susan", "my cousin with the red hair". This can also follow the datatype of HumanName.
Relationship to Proband	O		FamilyMemberHistory.relationship	For the Pedigree conceptual model, GA4GH has created KIN, a new family relations ontology. Also see HL7 v3 Value Set FamilyMember and CodeSystem: RoleCode.

Consanguinity	O		Need new FHIR resource	In clinical genetics, consanguinity is defined as a union between two individuals who are related as second cousins or closer. These relationships may not be in the same pedigree. Need new extension consanguineous[x] having two sub-attributes, consanguineousBoolean, then consanguineousInteger, to curate the degree of consanguinity.
Date of Birth	O		FamilyMemberHistory.born[x]	In most cases the DOB can be just the birth year. If not known, age or age range is recommended.
Age	O		FamilyMemberHistory.age[x]	System should alert about age when the pedigree is over a year old.
Estimated age	O		FamilyMemberHistory.estimatedAge	If true, indicates that the age value specified above is an estimated value. Clinicians often prefer to specify an estimated age rather than an age range. boolean
Age range	O		2.24.0.8 Range is a FHIR data type that can be assigned to *ageRange*.	Represented as two elements of lower age range and upper age range.
Deceased	O		FamilyMemberHistory.deceased[x]	Deceased flag or the actual or approximate age of the relative at the time of death. boolean|Age|Range|date|string
Disease or Condition	O		FamilyMemberHistory.condition	This is a repeating field to allow a system to represent more than one disease or condition per person. Each disease or condition entered should be linked to its related data field - Affected, Code, Age of Onset, or Contributed to Death, if such data has been entered in those fields.
Disease or Condition Affected	O	Col. 6 Affected status		Links to Disease or Condition data element When using PED, a single disease is the focus for all in the pedigree. PED tags people as affected or unaffected for that single disease.
Disease or Condition Code	O		FamilyMemberHistory.condition.code	Links to Disease or Condition data element If a condition is included, a code should be used (SNOMED, ICD-10-CM, HPO) For recording disease outcomes and interventions, see *FamilyMemberHistory.condition.outcome* in B.2 below.
Disease or Condition Age of Onset	O		FamilyMemberHistory.condition.onset[x]	Links to Disease or Condition data element Either the age of onset, range of approximate age or descriptive string can be recorded. For conditions with multiple occurrences, this describes the first known occurrence. Age|Range|Period|string
Disease or Condition Contributed to Death	O		FamilyMemberHistory.condition.contributedToDeath	Links to Disease or Condition data element This condition contributed to the cause of death of the related person. If contributedToDeath is not populated, then it is unknown. boolean

Race, Ethnicity, or Ancestry (REA)	O		FamilyMemberHistory.extension:Observation Or Observation.extension (Ancestry) Codes for the various REA ontologies are needed.	Several REA ontologies exist, each designed with a specific need and philosophy. Data collection tools should allow for multiple values. Many sites will be bound by local requirements. (See more on this topic in A.3 below.) **The Human Ancestry Ontology (HANCESTRO)** provides a systematic description of the ancestry concepts used in the NHGRI-EBI Catalog of published genome-wide association studies. **The ClinGen Ancestry & Diversity Working Group** is developing standards and guidelines for clinical genetics about the interpretation, collection, and use of REA.
Multiple birth status	O		Need new FHIR resource Patient.multipleBirth[x] will not work	Need extension that includes a boolean datatype, then number of the multiple (to represent twin, triplet, etc.)
Adoptive status	O		Extension.value[x] for *adoptionInfo*	Value should be boolean. True means the adopted person is in the pedigree, but not the biological child of the mother/father IDs.
Observation including genetic markers	O		FamilyMemberHistory.extension:Observation Or Observation-genetics Profile	Allows capturing risk-relevant observations about the relative that aren't themselves a specific health condition; e.g. Certain ethnic ancestries that are disease-relevant, presence of particular genetic markers, etc See “Adding genetic observations” in B.4 below.

The Pedigree Conceptual Model

To support the interoperability of this family health history data set, we developed the GA4GH pedigree conceptual model. This model defines core classes for Pedigree, Individual, and Relationship data that are interoperable with other standards, including HL7 FHIR (implementation guide) and GA4GH Phenopackets (currently defined on a branch). Relationships between individuals are standardized using concepts from the newly developed Kinship Ontology. To allow existing workflows and tools to gracefully add interoperability with this standard, we developed an open-source pedigree data interoperability library, pedigree-tools.

APPENDIX A: Additional Information on Minimum Dataset Data Elements

A.1. Methods used creating this document

To get started, representatives from several of the GA4GH Driver Projects submitted FHH use cases that describe how FHH data is collected and used within their activity. For data interoperability, they shared what systems need to exchange pedigree data, and how a GA4GH standard would help. They described what computations they need to make from the pedigree model, or operations they need to perform on the pedigree representation. We also captured from them what data elements, relationship representations, or other requirements they need that are not supported by PED. And finally, we asked about any information which helps us understand their technical environment e.g. EMR connections, data collection tools, etc.

A table of data elements for a recommended FHH minimum core data set was created and based initially on the original data set published in 2007 (as mentioned previously). The GA4GH Pedigree subgroup then updated the data set to match what is currently used in PED, note what is supported by HL7 FHIR *FamilyMemberHistory* profiles, and highlight gaps and other issues. A data modeling process based on the minimum data set was completed to confirm how the data should be represented in an interoperable data message.

A.2. Capturing the data elements

Multiple formats and methods exist for the capturing of these data elements. Each has their strengths and specific design purposes. If a form, survey, or questionnaire is used, the HL7 FHIR Structured Data Capture (SDC) project offers an implementation guide around the use of questionnaires and data elements to standardize the capture of information via user-facing forms. Chatbot technology has been applied to the questionnaire format to guide people in gathering a complete history. Other data capture methods may include the graphical creation of a pedigree. These tools follow the genealogical method of building a family tree, but then add medical histories for each family member on that tree. The Pedigree Standardization Work Group (PSWG) of the National Society of Genetic Counselors (NSGC) developed a system for a clinical pedigree nomenclature. The pedigree drawing symbols they use are incorporated into their professional training.

The different formats and methods available are to support the creator of the family health history record. These creators include patients and family members with self-reported information, to genetic counselors, to clinical health providers putting data into an electronic medical record, to self-reported data that is validated by the EMR, and to the genomics researcher and clinical trial. All of these creators and their use cases must be supported, so that family health history data can be technically interoperable between all compliant systems.

A.3. Issues on collecting Race, Ethnicity, or Ancestry

We cannot make a recommendation on how to use Race, Ethnicity, or Ancestry at this time. After speaking with several organizations and experts on this question, we learned there exists differing needs and use cases that have resulted in multiple REA ontologies. At least one ontology uses an ethnicity-language combination in identifying ethnolinguistic tribal affiliation. Identification codes for the various ontologies should be created. Another helpful FHIR resource could be Extension.extension:Source, which could

capture in what way the ancestry was reported (e.g. Patient reported, Genetic test/markers, Other, Unknown).

A recent paper published in The American Journal of Human Genetics titled *Clinical genetics lacks standard definitions and protocols for the collection and use of diversity measures* states "... there was no consensus on the relevance of REA, including how each of these measures should be used in different scenarios and what information they can convey in the context of human genetics. A lack of common definitions and applications of REA across the precision medicine pipeline may contribute to inconsistencies in data collection, missing or inaccurate classifications, and misleading or inconclusive results."

Many in the research world use gnomAD and ancestry informative markers (AIM) to infer ancestry. The Genome Aggregation Database (gnomAD) Consortium sponsored by the publication Nature, and the Broad Institute gnomAD browser includes exomes and genomes from European, Latino African and African American, South Asian, East Asian, Ashkenazi Jewish and other populations. They assigned ancestry to all samples for which the probability of that ancestry is > 90% according to the random forest model. All other samples were assigned the other ancestry (oth).

Concerning the use of AIM, according to the National Human Genome Research Institute "Ancestry informative markers refers to locations in the genome that have varied sequences at that location and the relative abundance of those markers differs based on the continent from which individuals can trace their ancestry. So by using a series of these ancestry informative markers, sometimes 20 or 30 or more, and genotyping an individual you can determine from the frequency of those markers where their great, great, great, great ancestors may have come from. These are generally resolved to the three major continents: Africa, Asia, and Europe."

Also see -

- US Federal Drug Administration (FDA, contains nonbinding recommendations) Collection of Race and Ethnicity Data in Clinical Trials
- US Centers For Disease Control and Prevention (CDC) Race Category Value Set
- US Core Implementation Guide (v3.1.1: STU 3) based on FHIR Release 4 Detailed Race Value Set and Race & Ethnicity - CDC (mostly a list of countries and hundreds of US native indian tribes)

A.4. Biological (Phenotypic) Sex and Person Gender

For more information on this topic, see the Context Definitions in HL7's Gender Harmony Project. The site explains that "It is expected that these context of use will be an attempt to represent single or combinations of things such as chromosomal allosomes, genotype, hormone levels, secondary sex characteristics, desired perception, surgical status or planned outcome, organ function, etc." This is a work in progress.

APPENDIX B: Other Data Elements to Consider

These data elements are not part of the minimum data set recommendation, but could be very helpful in specific use cases.

B.1. Interoperability capability of the captured pedigree

- *Data sharing status* - interoperability characteristics with other systems.
- *Consent status* - may be required in some use cases, especially if names and DOB's are shared.
- *FamilyMemberHistory.dataAbsentReason (FHIR)* - describes why the family member's history is not available.
- Other available codes are as follows -

Code	Display	Definition
subject-unknown	Subject Unknown	Patient does not know the subject, e.g. the biological parent of an adopted patient.
withheld	Information Withheld	The patient withheld or refused to share the information.
unable-to-obtain	Unable To Obtain	Information cannot be obtained; e.g. unconscious patient.
deferred	Deferred	Patient does not have the information now, but can provide the information at a later date.

B.2. Post-coordination of data elements

- Representation of marital or partner status and other relationship qualities (e.g., estranged, close, household member)
- *Extension.extension:Source (FHIR)* - In what way the disease, condition, race, ethnicity, ancestry is reported (e.g. Patient reported, Genetic test, EMR record, Public records like death certificate or disease registries, Clinical trial or research, Other, Unknown)
- A '*Severity*' option for diseases and conditions
- *FamilyMemberHistory.condition.outcome (FHIR)* - Indicates what happened following the condition. If the condition resulted in death, the deceased date is captured on the family member
- An '*Unknown*' and level of '*Certainty*' option for data elements that are not certain
- A '*Sensitivity*' field and management options of the sensitive issue

B.3. Use case specific data elements

- Non-diagnosed health status
- Non-genetic laboratory results
- Fetus as an Individual type
- Age of last follow-up - important to track disease progression in a research use case

- Relevant environmental/social data
- Prophylactic surgery date/type - important for censoring age for pedigree analysis in the cancer space where prophylactic surgery is recommended management for people at risk e.g. mastectomy, oophorectomy, hysterectomy
- Presymptomatic screening initiation date/type, important also for establishing a censoring age for colon cancer in particular, and suspicious polyps are removed during colonoscopy

B.4. Adding genetic observations

One may choose from a long list of coded information, from variants to findings, in the HL7 Genomics Reporting Implementation Guide (STU1). The guide is officially labeled as “not fully complete” and only for trial use, but it is very robust and ready for piloting.

General Genomic Reporting	Guidance and examples for the general structure of genomic reports, how to report overall interpretations and how to report genotypes, haplotypes and different types of variants. This content will be leveraged by all genomic reporting implementations.
Variant Reporting	Guidance on expressing information about variants gleaned from various sequencing approaches including direct sequencing, shotgun sequencing, array-based testing, etc.
Pharmacogenomic Reporting	Additional guidance and examples related to genomic testing done for the purpose of assessing genomic variations' implication on the use of medications - both for oncology and for general patient treatment
Somatic Reporting	Additional guidance related to genomic testing done on somatic (non-germline) tissues, including assessments of tumors

B.5. Adding analysis results

PedigreeAnalysisResults is a class that represents the results of analysis done to the data captured in the family history pedigree. Originally from the HL7 V3 Family History/Pedigree Implementation Guide, these data elements could be retrofitted to be included in pedigree data sharing which contains advanced risk assessment. A FHIR profile for *PedigreeAnalysisResults* would need to be developed.

PedigreeAnalysisResults contains several options to represent the actual results:

- *AnalysisResult* **-** This class is a catcher for any analysis that cannot be represented through the other classes in this choice box.

- *Probability* **-** The value holds a probability of having what is represented in the *PedigreeAnalysisResults*.code attribute (e.g., disease, variation). The code attribute holds a value that indicates that this is a probability observation.

- *PercentageRisk* **-** The value holds a percentage risk of having what is represented in the *PedigreeAnalysisResults*.code attribute (e.g., disease, variation). The code attribute holds a value that indicates that this is a percentage risk observation.

- *Relative Risk* **-** The value holds a relative risk of having what is represented in the *PedigreeAnalysisResults*.code attribute (e.g., disease, variation). The code attribute holds a value that indicates that this is a relative risk observation.

- *Polygenic Risk Score* **-** The value is calculated from genetic observations in B.4 (Variant Reporting - single nucleotide polymorphisms) and can be integrated into disease risk as part of a cosegregation analysis, penetrance estimation, and risk assessment.

➔ *code:* identifies the disease or variation for which the probabilities/risks/etc. are calculated. Note that this class can be populated as many times as needed, for each clinical condition which is the is the focus of the risk calculations
➔ *negationInd:* can be used to represent the fact that there is no risk found for this family history and the clinical condition in the code attribute
➔ *methodCode:* holds the type of the algorithm used to analyze the pedigree (e.g., BRCAPRO)

Lifestyle and environmental risk scores could also be considered.

APPENDIX C: Supplemental Information

C.1. Family History in GA4GH Phenopackets

The Phenopacket Schema represents an open standard for sharing disease and phenotype information to improve our ability to understand, diagnose, and treat both rare and common diseases. The Phenopackets Pedigree element represents a PED file, which describes the family relationships of each specimen along with their gender and phenotype. PED files are typically used by software for genetic linkage analysis. A pedigree node extends the FamilyMemberHistory-Genetic profile.

https://phenopacket-schema.readthedocs.io/en/1.0.0/family.html#
https://phenopacket-schema.readthedocs.io/en/1.0.0/pedigree.html#rstpedigree

C.2. External Family Health History Efforts

Global Genomic Medicine Collaborative (G2MC) website on Family Health History - By better understanding the needs in each individual country, the FHH project aims to promote the benefits of family health history for patient self-knowledge, family sharing, public screening, pre-screening for genetic testing in a clinical setting, and research.

National Human Genome Research Institute (NHGRI) Family Health History

The Office of the National Coordinator for Health Information Technology (ONC) - Representing Family Health History for Clinical Genomics

GA4GH Family History Tool Inventory is a catalogue of family history tools currently available for documenting family health history information.