# Phenopcakets proposal - GCAT

This is a description of the exercise of converting the phenotypic data from a dataset from GCAT project to Phenopackets.

Find here the definition schema https://github.com/clauw87/phenopackets/blob/master/gcat.proto and one example of phenopackets structure for 1 PhenopacketSet corresponding to 1 subject: https://github.com/clauw87/phenopackets/blob/master/gcat_pheno.

## 1 Description of data

The dataset used for this exercise has the following data types:

1. Lifestyle data [BASELINE] Data comes from one one-time survey at the beginning of the project and includes diverse metrics of physical activity, diet, smoking and alcohol drinking habits with their time intervals as well as categories and scores based on those metrics such as WHO alcohol behavior category based on gr alcohol consumption, fagerstrom score for smoking and predimed score for mediterranean diet attachment.
2. Diagnoses [TIMECOURSE] Data comes from EHR and includes diagnostic codes (ICD9 diagnostics)(including diseases and other clinical findings, signs or symptoms), categories of the medical unit (hospital, etc).
3. Medical Interventions [TIMECOURSE] Data comes from EHR and includes procedure codes (ICD9 procedures) and codes and types, names for the medical unit where procedure as performed
4. Medications [TIMECOURSE] Data comes from pharmacy system records and includes codes of drugs (UMLS ATC codes) and diary defined doses.
5. Medical Measures [TIMECOURSE] Data comes from measures taken in medical setting and includes anthropometric metrics and medical or laboratory measures as well categorical variables and scores based on those such as BMI, or "disease risk score"

## 2 Reuse of base phenopacket shema and new messages

For this exercise, we used the current version of Phenopacket schema v1.0 https://github.com/phenopackets/phenopacket-schema, as well as new messages proposed for v1.1 https://github.com/phenopackets/phenopacket-schema/compare/v1.1. Messages from base.proto v1.0 and new (v.1.1) reused for this exercises as well as the modified and custom messages are here https://github.com/clauw87/phenopackets/blob/master/gcat_all.proto.

In summary, modifications of base schema for this definition schema included the following:

- All timecourse data from each individual was stored as a *EncounterSet* (based on new(now deprecated) message TimeCourse) associated to that individual, where each Encounter (Phenopacket-like) is defined by its date (i.e data is in same Encounter if they shared date value)
- One-time point data (survey data) was stored as one Encounter with date, *time_at_encounter*, and individual age, *age_at_encounter* at survey time, although the elements referenced there include their own *TimeElement*
- Lifestyle data on Diet, Smoking and Alcohol Drinking and Physical Activity was stored as message *ExposureExtended*, based on new base.proto message *Exposure* including a *modifier* (OntologyClass) to indicate current or past behavior as well as messages *TimeElement* (age_at_start and, intervals for past behaviors), a *Quantity* (OntologyClass) and *Frequency* (OntologyClass)

- Timecourse data (Measurements, Findings, Diagnoses, Intervention, Medications) was stored as messages *Clinical Findings* (based on base schema PhenotypicFeatures), *Diagnoses* (based on base schema Diseases), and *Procedures* (based on new message Procedures) and *Medications* (based on new message PharmaceuticalTreatments within new message *MedicalAction*), adding a new message *Origin* describing the medical units where each of them occurred
- A *frequency* (ontologyClass) block added to *Dose Interval* within *Medications* in addition to *Quantity* and *Interval* to express medication routines as 50 (value) mg (units) "daily"/"every 8 hours" (frequency) for 2 weeks (interval). This could be later converted/ interoperable through summary doses per week, year, etc.
- *Quantity* message and *frequency* (OntologyClass) were included in *ExposureExtended* message for Lifestyle/ behavior data
- New message *Measurements*, including *parameter* (OntologyClass) and *Quantity* message was used for quantitative measures (anthropometric measures and medical measures)
- New messages *Scores* and *Categories*, reusing OntologyClass, were used to describe Findings that are better described as numeric values or categorical values associated to an ontology term than as *PhenotypicFeature* or *Measurements*

# 3 Definition schema

Here is a pdf version of the gcat.proto messages. Highlighted are the new messages proposed for this exercise. Note, messages now in test are designated as "org.phenopackets.schema.v1.1.core".

https://github.com/clauw87/phenopackets/blob/master/gcat_highlighted.pdf

# 4 Wonderings/Concerns

- TimeCourse vs bulk of phenopackets with many TimeElements. Does TimeCourse make sense even if it is not longitudinal design (i.e, encounter times are not in design but different for subjects, medical visits). Alternative would be maybe plain old phenopackets for each encounter adding then the TimeElement of each. Is this structure more efficient/better for later use of data.

- History of-data. Problem with survey data (hstory of disease, events, lifestyle, exposures) is that TimeElement of the described feature (e.g Exposure) don't math time_at_encounter, which is the survey time instead (what is true though is age and length of time intervals and time till survey time if exposure is current). Is there a better way of representing this kind of data?

- Similarly, though less troublesome, Medical Actions include sometimes an TimeElement interval and not a timepoint. In intervals, only time_at_start will match time_at_encounter.

- Duplicated TimeElements? Should TimeElement be required in the case of Diagnoses, where age and time will match exactly with age and time_at_encounter? Would that be better for findability/interoperability across resources?
- Seemingly duplicated info. Should be removed? Data that are derived from some other variable should be kept or removed?, e.g, scores or categories or phenotypic features based on stored measures, modifiers like past behaviors. Duplicated info, in these cases can be good for findability/interoperability with other resources.
- Metadata. Are these standard enough? should we map to some other standards?
- Units: IS units, non-IS units if necessary can defined in Metadata message?, e.g smoking fagerstrom score, MET
- UMLS ATC codes for medications, ICD9 codes for diagnoses and procedures. Are these standard, interoperable enough? Should we map to more used ontologies? e.g ICD9 codes include diagnostic codes for both diseases and not not diseases (clinical findings, signs), and phenopacket's idea is that Disease message include only diseases while PhenotypicFeature fits the others. To do this, we would need to do some selection or mapping NCIT Diseases or NCIT Clinical Findings, or similar.