Building a clinically significant rare disease data master approach and workflows

- KNIME 2019 FALL SUMMIT -

Sebastien Lefebvre
Senior Director – Data Sciences, Genetics and Bioinformatics
Alexion pharmaceuticals Inc
Rare disease challenges

Current challenge:

- There is no single authority providing a comprehensive list of rare disease definitions. A rare disease definition is the backbone to which all clinically relevant information can be mapped.

- There is a poor understanding and alignment around what makes up a rare condition both in academia, clinics and healthcare.

- The belief is that there are between 7,000 to 9,000 rare conditions depending which source you consult. Our complete set contains over 12,000 conditions with 7,230 from Orphanet with many still being defined.

- There are clear gaps in the medical coding (SNOMED-CT and ICD-10) coverage of rare diseases. There are 128 ICD-10 codes mapped to a rare condition.

- There are 161 conditions with approved drugs.

Some of the key sources:
Analysis approach

Why, how, what, outcome, value

Why we use KNIME:
1. Agility → build recipes that are easy to understand/modify
2. Knowledge capture and sharing → KNIME analytics app and server
3. Bring the analysis to the data → easily share “recipe”/analysis
4. Collaboration internal/external → open source desktop app
5. Easily deployable → KNIME server
Rare disease definitions

START WITH MAPPING THE UNIVERSE OF RARE DISEASES

Playbook:

1- Focused on key rare disease questions

2- Driven by a comprehensive data strategy

3- Near real-time tailored evidence data graph

4- Near real-time evidence based analytics pipelines

Each link represents a set of evidences that are extracted using a complex data preparation pipeline.
Rare disease definition with **clinical utility**

Build from many trusted curated sources and validated by expert curation

**Mapping approach:**

1. define the gene-condition master
2. assess the causal relationships
3. link to expert review
4. link to approved drugs/treatments
5. link to trials
6. link to biomarkers and lab tests

Each **link** represents a set of evidences that are extracted using a complex data preparation pipeline.
**Insights workflow architecture – tier approach**

Trying to avoid workflow dependency chaos

**Level-1**
Source to Normalized

- Starts with:
  - Raw sources
  - Specialized data API calls
  - Curated sets

- Generates:
  - Tabular sets
  - Data master Index with lineage

**Level-2**
Normalized to Useful

- Starts with:
  - Data master Index files
  - Tabular sets
  - Specialized data API calls
  - Curated sets

- Generates:
  - Derived datasets

**Level-3**
Useful to Analysis

- Starts with:
  - Derived datasets
  - Level-1 generated datasets
  - Specialized data API calls
  - Curated sets

- Generates:
  - Analysis report and supporting datasets

**Level-4**
Analysis to Insight

- Start with:
  - Analysis report and supporting datasets
  - Level-1,2 generated datasets
  - Specialized data API calls
  - Curated sets

- Generates:
  - Insight through visualization

** We have level-0 pipelines for downloads
Rare disease master (level-0,1)

Normalize and Master

RDM core

data_masters
  disease_master
    disease_ages
      disease_alias_index
        disease_clingen_index
        disease_drugbank_index
        disease_genereviews_index
        disease.omim_index
        disease_orpha_descriptions
        disease_orpha_index
        disease.prevalence
      drug_master
        drug_drugbank.disease_index
        drug_drugbank_index
        drug_drugbank.prices_index
        drug_drugbank.products_index
        drug_drugbank_trial_index
    gene_master
      gene.clingen_index
      gene.genereviews_index
      gene.omim_index
      gene.orpha_index
    trial_master
Rare disease master (level-0,1)
Rare disease master journey

Connecting the right data navigating the life sciences data maze

What we have done:

1. Built the rare disease all inclusive set
   • Include all type of matching “exact”, “broader/nearer term”, “fuzzy”
2. Got clinician feedback on gene-to-condition mapping
3. Built the clinically relevant rare disease master set by identifying the right source of information
   • Only include “exact” curator matching
   • Limit the sources to only ones trusted by clinicians

What we plan to do next:

1. We believe our master set is 80% clinically actionable and we need to improve that score
   • Improve feedback loop from clinicians → KNIME web portal
2. Increase our confidence using the complete definition of rare diseases
   • Focus curators on gaps → Use NLP (e.g. clustering, Ontology signatures) against all available unstructured information from trusted sources to identify gaps and guide curators
Analysis collaboration

KNOWLEDGE SHARING AT SCALE

Alexion analyst desktop

Alexion viewer

Partner viewer

Partner analyst desktop

KNIME server (Knime.alexion.cloud with SSO)

KNIME-portal Knime.alexion.cloud

Alexion common

Partnership common

KNIME: /user_Sandbox

KNIME: /user_sandbox

KNIME: /common_QC

KNIME: /collaboration_QC

Alexion viewer

Partner viewer

Alexion analyst desktop
Example of clinical utility: gene to treatment

Removing the google search and addressing the clinical knowledge gap

Patient diagnostics:
- Variant/Gene
- Condition

Match against the RDM

Clinical utility
- Diagnostic information
- Available treatment options

Rare Disease Master (RDM):
- Clinical diagnostics & description [e.g. GeneReviews, ORPHANET]
- Gene & Variant [e.g. Clingen]

Treatment:
- Drug-Conditions [e.g. DB+, DailyMed]
- Drugs-Trials [e.g. DB+]
- Trials-Conditions [e.g. CT.gov]
- Trials-Interventions [e.g. CT.gov]
- Expert Reviews-Conditions [e.g. NCBI]
- Biomarkers & lab tests [e.g. Excelra]
DOCTORS AND PATIENTS GET QUICK, LIFE-SAVING INFORMATION

What we are enabling with the rare disease master

• Doctors and patients get actionable, coordinated information in days not weeks.
• Doctor and patient engagement creates a knowledge exchange – flywheel/network effect that makes solution better over time.
• Create patient-centric communities around genetics, symptoms and treatments that work – potential to scale through crowdsourcing
• Provide re-diagnosis and updated treatment guidelines to doctors and healthcare systems
Q&A