Challenges and opportunities of

Standardizing specialized data on hematologic malignancies

Ana Heredia[1], Michel van Speybroeck[2], Laura Jamilis[1], Rubén Villoria[1]
[1]GMV, Valencia, Spain; [2]Janssen, Beerse Belgium

Background

OMOP has been the common data model of choice to standardize data in the context of the IMI-funded **HARMONY Alliance**, aiming to set up a pan-European, anonymized, Big Data repository of longitudinal data on **hematologic malignancies** provided by partners from both the **public and private sector**. Four data sources have been mapped so far, including The Cancer Genome Atlas TARGET-AML public clinical data. All four sources correspond to completed prospective trials conducted between 1995 and 2017, period spanning the publication of three updates of the European LeukemiaNET recommendations for diagnosis and management of acute myeloid leukemia in adults[1] (ELN) and two versions of The World Health Organization Classification of myeloid neoplasms[2] (WHO) although some of the sources still make use of the older and less restrictive French American British (FAB) classification (Figure 1).

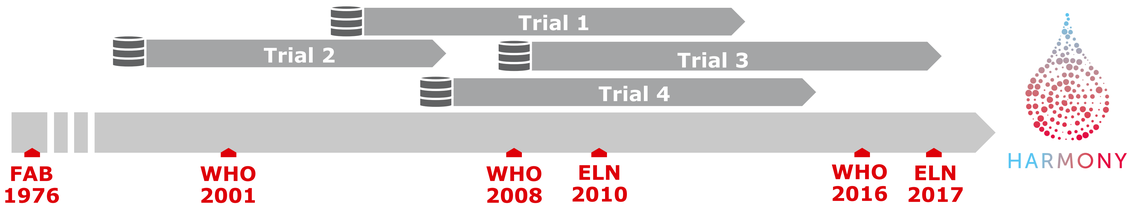

Figure 1. Global representation of the status of data provided to HARMONY Alliance based on the chronology of data collection and evolution of acute myeloid leukemia guidelines.

Major challenges

Most of the **disease-specific data** is present in registries or trial case report forms (CRFs) using singular coding requiring manual mapping. The information collected largely resembles **disease-focused guidelines** pre-coordinating concepts that are seldom combined in other situations, especially remarkable in the area of cytogenetics. These guidelines also define outcome measures considering the most relevant events during the course of the disease and establish very unique prognostic scoring systems which might not entirely overlap with those defined for other conditions and are therefore **not fully represented in the standardized vocabularies**.

The fact that absence or negative entities are explicitly captured in the respective datasets **contradicts model conventions** widely accepted in regular clinical practice, as does the recording of entities at levels of detail matching classification concepts e.g. exposure to broad classes of drugs or history of grouping disease and causes of death.

Apart from the coding and content of CRFs, the simultaneous analysis of multiple clinical trial data should consider the **population biases** introduced by inclusion and exclusion criteria, the representation of which in the mapping currently represents a challenge. Similarly, capturing adverse events at the moment would **not naturally fit the key-value model** in place for observations and measurements but require the use of fact relationships unifying the actual event with the grade, severity, whether it is expected or not and the drug that might have induced it (Figure 2).

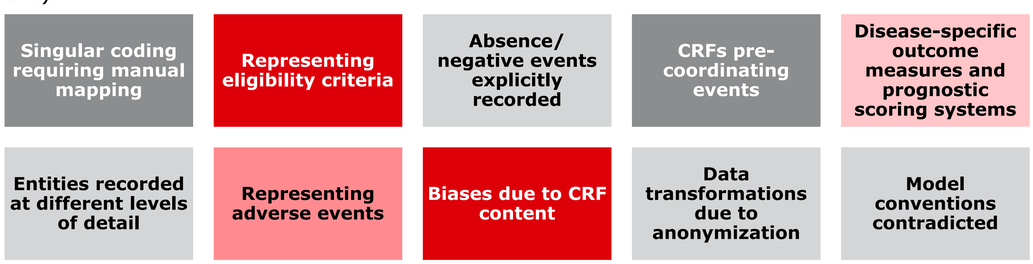

Figure 2. Major challenges faced in the process of standardizing specialized data on hematologic malignancies.

Methods and results

To understand to which extent custom concepts had to be created, a **comparison** was performed between the **standard and custom concepts** used in the mappings by clinical area.

Approximately half (50,8%) of the concepts present in the mapping logic were custom, mainly corresponding to information collected at the moment of diagnosis and being cytogenetics the major contributor to the need of custom concept creation, followed by targeted sequencing and survival representation (Figure 3).

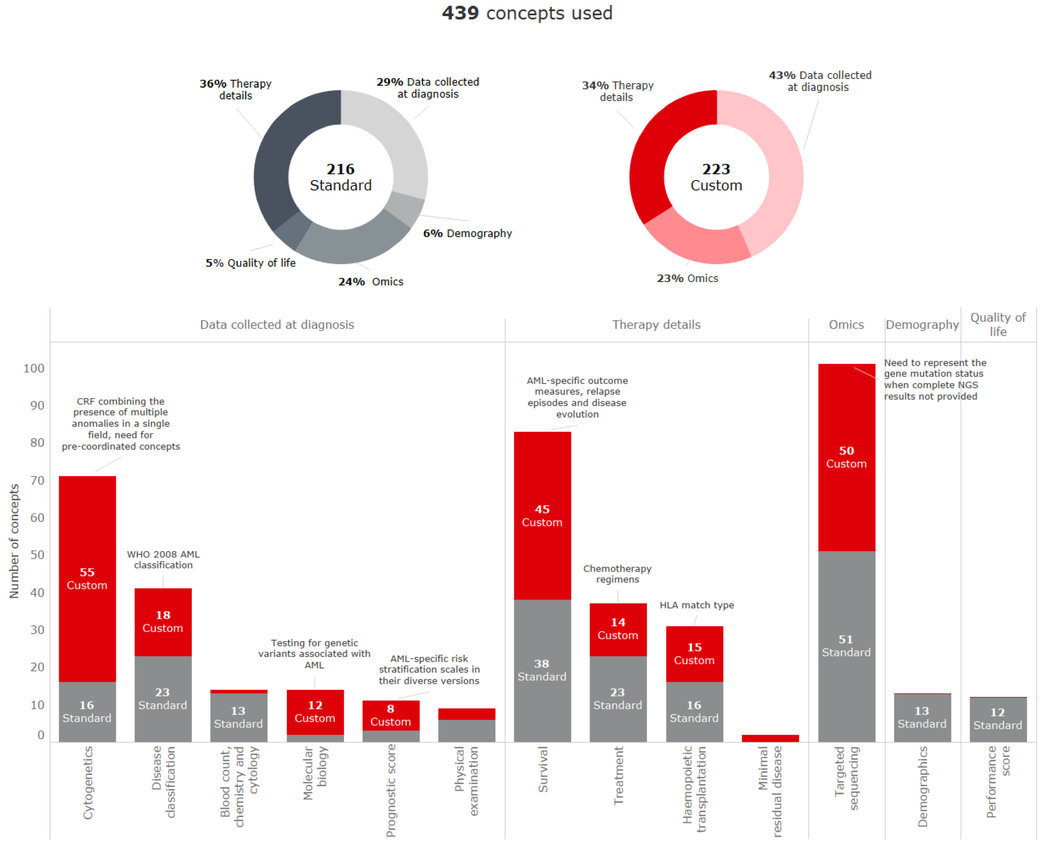

Figure 3. Concept type comparison considering the amount of existing and custom concepts as well as the contribution to the total of new concepts in percentage by clinical area.

Opportunities

Despite the hurdles, the analysis of data sources focused on hemato-oncology could result in **opportunities** for the community to

- continue growing and **widening the scope of the CDM** by serving as use cases for including new vocabularies as The Cancer Care: Treatment Outcomes Ontology (CCTOO)[3] covering outcome measures or
- **improving proposals** such as those for genomics and clinical trials, considering incomplete NGS data and inclusion/exclusion criteria.

Given that the creation of custom concepts is almost unavoidable, the establishment of a **unified procedure for taxonomy development and maintenance** could greatly facilitate the mapping tasks in multi-site collaborative projects.

References

1. Vardiman JW, Thiele J, Arber DA, Brunning RD, Borowitz MJ, Porwit A, Harris NL, Le Beau MM, Hellström-Lindberg E, Tefferi A, Bloomfield CD. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. Blood 2009;114(5):937-51.
2. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, Dombret H, Ebert BL, Fenaux P, Larson RA, Levine RL, Lo-Coco F, Naoe T, Niederwieser D, Ossenkoppele GJ, Sanz M, Sierra J, Tallman MS, Tien HF, Wei AH, Löwenberg B, Bloomfield CD. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood 2017;129(4):424-447.
3. Lin FP, Groza T, Kocbek S, Antezana S, Epstein RJ. Cancer care treatment outcome ontology: a novel computable ontology for profiling treatment outcomes in patients with solid tumors. JCO Clinical Cancer Informatics 2018;2: 1-14