the hyve

We empower scientists by building on open source software

Vocabulary mapping quality assessment in two European datasets - an Achilles extension

Maxim Moinat[1], Stefan Payralbe[1], Ilona Pinter[1], Marinel Cavelaars[1], Kees van Bochove[1], Rients van Wijngaarden[2]
*E-mail: maxim@thehyve.nl.
[1] The Hyve, Utrecht, The Netherlands; [2] The Pharmo Institute, Utrecht, The Netherlands

1. Background

At The Hyve, we have developed conversions of several **European datasets to the OMOP CDM**. This includes source data analysis, conversion workshops, development and execution of ETL scripts and most importantly data quality assessment. Each source dataset is unique, both in structure (syntactic) and in ontologies (semantic) used.

The **semantic mapping** is a challenging and imprecise task, which is especially pronounced for non-US datasets. The **OMOP standard vocabulary** was primarily built around US claims data and therefore in many cases does not support local European ontologies. The OHDSI community is working on support for European ontologies and already incorporated many in the OMOP vocabularies. A great effort is for instance the RxNorm Extension[1], which includes RxNorm-like concepts for international drug prescriptions.

Still, high quality vocabulary mapping poses a challenge. Here, we present an extension to the existing quality assessment tool, **Achilles**, to monitor **vocabulary mapping quality** with additional descriptive statistics. We applied this new extension for the quality analysis of the OMOP conversion of **two European datasets:**

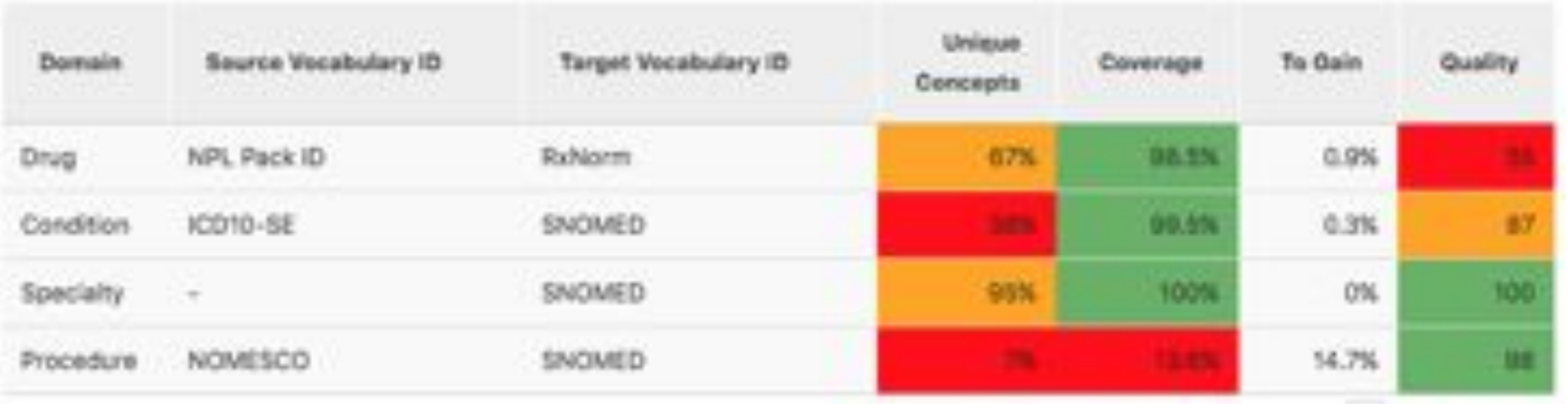
- **Swedish Electronic Health Records** (EHR), consisting of four national registries; patient care[2] (inpatient and outpatient), death registry, a subset of the drug registry (only cardiovascular drugs) and a subset of the socio-economic registry.
- **Dutch Primary Care** (PC), collected from a selection of primary care providers in the Netherlands. The primary care consists of episodes, medication, test results and referrals to secondary care.

2. Methods

To assess and track the quality of these vocabulary mappings, we built an extension of the Achilles R code. This extension, published on our Achilles Fork[3], scans every set of concept id, source concept and value in all CDM tables. For example, the *condition_concept_id*, *condition_source_concept_id* and *condition_source_value* in the CONDITION_OCCURRENCE table. This data is enriched with mapping information from the SOURCE_TO_CONCEPT_MAP.

The mapping coverage is calculated as number of records with a mapping (non-zero and non-null) divided by the total number of non-null records. Besides the coverage the following metrics are also outputted:

- Total number of distinct source codes
- Total number of distinct target concept_ids
- Total number of distinct persons
- Total number of rows
- Total number of rows with mapped concepts



Domain	Source Vocabulary ID	Target Vocabulary ID	Unique Concepts	Coverage	To Gain	Quality
Drug	NPL Pack ID	RxNorm	67%	98.5%	0.9%	55
Condition	ICD10-SE	SNOMED	38%	99.5%	0.3%	87
Specialty	-	SNOMED	99%	100%	0%	100
Procedure	NOMESCO	SNOMED	7%	13.6%	14.7%	98

▲ ***Figure 2:*** *Mockup of a new '**Mapping Quality Scorecard**' that assess the completeness and quality of vocabulary mappings.*

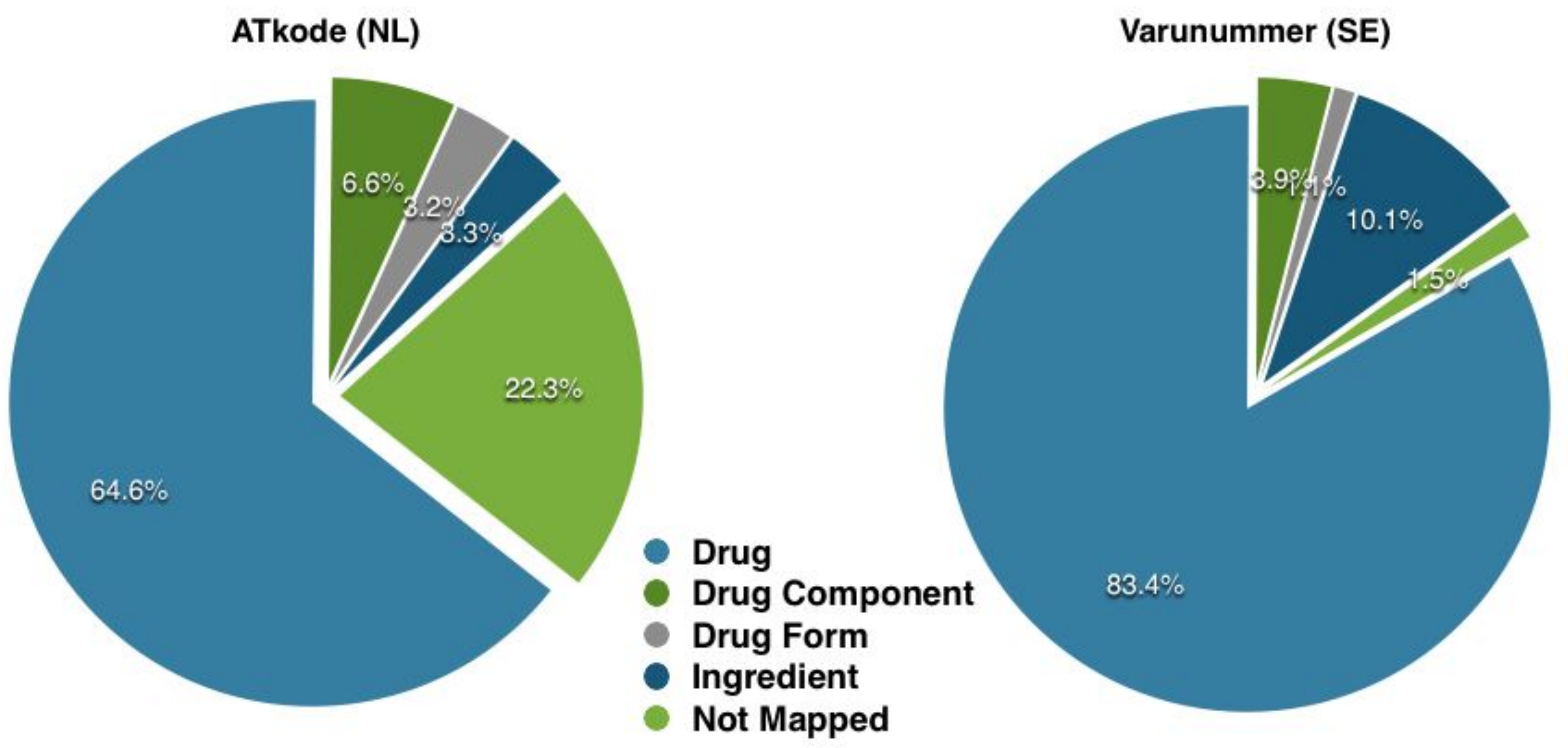
3. Results

The Achilles extension outputs a general **coverage** (Table 1) and a detailed metric per source vocabulary and **target concept class**. The detailed metric can reveal additional quality information, for example the RxNorm class to which drug concepts were mapped (Figure 1).

In addition, the user can choose to export a list of the most occurring mapped and unmapped source concepts. The first serves as a sanity check of the mappings and the second shows how much there is **to gain by mapping the top unmapped concepts**.

	Condition	Procedure	Drug	Measurement	Observation
Swedish EHR	99%	14%	83%	*N/A*	*N/A*
Dutch PC	96%	*N/A*	65%	100%	34%

▲***Table 1:* Coverage of the mapping to _concept_id** in five main OMOP CDM tables. Coverage is defined as the percentage of rows that have a mapping to a standard _concept_id (not to 0 or non-standard).



◂***Figure 1:*** **Mapping metrics for ATkode (Dutch drug items) and Varunummer (Swedish drug items)**, stratified per target drug concept class. Concept classes are merged for simplicity, 'Drug' encompasses all drug concepts with ingredient, component and form and the others only one of these three. It can be seen that more Swedish codes were mapped to an RxNorm concept than Dutch codes (1.5% and 22.3% not mapped respectively). However, of the mapped Swedish concepts, a relatively large proportion was mapped to an Ingredient and loses dosage form and strength information (10.1%). The Dutch codes only map for 3.3% to ingredients.

4. Discussion

In order to improve integration into the OHDSI quality analysis workflow, we have started to integrate the outputs into the **Data Sources component of Atlas**. This is work in progress and a first mockup is shown in Figure 2. Another useful feature would be to enrich the **Achilles Heel warnings** with information from these new vocabulary mapping metrics. For example by creating a drilldown with which concepts are not mapped and creating new warnings based on the vocabulary mapping metrics presented here. This quality analysis is complementary to the existing Achilles Heel quality analysis.

Lastly, the proposed mapping metrics are all quantitative; they report whether a mapping exists or not. The metrics cannot be used to determine whether an existing mapping is correct. For this purpose an **expert review** of the individual mappings should be provided. Usagi is a possible tool to extend with this capability, as it already allows an expert to search for and give a new suggestion for a mapping.

References: [1] C. Reich, http://www.ohdsi.org/web/wiki/doku.php?id=documentation:international_drugs, [2] Kodningskvalitet i patientregistret- slutenvård 2008, Socialstyrelsen, 2010 June [3] Github The Hyve, https://github.com/thehyve/Achilles/tree/vocab_mapping