**Combining the ATC Drug Classification System with the RxNorm Drug Nomenclature into a comprehensive Drug Ontology: Challenges and Achievements**

**Anna Ostropolets, MD1, Hamed Abedtash, PharmD, PhD2, Peter Rijnbeek, PhD3, Marcel de Wilde3, Alexander Davydov6, Olha Marushchak4, Christian Reich, MD5**

*1 Columbia University, New York City, NY, USA,2 Eli Lilly & Company, Indianapolis, IN, USA,3 Erasmus University Medical Center, Rotterdam, The Netherlands,4 Danylo Halytsky Lviv National Medical University, Lviv, Ukraine, 5IQVIA, Cambridge, MA, USA,6 Odysseus Data Services, Cambridge, MA, USA*

**Introduction**

The Anatomical Therapeutic Chemical (ATC) classification of drugs developed by the World Health Organization Collaborating Center for Drugs Statistic Methodology is employed routinely in pharmacoepidemiology, pharmacovigilance, pharmacoeconomy, and observational research as well as drug utilization monitoring programs. ATC classifies drug products by the target organ, pharmacological mechanism, therapeutic intent, chemical structure, or active ingredients into five levels1: The first level represents anatomical, the 2nd therapeutic (=indication) or pharmacologic, the 3rd and 4th chemical, or more detailed pharmacological or therapeutic attributes, and the 5th most granular level ­represents the chemical structure (active ingredients). Since the RxNorm terminology is the standard for marketed drug products published by the National Library of Medicine (NLM), a crosswalk between ATC and RxNorm concepts is essential to the observational studies on the drug exposure data in the US. Due to the complexity within the ATC hierarchical structure, merely matching the lowest ATC 5th level concepts to RxNorm ingredients results in misclassification of RxNorm drug concepts, since route, dose and indication information is lost in this transition leading to selection and measurement biases in data analysis. We aimed to develop and implement an automated RxNorm-to-ATC mapping process that would provide the best match between concepts from the two terminologies while concept attributes are preserved.

**Methods**

To create a set of comprehensive mappings, we extracted all the ATC codes along with their names, route of administration (ROA), daily defined dose (DDD) and additional notes from the official website (https://www.whocc.no/atc\_ddd\_index/) using a custom Python grabber based on the library “scrapy”. In our study, we used ATC version 2018-02-13 and RxNorm version 2018-08-12. From the original five-level ATC hierarchy, we sought to connect the lowest level codes in the ATC (ATC 5th) system to RxNorm Ingredients. Mapping these ATC codes requires to not only match the active ingredient, but also the other attributes such as indication, therapeutic usage, ROA (which is closely related to individual dose forms) or drug strength. For example, the corticosteroid prednisolone can be administered as an anti-inflammatory agent in the therapeutic usage of dermatology (D07AA03) or ophthalmology (S01CB02). It can have a Dose Form of nasal drops and sprays (R01AD02). It is also prescribed for oral or parenteral route of administration as an immunosuppressant (H02AB06), or as a topical vasoprotective (C05AA04).

ATC 5th codes that are unique for an ingredient we mapped directly to the corresponding RxNorm Ingredient. Ambiguous ATC 5th codes, which are those where more than one of them encodes the same ingredient, the additional attributes become relevant. For each of these, in addition to the ingredient cross-walk we assigned the appropriate RxNorm Dose Form concepts that matches a given ROA or therapeutic use. A portion of the ATC 5th codes did not explicitly state the therapeutic use or ROA; therefore, we inferred this information from the ATC 4th, 3rd, or 2nd levels above. For example, for the R01AD02 prednisolone we gleaned the relevant Dose Forms from ATC 3rd R01A "Decongestants and other nasal preparations for topical use". Finally, we split multicomponent drugs into sets of separate substances and processed them separately. For instance, G03AB08 “dienogest and estradiol” was broken into “dienogest” and “estradiol”, and subsequently mapped to RxNorm ingredients dienogest (RXCUI, 22968) and estradiol (RXCUI, 4083).

After deriving combination and Dose Form attributes, we selected the highest and most appropriate RxNorm concepts based on the similarity of attributes using an automated script. We then followed the RxNorm hierarchy to expand the mapping to all possible descendants. If descendants overlap, we established a ranking system to prioritize matching based on ATC attribute complexity: First, precise combinations where all components were defined (e.g., N02AJ13 tramadol and paracetamol); second, the combinations with broader groups (e.g. N02BE71 paracetamol, combinations with psycholeptics), and last, single-substance ATC 5th concepts (N02BE01 paracetamol). In this way, we eliminated pseudo-duplicates that might have been assigned other ATC 5th-level codes.

**Results**

Of all 4,964 ATC 5th-level concepts we mapped 3,809 (77%) to RxNorm concepts. Among the mapped ATC codes, 518 (10.44% of the total) were unambiguous and covered any RxNorm concept containing that ingredient with no additional attribute constraints, while the others underwent the above heuristic. Unmapped codes include unapproved ingredients in the US (e.g. A08AA06 etilamfetamine), combinations of ingredients not marketed here (e.g. A10BD12 pioglitazone and sitagliptin), and ingredients with non-typical ROA (e.g. V10AA03 yttrium (90Y) silicate colloid). Looking from the RxNorm coverage 28,133 out of a total of 33,903 of Clinical Drugs (83%) have an ATC ancestor. That is up from the (27.7%) RxNorm coverage that National Library of Medicine cross-walks provide. It includes not only monoingredient drugs, but also combinations and complicated and ambiguous groups of drugs such as vaccines and insulins. Since ATC hierarchy is not comprehensive, we do not expect to cover all drugs, but our approach allows to automatically expand the mappings once ATC is updated.

We noticed high level of variations in mapping distribution per ATC codes ranging from 2 to 23,499 drug concepts in RxNorm (average 443 concepts per one ATC code). Compared to the NLM-provided cross-maps, our approach yielded 88% overlap (Figure 1), and covered 371 additional concepts as we also processed multi-component drugs and combinations2.



***Figure 1. The comparison of ATC-to-RxNorm mapping performance between the NLM-provided relationships and yielded from our approach.***

**Conclusion**

Our semi-automated mechanism preserves the correct mapping between ATC and drugs, but also maintains a semantically correct assignment of ingredients. This will allow OMOP vocabulary users to use ATC as the standard classification system for drug products. We are currently working on the mapping to extend the mapping to remaining ATC codes that will mostly cover indications and non-therapeutic agents (e.g., media contrasts).

**References**

1. WHO Collaborating Centre for Drug Statistics Methodology, Guidelines for ATC classification and DDD assignment 2019. Oslo, Norway, 2018.

2. Bodenreider O, Rodriguez LM. Analyzing U.S. prescription lists with RxNorm and the ATC/DDD Index. AMIA Annu Symp Proc. 2014;2014:297-306. Published 2014 Nov 14.