Applying ATC and DDD methodologies to real-world data

Dr Kerry Atkins
Secretary, Drug Utilisation Sub Committee of the Pharmaceutical Benefits Advisory Committee
Director, Drug Utilisation Section, Technology Assessment and Access Division. Australian Government Department of Health and Aged Care

Outline

- Identifying appropriate RWD sources.
- Assessing data quality and suitability.
- Data preparation for mapping to ATC codes and DDD.
- Automated methods for text recognition and matching.
- Assessing the representativeness of RWD for DDD analysis.
- Worked examples for data processing steps.

Identifying appropriate RWD sources for drug utilisation studies

- Procurement or sales data, e.g. IQVIA.
  - Can be cost prohibitive.
- Data on the dispensed use of medicines:
  - Administrative databases from national claims (reimbursement) and electronic health databases
    - usually most comprehensive source with standardised drug names, disease codes and information to derive DDDs (dose, treatment duration and co-administered use).
      - e.g. for US data, request from the Centers for Medicare & Medicaid Services.
      - For Europe, request from European Medicines Agency or Eurostat, statistical office of the European Union.
  - Community pharmacies and retailers selling medicines
    - In particular to examine private consumption.
  - Private hospitals
  - Private health insurance companies.

Potential insights on different socio-economic factors compared to populations using medicines reimbursed through public health systems

Assessing the quality of RWD (1)

Guidelines:

- European Medicines Agency
- U.S Food and Drug Administration guidelines and frameworks
- Health Canada

Key considerations:

Provenance – i.e. nature of the data source. Location(s) of data collection, patient details, data collection methods, clinical coding and data management methods.

Relevance

- Are the patient characteristics and health care system included in the data asset consistent with the target population of interest for research?
- How contemporary is the data, does it reflect current clinical practice including recommended dosing?

Assessing the quality of RWD (2)

Check for the inclusion of required data elements:

- Medicine name. What naming convention(s) are used?
  - e.g. official standard international nonproprietary name (INN), United States adopted name (USAN) and the British approved name (BAN). Is standardisation required to allow mapping to ATC and DDD?
- Is information about the medical condition/diagnosis available and how is this recorded?
  - e.g. International Classification of Diseases (ICD), SNOMED CT system.
  - Does the indication used as the basis to assign an ATC or DDD from the ATC/DDD source match the medical conditions that are treated by the medicines captured in the data asset?

Data preparation

Text normalisation and preprocessing:

- Remove any **punctuation** (e.g., commas, hyphens etc.).
- Convert to **relevant case**.
- Remove **extra whitespace**.
- **Normalize text** (e.g., special characters, accents, and diacritics by converting them to their base forms).
- Perform **tokenization**: Split text into individual words or tokens, which can help in further processing.
- **Remove Stop Words** (e.g., "and", "the", "of")

Unit conversions:

Ensure amount of drug consumed/supplied data is consistent with the source DDD units.

Data cleaning

Example data processing code (Python software):

```
import pandas as pd
import re
# Example data for the table
data = {
    'PIN': [1, 2, 3, 4],
    'drug_name': ['ASPRIN', 'Ibuprofene 200mg',
'Paracetemol/ 500mg', '    Metformine 1000mg']
}
# Create DataFrame
df = pd.DataFrame(data)

# Function to clean drug names
def clean_drug_name(drug_name):
    # Convert to lowercase
    drug_name = drug_name.lower()
    # Remove punctuation
    drug_name = re.sub(r'[^\w\s]', '', drug_name)
    # Remove extra whitespace
    drug_name = re.sub(r'\s+', ' ', drug_name).strip()
    # Remove dosage information (e.g., 500mg)
    drug_name = re.sub(r'\d+mg', '', drug_name).strip()
    return drug_name

# Apply the cleaning function to the drug_name column
df['cleaned_drug_name'] = df['drug_name'].apply(clean_drug_name)

# Output table with PIN and cleaned drug_name
output_table = df[['PIN', 'cleaned_drug_name']]

print(output_table)
```

	PIN	cleaned_drug_name
0	1	asprin
1	2	ibuprofene
2	3	paracetemol
3	4	metformine

Deal with invalid drug names in next step

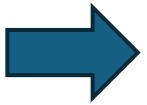
Mapping ATC codes and DDDs to data records

- Look for common issues, particularly for human-entered data into open-ended, free-text response fields:
  - misspelt generic drug names
  - entry of brand names
  - inclusion of multiple drugs or comments in the drug name record.
- Standardise generic drug names where possible to be consistent with the ATC source. e.g. Aspirin is included as ‘acetylsalicyclic acid’ in WHO ATC/DDD Index.
- Check ATC source for multiple codes and remove irrelevant codes as needed. e.g. Aspirin (acetylsalicylic acid) has 3 codes: A01AD05, B01AC06, N02BA01
- Attempt an initial matching to the ATC source information and assess mapping performance.
- For remaining unmatched records, consider automated methods to reduce the need for manual curation, which can be time intensive, to translate drug names for ATC and DDD matching. Common approaches include:
  - **Fuzzy matching:** handle partial matches and typographical errors (e.g. use of algorithms such as Levenshtein distance, Jaccard similarity, Soundex).
  - **Rule-Based Matching**: Custom rules to handle common variations and abbreviations. For example, "Acetaminophen" and "Paracetamol" can be mapped to the same drug name using predefined rule.
  - **Ontology-Based Matching**: Using medical ontologies like RxNorm, which standardize drug names and their relationships to match different variants of drug names to a common reference.

Text matching drug names

```
import pandas as pd
from fuzzywuzzy import process

# Example data for source
data_table_one = {
    'PIN': [1, 2, 3,],
    'drug_name': ['Asprin', 'Paracetemol']
}

# Example data for reference table with ATC5 codes
data_table_two = {
    'drug_name': ['Aspirin', 'Paracetamol'],
    'ATC5_code': [' N02BA01', 'N02BE01']
}

# Create DataFrames
df_table_one = pd.DataFrame(data_table_one)
df_table_two = pd.DataFrame(data_table_two)

# Function to perform fuzzy matching and get ATC5 code
def get_atc5_code(drug_name, reference_table):
    match = process.extractOne(drug_name, reference_table['drug_name'])
    if match and match[1] > 80:  # Threshold for matching
        return reference_table.loc[reference_table['drug_name'] == match[0],
'ATC5_code'].values[0]
    else:
        return None

# Apply the function to get ATC5 codes
df_table_one['ATC5_code'] = df_table_one['drug_name'].apply(lambda x: get_atc5_code(x,
df_table_two))

# Output table
output_table = df_table_one[['PIN', 'drug_name', 'ATC5_code']]

print(output_table)
```

Using fuzzy logic to match invalid drug names to the ATC code source



PIN	drug_name	ATC5_code
1	Asprin	N02BA01
2	Paracetemol	N02BE01

Calculating DDDs

Example: Are patients exceeding the recommended dose of aspirin of 4000 mg per day for the treatment of general headache?

PIN	drug_name	quantity_consumed	supply_days	ICD10
1	Aspirin	1.6	4	R51.9
2	Aspirin	6.4	2	K12.1
3	Aspirin	1.5	3	R51.9
4	Aspirin	8.5	5	R51.9
5	Aspirin	0.2	1	K12.1

Diagnosis codes for stomatitis

```
import pandas as pd
# Sample data for the table
data = {
    'PIN': [1, 2, 3, 4, 5],
    'drug_name': ['Aspirin', 'Aspirin', 'Aspirin', 'Aspirin', 'Aspirin'],
    'quantity_consumed': [1.6, 6.4, 1.5, 8.5, 0.2],
    'supply_days': [4, 2, 3, 5, 1],
    'ICD10': ['R51.9', 'K12.1', 'R51.9', 'R51.9', 'K12.1']
}
# Create DataFrame
df = pd.DataFrame(data)
# Exclude patients with ICD10 value K12.1
df_filtered = df[df['ICD10'] != 'K12.1']
# Define the DDD value
DDD_value = 3
# Calculate defined daily dose per day
df_filtered['DDD_per_day'] = (df_filtered['quantity_consumed'] /
df_filtered['supply_days']) * DDD_value
print(df_filtered)
```

PIN	quantity_consumed	supply_days	ICD10	DDD_day
1	1.6	4	R51.9	1.2
3	1.5	3	R51.9	1.5
4	8.5	5	R51.9	5.1

References

- Hollingworth, S.; Kairuz, T. Measuring Medicine Use: Applying ATC/DDD Methodology to Real-World Data. Pharmacy 2021, 9, 60.
- Ioakeim-Skoufa I, Atkins K, Hernández-Rodríguez MÁ. Optimizing real-world evidence studies for regulatory decision-making and impact assessment in pharmacovigilance. Br J Clin Pharmacol. 2025 Jan 17. doi: 10.1111/bcp.16393. Epub ahead of print. PMID: 39821103.
- Kellmann, A.J., Lanting, P., Franke, L. et al. Semi-automatic translation of medicine usage data (in Dutch, free-text) from Lifelines COVID-19 questionnaires to ATC codes. Database (2023) Vol. 2023: article ID baad019; DOI: https://doi.org/10.1093/database/baad019
- Ostropolets, A., Abedtash, H., Rijnbeek, P. et al. Combining the ATC Drug Classification System with the RxNorm Drug Nomenclature into a comprehensive Drug Ontology: Challenges and Achievements. Observational Health Data Sciences and Informatics symposium, 2019. Accessed at: https://www.ohdsi.org/2019-us-symposium-showcase-20/
- Quint, J; Brownrigg, A. What Should Clinicians Know About How Coding Influences Epidemiological Research? AMA Journal of Ethics January 2025, Volume 27, Number 1: E51-57.
- WHO ATC/DDD Toolkit, ‘Sources of drug utilization data’. Accessed at: https://www.who.int/tools/atc-ddd-toolkit/data.