Accuracy of drug-related information on Wikipedia

Luca Bartek

18th February, 2016
Why are we interested?

• Free, well-known, multilingual resource
• Wikipedia has over 10,000 drug-related pages
• Wikipedia is on the first page of drug-related Google search results $\sim$ 80% of the time$^1$
• 51% of students of Medicine use Wikipedia as source$^2$
• Has been researched previously – but only small-scale$^3$

What kind of information is available on Wikipedia?

**Title.** Also ID. E.g. http://en.wikipedia.org/wiki/Aspirin

- **Main content,** contains most information, freely editable
- **Infobox** (drugbox). Contains specific information for each type of article, has a template.
What data was retrieved and how?

• Wikipedia API (mediawiki.org)
• Get a list of pages in which an “Infobox drug” Template is embedded
• Use list to retrieve timestamps for all revisions, and the content of section 0 for the latest revision
• Extract physical, chemical properties and database ID’s including those which can be confirmed by OPS
What data was retrieved?

Physical & Chemical Properties
• Smiles
• Standard & Non-Standard InChIs
• Standard & Non-Standard InChIKeys
• Molecular weight
• Molecular formula

Database ID’s
• CAS number
• DrugBank ID
• ChEMBL ID
• ChEBI ID
• PubChem ID
• KEGG ID
• UNII ID
• ATC code
How to verify all this data?

• Does all the information match with each other?
  – Can check for physical/chemical data

• Does this data actually correspond to given drug? Do the ID’s also match?
  – A reliable “gold standard” is needed – enter Open PHACTS

• Does the number of edits / age of the article have an effect on accuracy?
Question 1 – Internal consistency

Completeness of Wikipedia for each examined property

- CAS number: 87%
- ATC code: 53%
- ChEBI: 22%
- PubChem: 27%
- DrugBank: 79%
- ChemSpiderID: 59%
- UNII: 36%
- KEGG: 36%
- ChEMBL: 91%
- molecular weight: 87%
- molecular formula: 83%
- smiles: 65%
- InChI: 62%
- InChIKey: 73%
- StdInChI: 73%
- StdInChIKey: 73%
Question 1 – Internal consistency

6013 molecules

856 enzymes, mixtures, proteins, antibodies

5157 small molecules

230 molecules with no SMILES string

4927 with SMILES

17 molecules with invalid SMILES

4910 with valid SMILES
Question 1 – Internal consistency

4910 with valid SMILES

4249 with StdInChI

3970 molecules (93%) correct

3959 with InChI and InChIKey correct

4275 with StdInChIKey

4023 molecules (94%) correct

3924 with MolFormula defined

3796 with correct MolFormula (97%)

3700 with Molweight within 1g (97%);
3600 with Molweight within 0.1g (95%)

3700 molecules are fully complete and consistent:
72% of all examined small molecules
Question 2 – External verification

Input for Open PHACTS?
Need a URI.

Drugs matched via
identifiers.org \(\rightarrow\) HMDB

• Identifiers.org URI’s created in the format
http://info.identifiers.org/wikipedia.en/drug

• Used as an input to Map URI API call

• Retrieve HMDB URI

• Successful for 1158 drugs

Total of 4031 drugs matched between
Open PHACTS and Wikipedia.

Drugs matched via Concept
search

• Compound names used as an input for Get
concept description API call, using the UUID
for Chemical viewed structurally

• Exact matches are annotated:
  <em>Drug</em>

• 2609 drugs successfully matched with this
method

• Another 264 matched manually (more than
1 <em>)
Question 2 – External verification

• Previously acquired URI’s used as input
• Retrieving ID’s – Using the Map URI Call
• `targetUriPattern` parameter used to filter results
• The identifiers were taken from the result URI’s via regular expressions
• Acquired identifiers:
  - PubChem
  - CAS
  - ChEBI
  - ChemSpider
  - DrugBank
  - ChEMBL
  - OPS
  - (KEGG)
  - (HMDB)
Question 2 – External verification

• Retrieving physical/chemical data using the Compound information API call
• OPS URI is used as an input (no OPS URI = no compound information)
• The following properties were extracted:
  - Smiles
  - Molecular weight
  - Molecular formula
  - InChIKey
  - InChI
Question 2 – External verification

Completeness of Open PHACTS for each examined property

- PubChem: 33%
- CAS: 34%
- KEGG: 21%
- ChEBI: 29%
- ChemSpider: 100%
- HMDB: 34%
- DrugBank: 34%
- ChEMBL: 91%
- OPS-uri: 99%
- smiles: 97%
- Molecular weight: 96%
- Molecular formula: 96%
- InChIKey: 97%
- InChI: 97%
Question 2 – External verification

Method:

• For ID’s: string matching

• SMILES:
  – A molecule is created from both smiles
  – It is re-converted into standard InChIs

• InChI & InChIKey: string matching

• Molecular formula: an array with the atom types and the numbers in each element

• Molecular weight: rounded to whole numbers
Question 2 – External verification

% of Correct Properties on Wikipedia Taking Open PHACTS as a Gold Standard

<table>
<thead>
<tr>
<th>Source</th>
<th>0%</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubChem</td>
<td>84%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAS</td>
<td>94%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChEBI</td>
<td>96%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChEMBL</td>
<td>87%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChemSpider</td>
<td>87%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>molformula</td>
<td>95%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>molweight</td>
<td>94%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InChI</td>
<td>87%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InChIKey</td>
<td>88%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smiles</td>
<td>84%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Question 2 – External verification

- Lowest accuracy: SMILES
  - 64% of cases: stereochemistry
  - Salt vs. parent drug e.g. Edrophonium vs Edrophonium Chloride
    - © vs @ in smiles
- ID with lowest accuracy:
  - PubChem
Question 3 – What influences completeness/correctness on Wikipedia?

- Two scores created
- CompleteScore:
  - Measure of how complete data is
    \[ \text{CompleteScore} = \# \text{ data defined} \]
- WikiScore
  - Measure of both quality and completeness of data
    \[ \text{WikiScore} = (\# \text{ correct}) \times 1 + (\# \text{ incorrect}) \times (-1) \]
Question 3 – What influences completeness/correctness on Wikipedia?

Average WikiScore and CompleteScore vs. Year of Article Creation
Question 3 – What influences completeness/correctness on Wikipedia?

Number of revisions vs. Year of Article Creation
Question 3 – What influences completeness/correctness on Wikipedia?

WikiScore and Completescore Vs. Number of Revisions
Main outcome & What’s next

• Confidence in the accuracy of Drug related information on Wikipedia

• Mapping between OPS and Wikipedia for over 4000 drugs
  – can be added to OPS mapping
  – OPS number can be added to Infoboxes

• Flagged some typical issues with Wikipedia data
  – update Wikipedia via API or publishing results

• Identified some issues where OPS mapping is imperfect
Acknowledgements

- Stefan Senger (external advisor)
- Tell Tuttle (project supervisor)
- Biovia (provided Pipeline Pilot license)
Thank you for your attention.