There is a Need to Integrate Clinical Use with Active Ingredients, Pharmaceutical Products & Associated Information at the Molecular Level

Tudor I. Oprea

8/29/2017
OHDSI
Collaborator meeting via Webex
Albuquerque, NM

http://targetcentral.ws/
http://pharos.nih.gov
http://drugcentral.org
http://newdrugtargets.org
Initially designed to answer “how many drugs are out there”…

- The Two Cultures: what patients and docs call “drugs” (products) vs. what scientists call “drugs” (active pharmaceutical ingredients)
- Also wanted to know how many drug targets there are………

- Total number of active ingredients: ~4500
- This includes API approved for human use worldwide, FDA approved and discontinued
- ~1500 are currently marketed and FDA approved, ~300 are discontinued
Several online resources contain important drug information.

To facilitate data analysis we have mapping of active ingredients to most relevant drug information resources online.

Most mappings were done using generic names and structure.

These drug resources provide information on regulatory status, publications, pharmacology, biological activity profiles, etc.
Mapping of drugs to external resources ranges from 13% (PDB Ligands) to 100% (CAS registry numbers)
DAILYMED DRUG LABELS (FDA)

- Drug labels in SPL (Structured Product Label) format
- Updated Daily
- Text in sections annotated with LOINC codes
  - Summary of clinical trial results
  - Contraindications, adverse events, warnings, therapeutic dose, etc.
- Table with active/inactive ingredients, strength, route of administration
  - NDA, ANDA, UNII identifiers
- DailyMed is the main source of information on pharmaceutical products. We use custom processing pipelines that extract text from SPL separated by LOINC sections.
- Dose, formulations and active ingredients tables are parsed and mapped to the main active ingredients table.
- Pharmaceutical formulations containing herbals, allergens, etc. products are discarded
- We do not process homeopathic labels and SPL files for devices.
Active ingredients in Rx products only form more than 82% of the total number of active ingredients.

However, when compared total number of pharmaceutical products OTC only active ingredients have 46% share.

HOW MANY APIs PER PRODUCT?

Most of the pharmaceutical products contain 1 active ingredient,

Most of the products with 2 or more active ingredients are usually OTC.
There are almost as many “OTC” as Rx drugs, but with far less APIs.

Over 5000 drug labels contain acetaminophen (84 unique API fixed-dose combinations).

<table>
<thead>
<tr>
<th>Type</th>
<th>OTC</th>
<th>PRESCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>APIs</td>
<td>284</td>
<td>1,562</td>
</tr>
<tr>
<td>Drugs (”drug labels”)</td>
<td>46,770</td>
<td>43,172</td>
</tr>
</tbody>
</table>


8/29/17 revision
Reckitt Benckiser sells:
- Nurofen Back Pain,
- Nurofen Period Pain
- Nurofen Migraine Pain and
- Nurofen Tension Headache

at twice the price compared to Nurofen, even though it contains exactly the same active ingredient (342mg of ibuprofen lysine, equivalent to 200mg of ibuprofen).
TAKE HOME MESSAGE 1

PHARMACEUTICAL PRODUCTS ARE AN EQUALLY IMPORTANT COMPONENT OF DRUG RESEARCH
**PHARMACOLOGIC CLASSIFICATIONS**

- DrugCentral integrates pharmacologic classifications from ATC, MeSH, ChEBI, and FDA.
- These provide systematic groupings of drugs based on common therapeutic applications and mechanism of action.

---


8/29/17 revision
Because most of the drugs' Mechanism of Action is mediated by protein targets, DrugCentral collects and combines data on biological activity profile from multiple sources.

- The ChEMBL database is the primary source of MoA data.
- Median target binding data shows that drugs targeting GPCR, NR, and Kinases are among the most potent drugs with potency in low nM range.
Drugs distributed by ATC codes (levels 1-2). Concentric rings indicate ATC levels. Histograms represent the number of drugs distributed per year of first approval. Maximum scale: 100.
## Commercial Impact of Target Classes

<table>
<thead>
<tr>
<th>Target Class</th>
<th>Targets</th>
<th>APIs</th>
<th>Sales (B USD)</th>
<th>Market Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPCR</td>
<td>72</td>
<td>372</td>
<td>889.17</td>
<td>27.42%</td>
</tr>
<tr>
<td>Enzyme</td>
<td>88</td>
<td>234</td>
<td>683.14</td>
<td>21.06%</td>
</tr>
<tr>
<td>Nuclear receptor</td>
<td>16</td>
<td>111</td>
<td>340.13</td>
<td>10.49%</td>
</tr>
<tr>
<td>Transporter</td>
<td>18</td>
<td>82</td>
<td>323.99</td>
<td>9.99%</td>
</tr>
<tr>
<td>Ion channel</td>
<td>23</td>
<td>167</td>
<td>281.11</td>
<td>8.67%</td>
</tr>
<tr>
<td>Kinase</td>
<td>43</td>
<td>49</td>
<td>240.46</td>
<td>7.41%</td>
</tr>
<tr>
<td>Cytokine</td>
<td>9</td>
<td>12</td>
<td>184.29</td>
<td>5.68%</td>
</tr>
<tr>
<td>Other</td>
<td>43</td>
<td>68</td>
<td>300.83</td>
<td>9.28%</td>
</tr>
</tbody>
</table>

What are the most lucrative targets from a therapeutic perspective? We evaluated data from IMS Health on drug sales from 75 countries, including Europe, North America and Japan, aggregated over a 5-year period (2011–2015). After excluding botanicals, traditional Chinese and homeopathic medicines and drugs perturbing non-human targets, we identified 51,095 unique products. These were mapped to 1,069 active pharmaceutical ingredients from DrugCentral, corrected by number of APIs per product, then by number of Tclin targets per API.
## Top 20 Drug Targets by Revenue

<table>
<thead>
<tr>
<th>Gene</th>
<th>Protein Target</th>
<th>Action</th>
<th>Sales (B USD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNF</td>
<td>Tumor necrosis factor</td>
<td>Immunosuppressants</td>
<td>163.39</td>
</tr>
<tr>
<td>INSR</td>
<td>Insulin receptor</td>
<td>Hypoglycemic agents</td>
<td>143.55</td>
</tr>
<tr>
<td>NR3C1</td>
<td>Glucocorticoid receptor</td>
<td>Anti-inflammatory</td>
<td>142.75</td>
</tr>
<tr>
<td>HMGCR</td>
<td>3-hydroxy-3-methylglutaryl-coenzyme A reductase</td>
<td>Hypolipidemic agents</td>
<td>122.55</td>
</tr>
<tr>
<td>ATP4A/ATP4B</td>
<td>Proton Pump (K⁺ ATP-ase)</td>
<td>Anti-ulcer agents</td>
<td>118.16</td>
</tr>
<tr>
<td>AGTR1</td>
<td>Type-1 angiotensin II receptor</td>
<td>Antihypertensive agents</td>
<td>99.98</td>
</tr>
<tr>
<td>ADRB2</td>
<td>Beta-2 adrenergic receptor</td>
<td>Bronchodilators</td>
<td>90.02</td>
</tr>
<tr>
<td>OPRM1</td>
<td>Mu-type opioid receptor</td>
<td>Analgesics</td>
<td>87.97</td>
</tr>
<tr>
<td>PTGS2</td>
<td>Cyclooxygenase-2</td>
<td>Analgesics</td>
<td>84.04</td>
</tr>
<tr>
<td>DRD2</td>
<td>D2 dopamine receptor</td>
<td>Antipsychotic agents</td>
<td>74.91</td>
</tr>
<tr>
<td>CHRM[1-5]</td>
<td>Muscarinic acetylcholine receptor</td>
<td>Anticholinergics</td>
<td>64.16</td>
</tr>
<tr>
<td>SLC6A4</td>
<td>Sodium-dependent serotonin transporter</td>
<td>Antidepressants</td>
<td>59.18</td>
</tr>
</tbody>
</table>

### Gene | Protein Target | Action | Sales (B USD) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HTR2A</td>
<td>5-hydroxytryptamine receptor 2A</td>
<td>Antipsychotics</td>
<td>57.58</td>
</tr>
<tr>
<td>CACNA1S/CACNA1C/CACNA1D/CACNA1F</td>
<td>L-type calcium channel</td>
<td>Antihypertensive agents</td>
<td>55.97</td>
</tr>
<tr>
<td>SLC6A2</td>
<td>Sodium-dependent noradrenaline transporter</td>
<td>antidepressants &amp; psychostimulants</td>
<td>55.72</td>
</tr>
<tr>
<td>VEGFA</td>
<td>Vascular endothelial growth factor A</td>
<td>antineovascularisation agents</td>
<td>55.15</td>
</tr>
<tr>
<td>HRH1</td>
<td>Histamine H1 receptor</td>
<td>antihistamines</td>
<td>53.55</td>
</tr>
<tr>
<td>IFNAR1/IFNAR2</td>
<td>Type I interferon receptor</td>
<td>immunostimulants</td>
<td>51.40</td>
</tr>
<tr>
<td>SCN[1,2,3,4,5,7,8,9,10,11]A</td>
<td>Voltage-gated sodium channel</td>
<td>antiarrhythmics &amp; antiepileptics</td>
<td>50.64</td>
</tr>
<tr>
<td>ESR1</td>
<td>Estrogen receptor</td>
<td>contraceptives / estrogen agonists</td>
<td>50.35</td>
</tr>
</tbody>
</table>

T. Oprea et al., *Nature Rev. Drug Discov.* poster, Jan 2017
By combining information for drug indications, targets, pharmacologic class, and structures, it is possible to get a quick overview for different areas of therapeutic interest, as an example drugs for diabetes.
ONTIOLOGY-BASED CAPTURE OF THERAPEUTIC INTENT FROM DRUG INDICATIONS

Pimavanserin
- treats
- is contraindicated in

Hallucinations
- symptom
- has disease context
- Parkinson's disease
disease

Delusions
- symptom
- has disease context
- Parkinson's disease
disease

Psychosis
- symptom
- has disease context
- Dementia
disease

Lumacaftor
- treats

Cystic fibrosis
- disease
- has mutated gene
- CFTR
- has mutation
- F508del
mutation

Medical intent
CURATION TOOL FOR ANNOTATING DRUG INDICATIONS

An atypical antipsychotic, mechanism of action of pimavanserin in the treatment of hallucinations and delusions associated with Parkinson's disease psychosis is unknown.

NUPLAZID™ is indicated for the treatment of hallucinations and delusions associated with Parkinson's disease psychosis.

Subjectively experienced sensations in the absence of an appropriate stimulus, but which are regarded by the individual as real. They may be of organic origin or associated with MENTAL DISORDERS.
Live presentation should follow (Oleg Ursu)
TAKE HOME MESSAGE 2

LINKING DRUGS TO TARGETS AND INDICATIONS GUIDES FURTHER RESEARCH
Most protein classification schemes are based on structural and functional criteria.

For therapeutic development, it is useful to understand how much and what types of data are available for a given protein, thereby highlighting well-studied and understudied targets.

- Proteins annotated as drug targets are **Tclin**
- Proteins for which *potent* small molecules are known are **Tchem**
- Proteins for which biology is better understood are **Tbio**
- Proteins that lack antibodies, publications or Gene RIFs are **Tdark**
Bioactivities of approved drugs (by Target class)

- **Tclin** proteins are associated with drug Mechanism of Action (MoA)
- **Tchem** proteins have bioactivities in ChEMBL and DrugCentral, + human curation for some targets
  - Kinases: $\leq 30\,\text{nM}$
  - GPCRs: $\leq 100\,\text{nM}$
  - Nuclear Receptors: $\leq 100\,\text{nM}$
  - Ion Channels: $\leq 10\,\mu\text{M}$
  - Non-IDG Family Targets: $\leq 1\,\mu\text{M}$

**ChEMBL**: database of bioactive chemicals
[https://www.ebi.ac.uk/chembl/](https://www.ebi.ac.uk/chembl/)

**DrugCentral**: online drug compendium
[http://drugcentral.org/](http://drugcentral.org/)
**D-T DEVELOPMENT LEVEL 2**

- **Tbio** proteins lack small molecule annotation cf. Tchem criteria, and satisfy one of these criteria:
  - protein is above the cutoff criteria for **Tdark**
  - protein is annotated with a GO Molecular Function or Biological Process leaf term(s) with an Experimental Evidence code
  - protein has confirmed [OMIM](https://omim.org) phenotype(s)

- **Tdark** ("understudied proteins") have little information available, and satisfy these criteria:
  - PubMed score (text mining) from [Jensen Lab](http://www.jensenlab.org) < 5
  - <= 3 Gene RIFs
  - <= 50 Antibodies available according to [antibodypedia.com](http://www.antibodypedia.com)

---

Fractional paper count

$$
\text{PubMed score} = \sum_{j \in D} \frac{n_{ij}}{n_j}
$$
We used name entity recognition software (from LJ Jensen’s lab) to evaluate nearly ~27 million abstracts (including ~2M full text articles) to derive a publication score per protein.
Nr of antibodies reflects our ability to characterize proteins. Understudied proteins have fewer such tools.
**TDL: EXTERNAL VALIDATION**

Tdark parameters differ from the other TDLs across the 4 external metrics cf. Kruskal-Wallis post-hoc pairwise Dunn tests

T. Oprea et al., *Nature Rev. Drug Discov. poster*, Jan 2017
“Counts by Name” == users accessing the STRING website and typing in a gene symbol. “Counts by Link” == users accessing the network for a gene in STRING by linking to it from another resource.

Data courtesy of Christian von Mehring, STRING-db
TAKE HOME MESSAGE 3

THERE IS A KNOWLEDGE DEFICIT

over 37% of the proteins remain understudied (Tdark)

~10% of the Proteome (Tclin & Tchem) can be targeted by small molecules
The absence of a quantitative language “is the flaw of biological research” or “The more facts we learn the less we understand”.

A biologist describing a Radio:

Src: Serendipitously Recovered Component (wire connecting to the antenna, which is) 
Mic: Most Important Component 
but you really need 
Ric: Really Important component (rectifier) 
and U-Mic (Undoubtedly Most Important Component) [the switch]

When little is known, don’t expect knowledge to accumulate quickly
CONCEPTUAL FALLACY: SEPARATION BY ORGAN/CASE

- Medicine maintains this separation for necessity: by organ (e.g., cardiology, ophthalmology), by disease category (e.g., oncology, infection).

- NIH Institutes are organized in a similar way.
- Many pharma companies are organized by Therapy Area.
- … yet genes / proteins / pathways do not observe such separation.
- The impact of this “mental divide” in science has yet to be understood.
PRE-CHRONIC KIDNEY DISEASE (5 YEARS)

A.B. Jensen et al., Nature Communications 2014 5:4022
DISEASES ARE CONCEPTS

- Diseases lack physical manifestation outside patients.
- **The search for cures has to be patient centered**
- …Animal models should be combined with patient data mining
- Remember David Horrobin’s papers…

Modern biomedical research: an internally self-consistent universe with little contact with medical reality?

David F. Horrobin
Illuminating the Druggable Genome Knowledge Management Center

~27 Million Papers
~7 million Patents
~62 million Patients

~15,000 Diseases

Seeking New Knowledge

~20,000 Proteins
~4,500 Drugs
DrugCentral is part of our translational informatics division.