Towards Personalized Medicine: Leveraging Patient Similarity and Drug Similarity Analytics

Ping Zhang, Fei Wang, Jianying Hu, Robert Sorrentino

Healthcare Analytics Research Group
IBM T.J. Watson Research Center
Disclosure and learning objective

- Disclosure: All authors are employees of IBM

- Learning Objective: After participating in this activity the learner should be better able to:
  - Recognize the benefits of adopting real-world evidence for personalized medicine
  - Measure drug-drug and patient-patient similarities from different aspects
  - Construct a heterogeneous graph to encode patient similarity, drug similarity, and patient-drug prior associations
  - Formulate a label propagation approach to spread the label information representing the effectiveness of different drugs for different patients over the heterogeneous graph
Outline

- Background Introduction
- Our Methodology
- Experimental Results
- Future Works
Outline

- Background Introduction
- Our Methodology
- Experimental Results
- Future Works
Moving towards personalized medicine

- Personalized Medicine: the right patient with the right drug at the right dose at the right time.
  - (for patients) the end of one size fits all drugs would result in safer and more effective treatments
  - (for doctors) reduce wasted time for patients and resources with futile treatments
  - (for pharms) lower cost marketing due to targeted patients, faster clinical trials, less focus on animal trials

The figure is from https://3dbiomatrix.com/a-new-dimension-in-personalized-medicine-2/
Real world data (RWD): an additional resource

- Personalized medicine appears to benefit from advances from "omics" (genomics, proteomics, metabolomics, etc.), but
  - “Omnics” information is not yet widely available in everyday clinical practice
  - Other than "omics", numerous external factors (e.g., environment, diet and exercise) affect response to medication

- RWD are clinical observations other than randomized clinical trials (RCT).
  - RWD are observations on human in the clinical stage, so there is less of a translational issue.
  - RWD is not only vast but also varied in type and source: electronic medical records (EMR), claims data, and even social media.
Patient similarity and drug similarity analytics

- Patient Similarity analytics: Find patients who display similar clinical characteristic to the patient of interest
- Drug Similarity analytics: Find drugs which display similar pharmacological characteristic to the drug of interest
- Resulting insights: medical prognosis, risk stratification, care planning (especially for patients has multiple diseases)
- Resulting insights: drug repositioning, side-effect prediction, drug-drug interaction prediction

How to leverage both patient similarity and drug similarity for personalized medicine?
Outline

- Background Introduction
- Our Methodology
- Experimental Results
- Future Works
Drug personalization problem: whether drug A is likely to be effective for specific patient B. To take into consideration the specific condition of patient B as well as the characteristics of drug A, we should leverage the information of:
- The patients who are clinically similar to patient B
- The drugs which are similar to drug A
- Prior associations between patients and drugs, which are measured by diagnosis of patients and therapeutic indications of drugs

Heterogeneous patient-drug graph $A$:

$$A = \begin{bmatrix}
S_p & R \\
R^T & S_d
\end{bmatrix}$$
Label propagation method

- For each drug $d$, we constructed a corresponding effectiveness vector (i.e., known but not completed “label” vector) $y = [y_1, y_2, \ldots, y_n, y_{n+1}, \ldots, y_{n+m}]^T$, where
  
  $$y_k = \begin{cases} 
1 & (k = 1, 2, \ldots, n), \text{if } d \text{ is an effective treatment for patient } k \\
1 & (k = n+1, n+2, \ldots, n+m), \text{if } d \text{ is the } (k-n)\text{-th drug} \\
0, & \text{otherwise}
\end{cases}$$

- $W$ is a normalized form of the similarity matrix $A$.

- In each propagation iteration, the estimated score of each node “absorbs” a portion ($\mu$) of the label information from its neighborhood, and retains a portion ($1 - \mu$) of its initial label information.

- The updating rule for node $i$ is given by
  
  $$f_i^{\text{after}} \leftarrow \mu \sum_{j=1}^{n} W_{ij} f_j^{\text{before}} + (1 - \mu) y_i$$

Consider the initial condition is $f^0 = y$, we have the equation

$$f^t = (\mu W)^{t-1} y + (1 - \mu) \sum_{i=0}^{t-1} (\mu W)^i y$$

$$\lim_{t \to \infty} f^t = (1 - \mu)(I - \mu W)^{-1} y$$

- $f^*$ – the possibility when a drug is effective for a patient
Calculations of similarities and association prior

- Drug similarity evaluation
  - Drugs with similar chemical structures would carry out common therapeutic function. Each drug was represented by an 881-dimentional PubChem fingerprint. Tanimoto coefficient (TC) of two fingerprints as chemical structure similarity. \( TC(A,B) = \frac{|A \cap B|}{|A \cup B|} \)
  - Drugs sharing common targets often possess similar therapeutic function. Average of sequence similarities of target protein sets as target protein similarity.

\[
sim_{\text{target}}(d_x, d_y) = \frac{1}{|P(d_x) \parallel P(d_y)|} \sum_{i=1}^{|P(d_x)|} \sum_{j=1}^{|P(d_y)|} SW(P_i(d_x), P_j(d_y))
\]

- Patient similarity evaluation: For simplicity and consistency, we used co-occurring ICD9 diagnosis code (i.e., TC of patient ICD9 diagnosis feature vector).

- Patient-drug association prior evaluation: measured by the TC between ICD9 diagnosis of patients and ICD9-format drug indications from MEDI database.
Outline

- Background Introduction
- Our Methodology
- Experimental Results
- Future Works
Data description

- Problem: identify personalized treatments for *hyperlipidemia*.
- Raw data: a 3-year longitudinal EMR of 110,157 patients. 8 cholesterol-lowering drugs. 273,525 low-density lipoprotein (LDL) lab-test records.
- Definition of an effective drug for a patient: we selected the patients who take only one cholesterol-lowering drug within a 60-day treatment window and remain “well-controlled” (i.e., LDL<130 mg/dL) for at least two consecutive lab assessments.

- Final data: 1219 distinct patients and 4 statin cholesterol-lowering drugs
- Patient similarities were calculated based on the ICD9 codes within the 90-day patient assessment window

<table>
<thead>
<tr>
<th>Drug</th>
<th>Patient #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atorvastatin</td>
<td>97</td>
</tr>
<tr>
<td>Lovastatin</td>
<td>221</td>
</tr>
<tr>
<td>Pravastatin</td>
<td>24</td>
</tr>
<tr>
<td>Simvastatin</td>
<td>877</td>
</tr>
</tbody>
</table>
Statin Drugs

Atorvastatin
- CYP3A4, HMGCR, SLCO1B1

Lovastatin
- CYP3A4, HMGCR, RASD1
- ABCC2, HMGCR, SLCO1B1

Pravastatin

Simvastatin
- CYP3A4, HMGCR

strongest: had a heart attack or have acute coronary syndrome and LDL is highly elevated.

lower “bad” (LDL) cholesterol by less than 30 percent

stronger: LDL reduction of 30 percent or more; have heart disease or diabetes
Averaged ROC comparison of three treatment recommendation strategies based on 50 independent 10-fold cross-validation runs

- Label Propagation w/o Drug Similarity information (auc=0.7734)
- Label Propagation with Drug Structure Similarity (auc=0.8021)
- Label Propagation with Drug Target Similarity (auc=0.8361)
Outline

- Background Introduction
- Our Methodology
- Experimental Results
- Future Works
Conclusions and future works

- To support personalized medicine, we propose a heterogeneous **label propagation** method by leveraging **patient similarity** and **drug similarity** analytics.

- For further investigation
  - Explore more sophisticated drug and patient similarity measures
  - Consider dosage information in EMR
  - Define “effective” drugs for patients with more clinical significance
  - Apply the method to more drugs and diseases
Thank You

The figure is from http://dnadestiny.yolasite.com/future-opportunities.php