Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality

Hua Xu, PhD
School of Biomedical Informatics, UTHealth

JAMIA Journal Club
October 2nd, 2014
The Issue of Drug Discovery

• Developing a new drug will take:
  – $800 million
  – 10 – 17 years
  – 10% success rate

• A productivity problem of pharmaceutical industry

Drug Repurposing

• Drug repurposing (repositioning or re-profiling) – find new indications of existing drugs
  – Find new uses of FDA-approved drugs
  – Rescue drugs previously failed in clinical trials

• Advantages: known pharmacokinetic, pharmacodynamic, and toxicity profiles
  – Lower risk
  – Lower cost
  – Less time
Computational Approaches to Drug Repurposing

- Large-scale compound databases containing structure, bioassay, and genomic information, e.g., NIH’s Molecular Libraries Initiative
- Computational approaches
- Electronic health records (EHRs)
EHRs Data

• Large EHRs – millions of records for more than a decade
• By 2014, all US hospitals will implement EHRs systems
• Types of data in EHRs
  – Structured
    • Administrative data
    • Billing codes: ICD9, CPT, ...
    • Lab tests
    • Computerized orders
    • ..... 
  – Unstructured
    • Admission notes
    • Discharge summaries
    • Clinic visit notes
    • Pathology notes
    • .....
EHRs for Drug Studies

• EHRs data
  – Rich treatment and outcome information
  – Longitudinal practice-based data

• Different types of drug studies
  – Pharmacoepidemiology
  – Pharmacoeconomics
  – Pharmacovigillance
  – Pharmacogenomics (with Biobanks)
Challenges for using EHR Data

• Data extraction
  – Much of detailed information is embedded in narrative text
  – Heterogeneous data sources, e.g., different terminologies

• Data abstraction and analysis
  – Discrepancy
  – Missing data
  – Confounding

• The need for informatics approaches
A study of metformin and cancer mortality using EHR and informatics

• We want to answer two questions:
  – Can EHR data be used to find new indications of existing drugs?
  – What is the role of informatics in this type of research?

• Specific aim – validate the association between metformin and improved cancer survival rate using EHR data
Metformin and cancer survival

• Reduced Cancer Mortality with Metformin Use Among Diabetics

HR met continuous = 0.58 (0.36-0.93)

Landman Diab Care 2010
Study Design

• Primary analysis – Vanderbilt

SD - Vanderbilt De-identified EHR

Vanderbilt Tumor Registry

Tumor subjects with Diabetes in EHR

Metformin

Other oral T2D meds

Insulin Only

Tumor subjects without Diabetes in EHR

Non-T2D

• Replication – Mayo Clinic
The Use of Informatics

• Cohort Identification
  – Type 2 Diabetes algorithm developed by eMERGE
  – MedEx for determining medication exposure

• Covariates extraction
  – Smoking status
  – Height and weight
Informatics – Cohort Identification

• Type 2 Diabetes algorithm developed by eMERGE
  – ICD9 codes
  – Medications
  – Lab test

• Evaluated at Vanderbilt and other sites, high performance (PPV>95%)
Informatics – Medication Exposure

• MedEx
  – Identify drug name and signature information
  – Identify other T2D drugs
  – Determine metformin and its daily dose
  – Available at http://code.google.com/p/medex-uima/

• Heuristic rules for drug exposure
Informatics – covariate extraction

• Smoking status
  – The default cTAKES smoking status module did not work well on Vanderbilt text
  – Customized to Vanderbilt narratives – a 93% PPV for smoking status

• Height and weight
  – Structured fields - 42% height and 36% weight were missing
  – A simple regular expression program reduced missing rate to 33% and 16% for height and weight respectively
Data Extraction Workflow - Vanderbilt

Synthetic Derivative – the de-identified EHR of Vanderbilt University Hospital
(n = 2.2 million patients as May 2013)

Cancer patients (excluding skin cancer and age < 18) diagnosed between 1995 and 2010, based on Tumor Registry
(n = 44,257)

Cancer patients from above, but excluding CHF (congestive heart failure) and CKD (chronic kidney disease) before tumor diagnosis date
(n = 42,165)

Excluded – CHF or CKD
n = 2,092

Type 2 Diabetes
n = 5,796

Excluded – uncertain diabetes status due to insufficient information
n = 7,452

Metformin
n = 2,218

Other Oral DM2 drugs
n = 903

Insulin monotherapy
n = 377

Excluded – ineligible
n = 2,298

Non-Diabetics
n = 28,917
Data Extraction Workflow - Mayo

1. EHR of Mayo Clinic
   \( n = 7.4 \text{ million patients as May 2013} \)

2. Cancer patients (excluding skin cancer and age < 18) diagnosed between 1995 and 2010, based on Tumor Registry
   \( n = 102,546 \)

3. Cancer patients from above, but excluding CHF (congestive heart failure) and CKD (chronic kidney disease) before tumor diagnosis date
   \( n = 96,169 \)

4. Excluded – CHF or CKD
   \( n = 6,377 \)

5. Type 2 Diabetes
   \( n = 8,939 \)

6. Excluded – uncertain diabetes status due to insufficient information
   \( n = 14,092 \)

7. Metformin
   \( n = 3,029 \)

8. Other Oral DM2 drugs
   \( n = 1,629 \)

9. Insulin monotherapy
   \( n = 1,462 \)

10. Excluded – ineligible
    \( n = 2,819 \)

11. Non-Diabetics
    \( n = 73,138 \)
## Data Set Statistics - Vanderbilt

<table>
<thead>
<tr>
<th></th>
<th>Type 2 Diabetics (T2D)</th>
<th>Non- T2D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Metformin</td>
<td>Other oral meds</td>
</tr>
<tr>
<td># of patients</td>
<td>2218</td>
<td>903</td>
</tr>
<tr>
<td>Age</td>
<td>62</td>
<td>64</td>
</tr>
<tr>
<td>Female (%)</td>
<td>42%</td>
<td>39%</td>
</tr>
<tr>
<td>Race (W/B)</td>
<td>88%/12%</td>
<td>90%/10%</td>
</tr>
<tr>
<td>BMI</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>A1C</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td>% of died</td>
<td>30% (658)</td>
<td>49% (442)</td>
</tr>
</tbody>
</table>

.....
Survival Analysis – all cancers

K–M Plot for Vanderbilt Overall Cancer N=32415

K–M Plot for Mayo Overall Cancer N=79258
Cox proportional hazards model – all and four individual cancers

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>Treatment 1</th>
<th>Hazard Ratio</th>
<th>Treatment 2</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Cancer</td>
<td>Metf vs Other</td>
<td>0.78(0.69, 0.88)</td>
<td>Metf vs Other</td>
<td>0.70(0.63, 0.77)</td>
</tr>
<tr>
<td></td>
<td>Metf vs Insulin</td>
<td>0.61(0.50, 0.73)</td>
<td>Metf vs Insulin</td>
<td>0.65(0.58, 0.73)</td>
</tr>
<tr>
<td></td>
<td>Metf vs None</td>
<td>0.77(0.71, 0.85)</td>
<td>Metf vs None</td>
<td>0.59(0.54, 0.65)</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>Metf vs Other</td>
<td>0.47(0.26, 0.86)</td>
<td>Metf vs Other</td>
<td>0.49(0.31, 0.77)</td>
</tr>
<tr>
<td></td>
<td>Metf vs Insulin</td>
<td>0.38(0.13, 1.05)</td>
<td>Metf vs Insulin</td>
<td>0.57(0.34, 0.95)</td>
</tr>
<tr>
<td></td>
<td>Metf vs None</td>
<td>0.77(0.49, 1.21)</td>
<td>Metf vs None</td>
<td>0.47(0.31, 0.71)</td>
</tr>
<tr>
<td>Colorectal Cancer</td>
<td>Metf vs Other</td>
<td>0.50(0.31, 0.81)</td>
<td>Metf vs Other</td>
<td>0.60(0.44, 0.83)</td>
</tr>
<tr>
<td></td>
<td>Metf vs Insulin</td>
<td>0.74(0.29, 1.90)</td>
<td>Metf vs Insulin</td>
<td>0.79(0.55, 1.13)</td>
</tr>
<tr>
<td></td>
<td>Metf vs None</td>
<td>0.57(0.41, 0.80)</td>
<td>Metf vs None</td>
<td>0.48(0.35, 0.64)</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>Metf vs Other</td>
<td>0.90(0.64, 1.28)</td>
<td>Metf vs Other</td>
<td>0.76(0.58, 0.99)</td>
</tr>
<tr>
<td></td>
<td>Metf vs Insulin</td>
<td>0.93(0.48, 1.80)</td>
<td>Metf vs Insulin</td>
<td>0.59(0.40, 0.85)</td>
</tr>
<tr>
<td></td>
<td>Metf vs None</td>
<td>0.80(0.63, 1.01)</td>
<td>Metf vs None</td>
<td>0.58(0.47, 0.73)</td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>Metf vs Other</td>
<td>0.51(0.26, 1.00)</td>
<td>Metf vs Other</td>
<td>0.55(0.40, 0.76)</td>
</tr>
<tr>
<td></td>
<td>Metf vs Insulin</td>
<td>0.41(0.05, 3.61)</td>
<td>Metf vs Insulin</td>
<td>0.39(0.25, 0.62)</td>
</tr>
<tr>
<td></td>
<td>Metf vs None</td>
<td>1.04(0.66, 1.67)</td>
<td>Metf vs None</td>
<td>0.69(0.52, 0.93)</td>
</tr>
<tr>
<td>Other than 4 main</td>
<td>Metf vs Other</td>
<td>0.75(0.64, 0.88)</td>
<td>Metf vs Other</td>
<td>0.67(0.59, 0.77)</td>
</tr>
<tr>
<td></td>
<td>Metf vs Insulin</td>
<td>0.59(0.47, 0.73)</td>
<td>Metf vs Insulin</td>
<td>0.53(0.46, 0.61)</td>
</tr>
<tr>
<td></td>
<td>Metf vs None</td>
<td>0.71(0.63, 0.79)</td>
<td>Metf vs None</td>
<td>0.57(0.51, 0.65)</td>
</tr>
</tbody>
</table>
Limitations

- Incomplete medication exposure information
- Imperfect phenotyping algorithms
- Did not adjust for cancer treatment regimens
- Limited sample size for individual cancers
Conclusion

• Large EHRs are valuable sources for drug repurposing studies
• Informatics is the key to speed up this type of research
• A new drug repurposing model for using EHRs and informatics
  – Rapid pilot studies for hypothesis generation
  – Quick replications of known signals
  – New knowledge discovery (with solid study design and data extraction methods)
Future work

• More drugs – screening hundreds of other drugs
• More data sources – EHRs, claims, disease registries ...
• More NLP tools for clinical phenotyping – customizable, high-performance, user-friendly
• More analysis methods – missing data, confounder identification, bias correction ...
Acknowledgement

Collaborators

Joshua C. Denny M.D., M.S.
Melinda C. Aldrich, Ph.D., M.P.H.
Qingxia Chen, Ph.D.
Hongfang Liu, Ph.D.
Neeraja B. Peterson, M.D.
Qi Dai, M.D.
Mia Levy, M.D., Ph.D.
Anushi Shah, M.S

Xue Han, M.P.H.
Xiaoyang Ruan, Ph.D.
Min Jiang, M.S.
Ying Li, M.S.
Jamii St. Julien, M.D., M.P.H.
Jeremy Warner M.D., M.S.
Carol Friedman, Ph.D.
Dan M. Roden M.D.

Funding

CPRIT R1307
NCI R01CA141307
NHGRI U01 HG004603 - VGER
NIH RC2GM092618 - VESPA
NIH U19HL065962 – PGRN
Vanderbilt CTSA
Thank you!

Questions?

hua.xu@uth.tmc.edu
josh.denny@vanderbilt.edu