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

Reichert JM. Trends in development and approval times for new therapeutics in the US. *Nature reviews. Drug discovery.* 2003;2(9):695-702.

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
  - Structured-based virtual screening (Ma DL et al, *Chem Soc Rev,* 2013)
  - Analysis of side effect profiles (Campillos M, *Science*, 2008; Lounkine E et al, *Nature*, 2012)
  - Genomic and gene expression data (Wang and Zhang, *Nat Biotechnol*, 2013)
  - Biomedical literature mining (Andronis C et al, *Brief Bioinfom*, 2011)
- 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

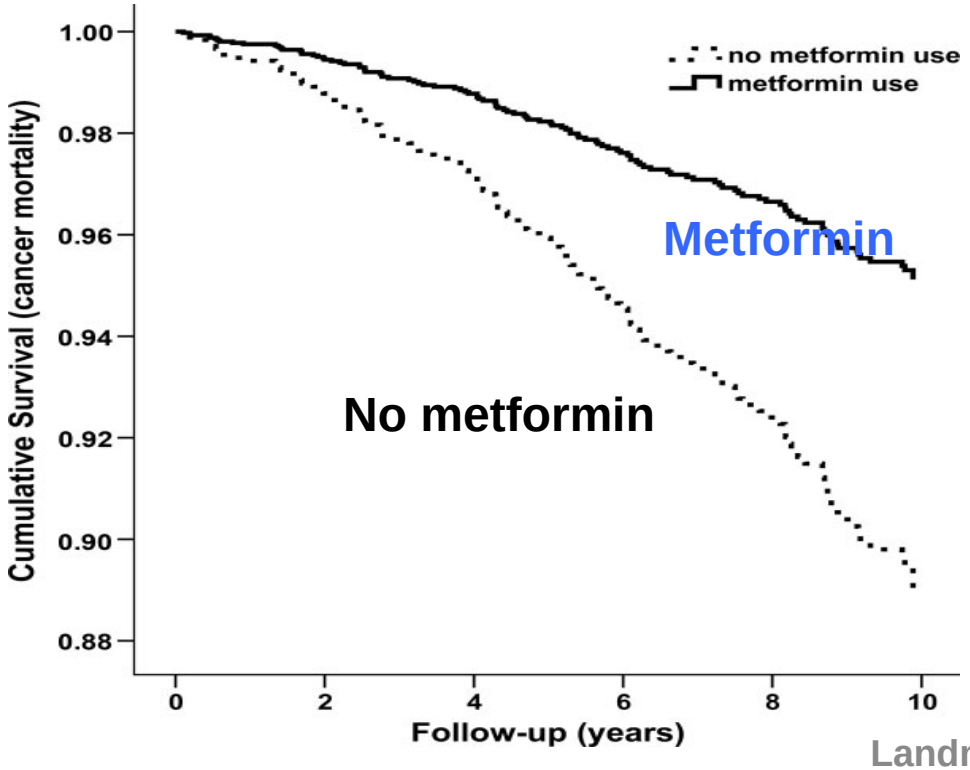
- Data extraction
  - Much of detailed information is embedded in narrative text
  - Heterogeneous data sources, e.g., different terminologies
- Data abstraction and analysis
  - Discrepancy
  - Missing data
  - Confounding
- The need for informatics approaches

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

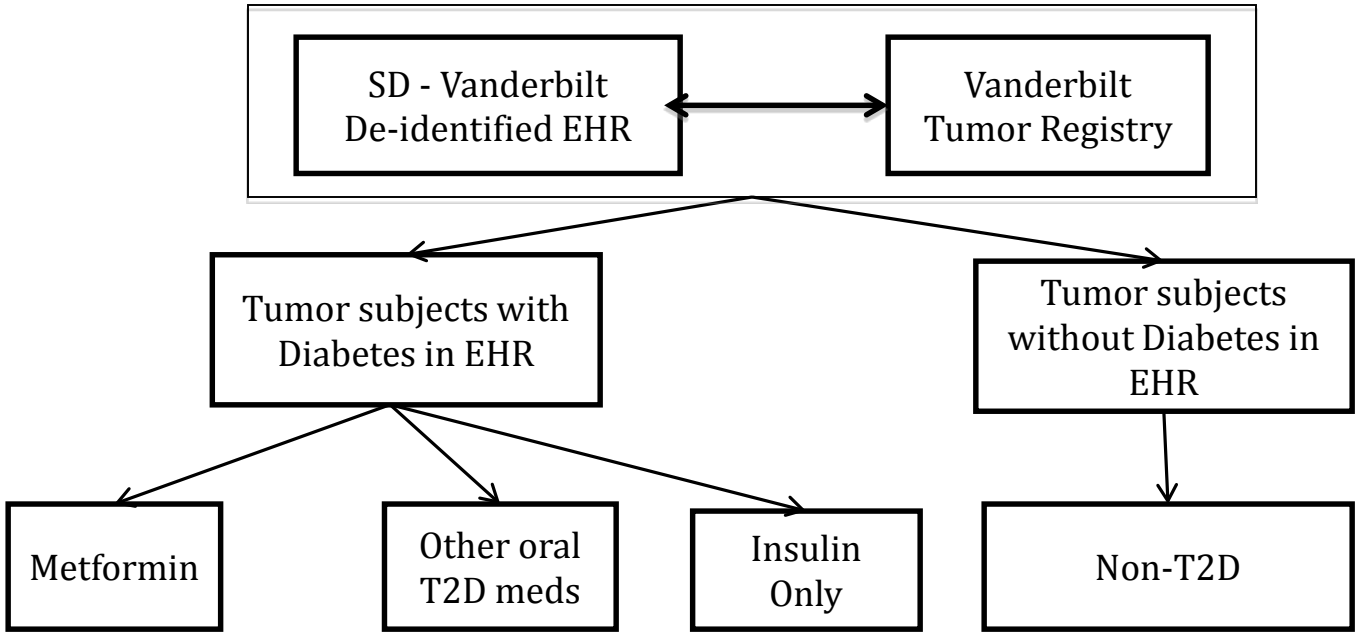
- Reduced Cancer Mortality with Metformin Use Among Diabetics



Landman Diab Care 2010

Study Design

- Primary analysis – Vanderbilt



- 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

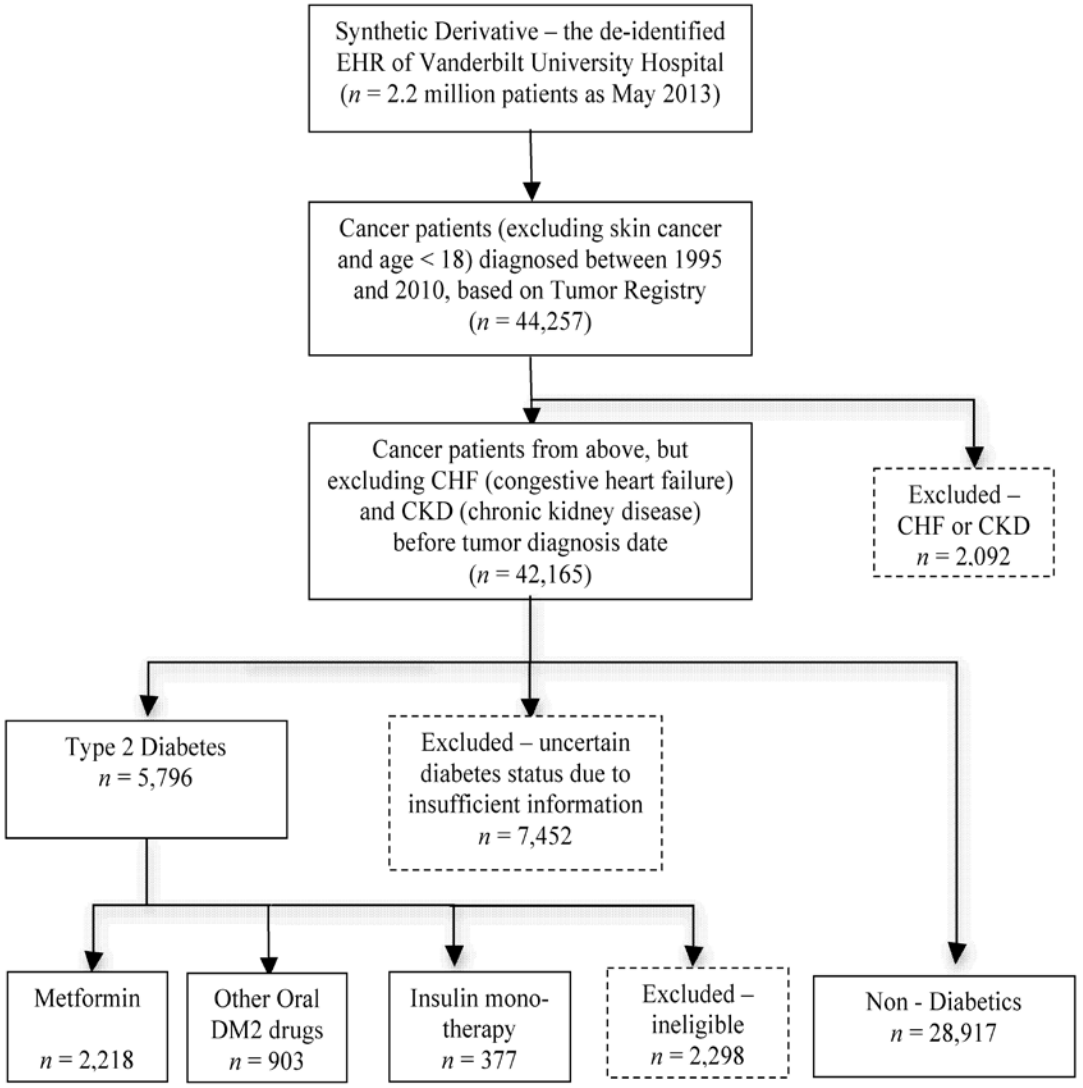
- MedEx
  - Identify drug name and signature information
  - Identify other T2D drugs
  - Determine metformin and its daily dose
  - Available at http://code.google.com/p/medex-uima/

- Heuristic rules for drug exposure

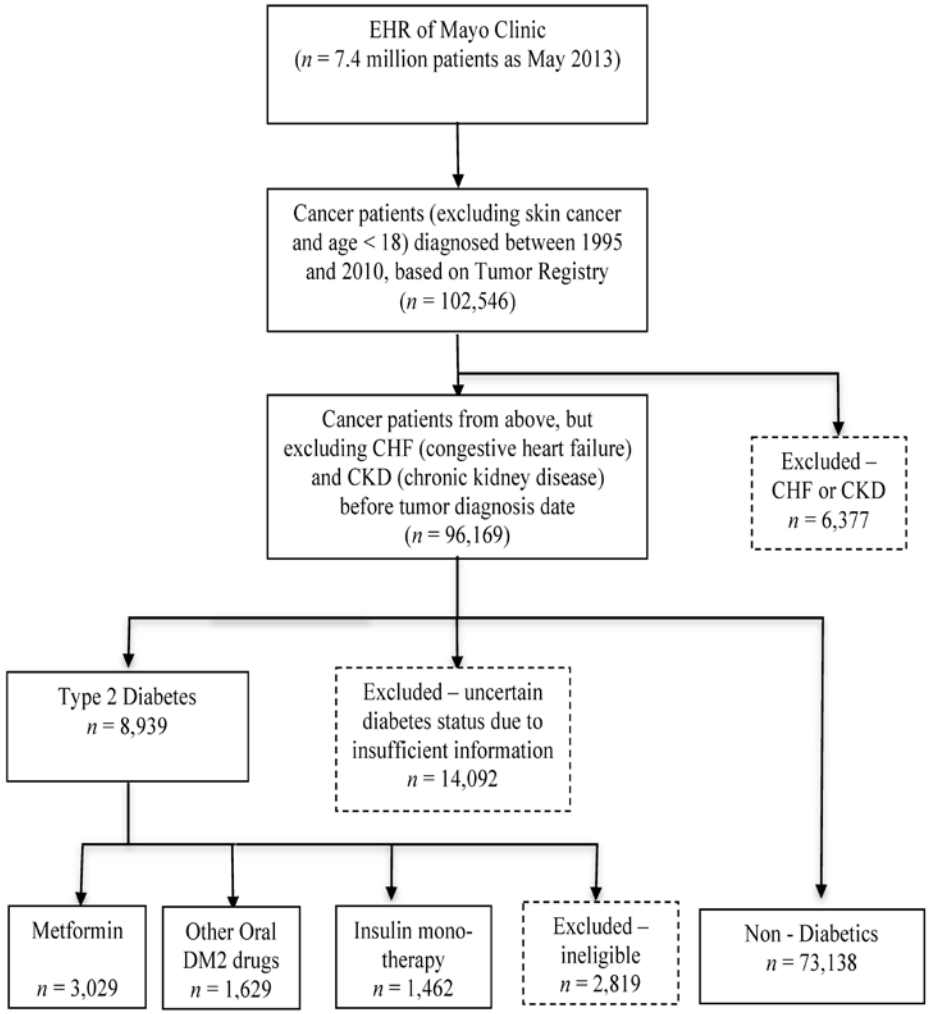
Informatics – covariate extraction

- 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



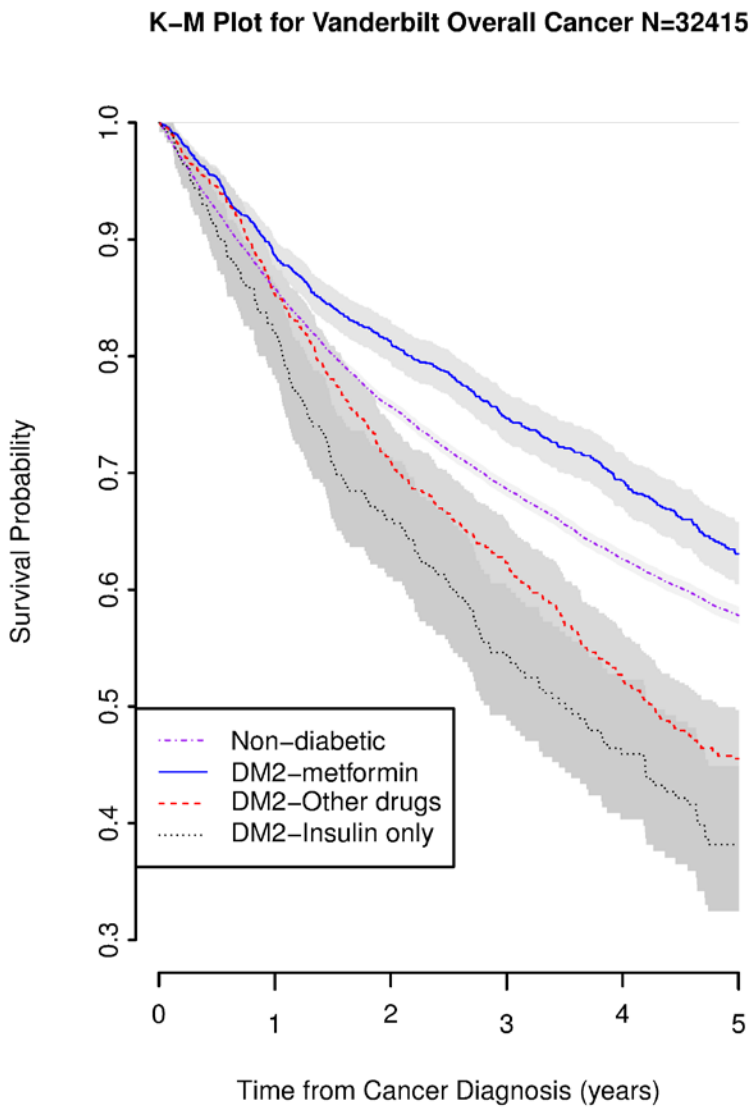
Data Extraction Workflow - Mayo

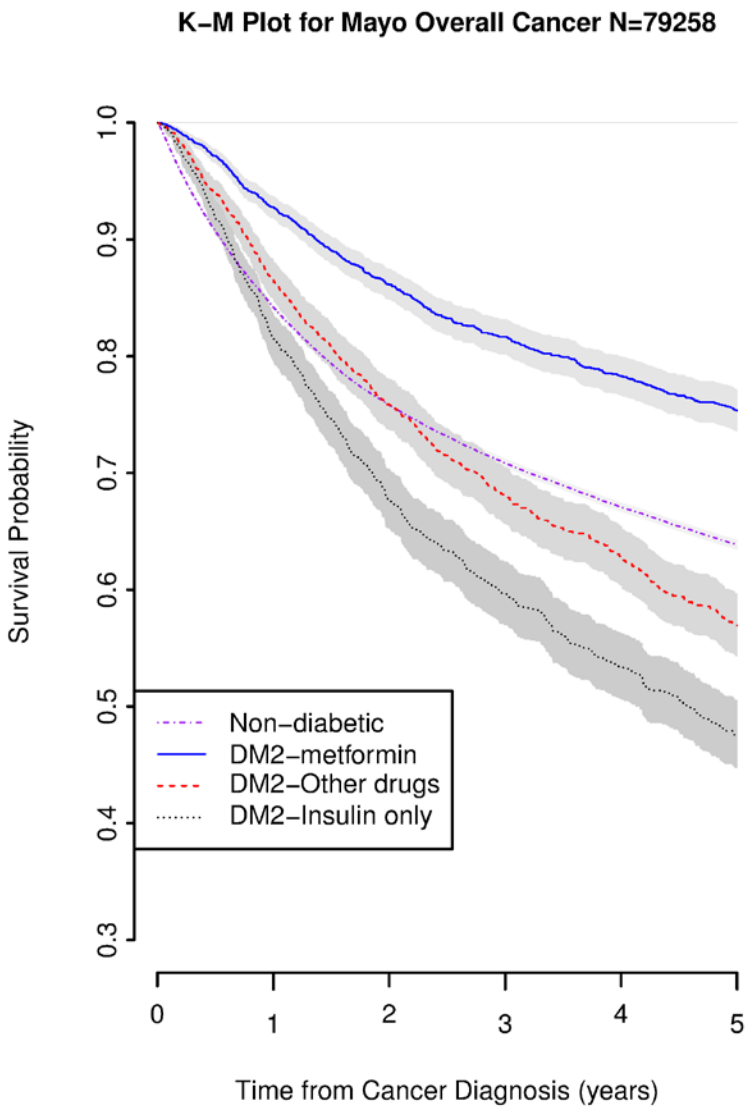


Data Set Statistics - Vanderbilt

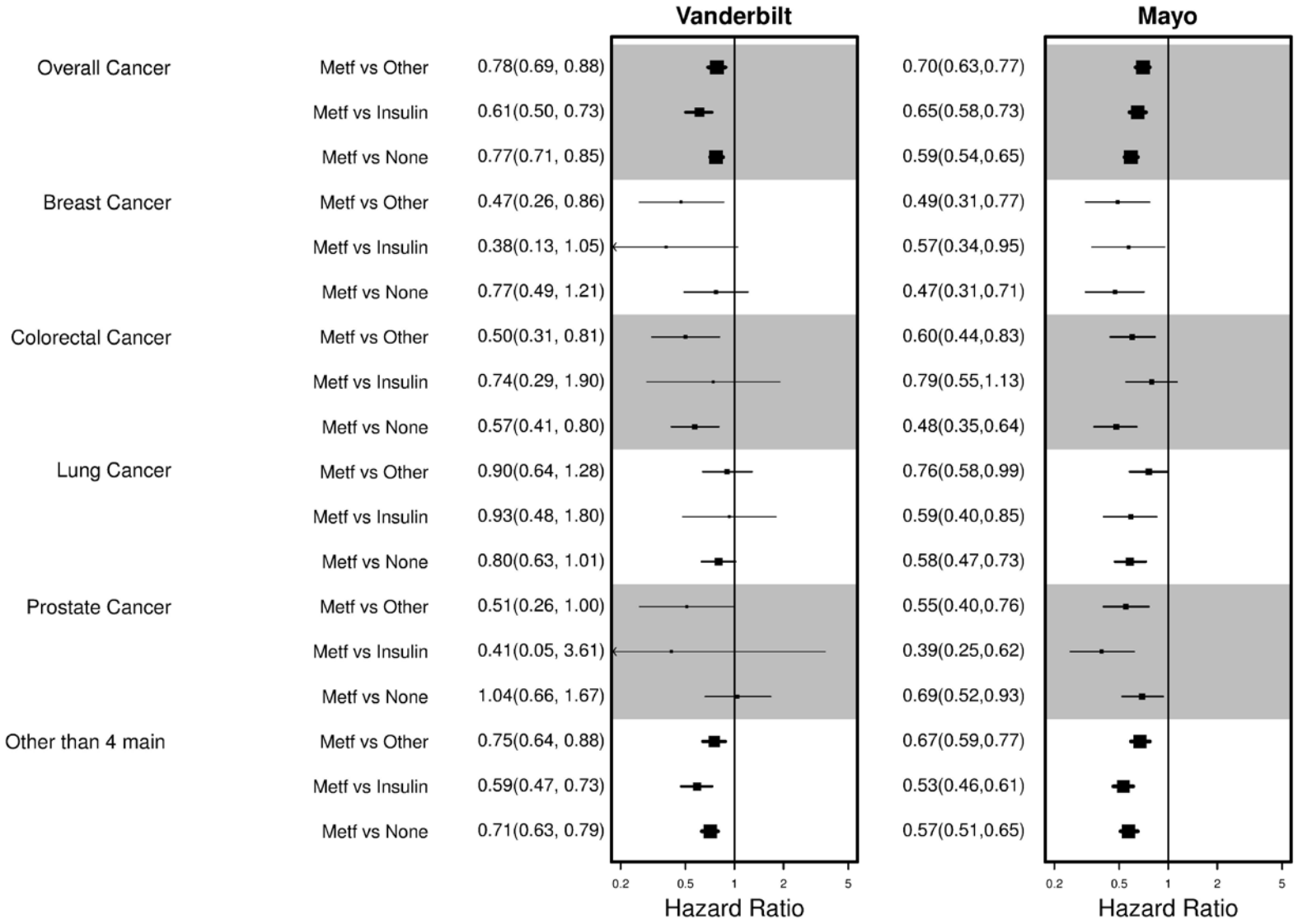
	Type 2 Diabetics (T2D)			Non- T2D
	Metformin	Other oral meds	Insulin	
# of patients	2218	903	337	28917
Age	62	64	59	58
Female (%)	42%	39%	42%	43%
Race (W/B)	88% / 12%	90% / 10%	88% / 12%	93% / 7%
BMI	31	31	29	27
A1C	7.5	7.5	7.6	N/A
…..				
% of died	30% (658)	49%(442)	59%(406)	33%(9449)

Survival Analysis – all cancers





Cox proportional hazards model – all and four individual cancers



ool of Biomedical
rmatics

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