Biomedical Informatics

discovery and impact

Facilitating phenotype transfer using the OMOP common data model in eMERGE

George Hripcsak, MD, MS

Biomedical Informatics, Columbia University

Medical Informatics Services, NewYork-Presbyterian

Columbia University Medical Center

NewYork-Presbyterian

Facilitating phenotype transfer using a common data model

Hripcsak G, Shang N, Peissig PL, Rasmussen LV, Liu C, Benoit B, Carroll RJ, Carrell DS, Denny JC, Dikilitas O, Gainer VS, Howell KM, Klann JG, Kullo IJ, Lingren T, Mentch FD, Murphy SN, Natarajan K, Pacheco JA, Wei WQ, Wiley K, Weng C

Journal of Biomedical Informatics,
accepted for publication

eMERGE Network

- Electronic medical records and genomics (eMERGE) Network
  - Funded by NIH's National Human Genome Research Institute (NHGRI)
- Combine DNA biorepositories with electronic health record systems for large scale, high-throughput genetic research in support of implementing genomic medicine
- 10 sites, 12 years, 136K patients, 64 phenotypes
  - PheKB.org repository

eMERGE Phenotype

- Generally a knowledge-engineered, rule-based definition of a disease or condition.
- Each site has its own local data model, terms
- Aim for high positive predictive value (PPV)
  - Precision
  - Genome-wide association studies require precision
- Primary site creates the definition and generally aims for >90% PPV
  - Secondary site implements and tests PPV
  - Rest of the network implements

Phenotype

- Can take months to create a new phenotype
- Comes with
  - Narrative description
  - Lists of terms (mostly ICD9), drug names
  - Graphical flow chart
  - Sometimes pseudocode
- Generally takes months to then implement it across the network
  - Effort is 2-3 weeks per site
- Much eMERGE research aims to improve phenotype development and sharing
  - Repeatable patterns, tools, specification language
  - Machine learning

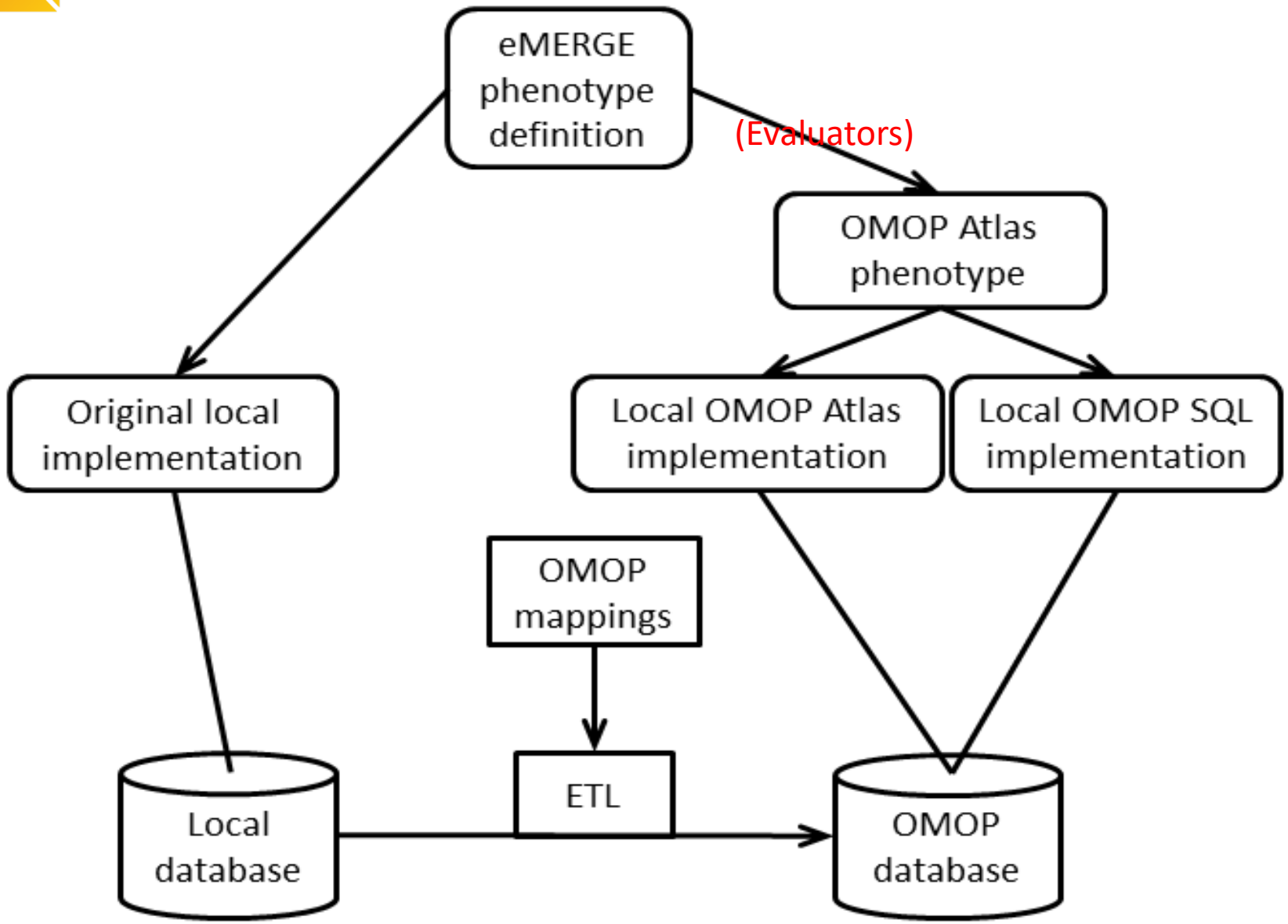
Study Design

- NHGRI eMERGE OMOP supplement 2016
- Site converts local database to OMOP
- Select phenotypes (structured data only)
  - Type 2 diabetes mellitus (T2DM)
    - Complex with many data types
  - Attention deficit and hyperactivity disorder (ADHD)
    - Simpler
- Evaluators convert eMERGE phenotype to OMOP (Atlas)
  - Generate Atlas JSON and SQL

Study design

- Share the new phenotype
  - Each site implements and runs it
- Ask each site
  - Time and effort to complete
  - Compare to original eMERGE phenotype
  - Record issues: coding, data, query, DBMS, software stack, organizational, other

Study design



Results: Database conversion

- All 10 sites converted database to OMOP
  - 4 to 12 months elapsed time
  - 2 sites report still converting lab and procedure
  - Lab data in local codes, so many did not convert
    - Instead map labs as needed
  - 5 sites installed the stack with Atlas
    - Reasons for not: security, DBMS, effort

Results: phenotyping

- 9 sites did phenotyping exercise
  - 7/9 T2DM and 6/8 ADHD ran phenotype in 1 day
  - Rest took 14 to 144 days elapsed time
    - Other priorities or had to reload data
- Prevalence of condition varied
  - 0.3%-22.4% T2DM
  - 0.1%-12.3% ADHD
  - Age groups, disease cohorts

Results: phenotype

- 5 sites compared OMOP to old phenotype
  - Reasons for not: joined after phenotype was shared, low expected case count, lost original results, change in privacy policy
- Agreement varied 100% to 43%

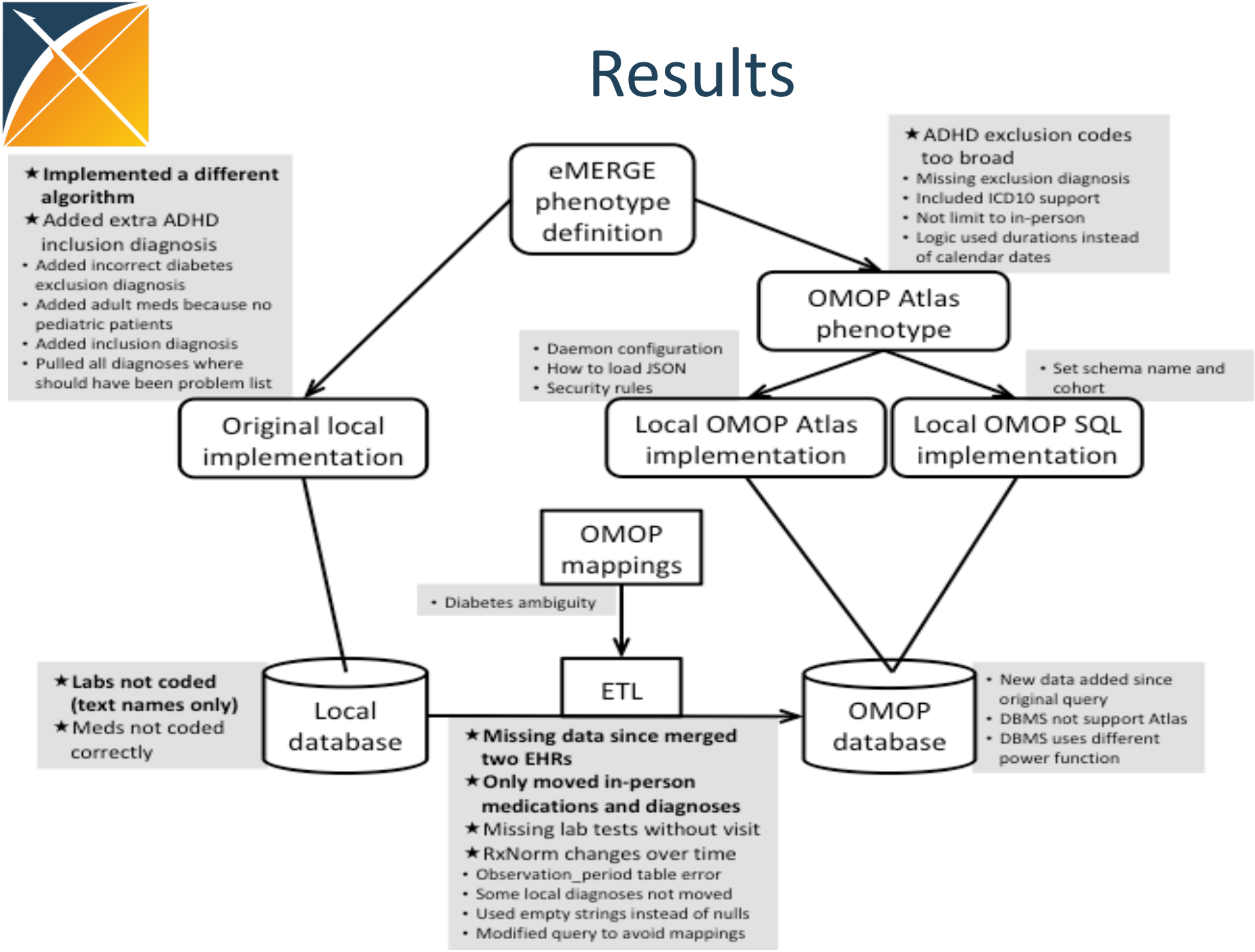
Results: T2DM

Overlap between original and OMOP phenotypes (number)				Positive specific agree-ment	Negative specific agree-ment
Overlap	Original only	OMOP only	Neither		
38	0	4	5465	**0.950**	**1.000**
1179	95	30	4086	**0.950**	**0.985**
242	381	250	4804	**0.434**	**0.938**
735	1165	18	396	**0.554**	**0.401**
3139	819	1588	19143	**0.723**	**0.941**

Results: ADHD

Overlap between original and OMOP phenotypes (number)				Positive specific agree-ment	Negative specific agree-ment
Overlap	Original only	OMOP only	Neither		
7	0	0	5500	**1.000**	**1.000**
23	11	1	5355	**0.793**	**0.999**
1761	507	48	12282	**0.864**	**0.978**
65	15	19	4861	**0.793**	**0.997**

Results


★Implemented a different algorithm
★Added extra ADHD inclusion diagnosis
• Added incorrect diabetes exclusion diagnosis
• Added adult meds because no pediatric patients
• Added inclusion diagnosis
• Pulled all diagnoses where should have been problem list
eMERGE phenotype definition
★ADHD exclusion codes too broad
• Missing exclusion diagnosis
• Included ICD10 support
• Not limit to in-person
• Logic used durations instead of calendar dates
OMOP Atlas phenotype
• Daemon configuration
• How to load JSON
• Security rules
• Set schema name and cohort
Original local implementation
Local OMOP Atlas implementation
Local OMOP SQL implementation
OMOP mappings
• Diabetes ambiguity
★Labs not coded (text names only)
★Meds not coded correctly
Local database
ETL
OMOP database
• New data added since original query
• DBMS not support Atlas
• DBMS uses different power function
★Missing data since merged two EHRs
★Only moved in-person medications and diagnoses
★Missing lab tests without visit
★RxNorm changes over time
• Observation_period table error
• Some local diagnoses not moved
• Used empty strings instead of nulls
• Modified query to avoid mappings

Results: local data

- ★**Labs not coded (text names only)**
- ★Meds not coded correctly

*__Bold__ >2%

*Plain 0.2-2%

. Plain <0.2%

Results: local ETL

- ★**Missing data since merged two EHRs**
- ★**Only moved inpatient diagnoses and meds**
- ★Missing lab tests without visit
- ★RxNorm changes over time

- Observation_period table error
- Some local diagnoses not moved
- Used empty strings instead of nulls
- Modified query to avoid mappings

Results: original implementation

★**Implemented a different algorithm**

★Used only inpatient diagnoses for inclusion

- Added incorrect exclusion diagnosis
- Added inclusion diagnosis not included in definition
- Added adult meds because no pediatric patients
- Pulled all diagnoses where should have been problem list
- Skipped some encounters

Results: Altas implementation

★ADHD exclusion codes too broad

- Erroneously missing one ADHD inclusion diagnosis
- Missing exclusion diagnosis
- Optimized to include ICD10 instead of just ICD9
- Logic used durations instead of calendar dates

Results: local Atlas implementation

- Daemon configuration
- How to load JSON
- Security rules

Results: local SQL implementation

- Set schema name and cohort

Results: OMOP mappings

- Diabetes ambiguity

Results: local OMOP database

- New data added since original query
- DBMS not support Atlas
- DBMS uses different power function

Findings

- Sharing of a single computable query uncovered differences among the original implementations despite starting from the same **narrative description, codes lists, pseudocode, and flowchart**
  - Sharing is hard

Findings

- The eMERGE network was able to convert its databases into the OHDSI OMOP Common Data Model
  - Primary challenge conversion of local laboratory test codes to the LOINC standard
  - ICD* and drugs straightforward

Findings

- Efficiency of sharing phenotypes improved dramatically with most sites able to execute the query within a day
- Is it worth it?
  - Cost of converting database to OMOP (4 months)
  - Savings in implementing phenotype (2 weeks)
  - Breakeven point about 10 to 20 phenotypes

Findings

- Agreement between the OMOP phenotype query and the original eMERGE query varied from perfect to mediocre
  - Problems in the original query
  - Problems in the OMOP query
  - Changes in data
  - Issues in the database
  - (More about data and database than logic)

Limitations

- Only 2 phenotypes
- Half sites could not compare to original
- Only structured data

Conclusion

- Implementing original phenotypes over a network of electronic health record databases had been labor intensive and error prone
- The potential for a common data model to improve efficiency and consistency

Thanks

- eMERGE Network
- Co-authors
- NHGRI
  - This phase of the eMERGE Network was initiated and funded by the NHGRI through the following grants: U01HG008657 (Group Health Cooperative/University of Washington); U01HG008685 (Brigham and Women's Hospital); U01HG008672 (Vanderbilt University Medical Center); U01HG008666 (Cincinnati Children's Hospital Medical Center); U01HG006379 (Mayo Clinic); U01HG008679 (Geisinger Clinic); U01HG008680 (Columbia University Health Sciences); U01HG008684 (Children's Hospital of Philadelphia); U01HG008673 (Northwestern University); U01HG008701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG008676 (Partners Healthcare/Broad Institute); U01HG008664 (Baylor College of Medicine); and U54MD007593 (Meharry Medical College).
  - In addition, this work was funded by R01LM006910, Discovering and applying knowledge in clinical databases; R01HG009174, Developing i2b2 into a Health Innovation Platform for Clinical Decision Support in the Genomics Era, OT2OD026553, The New England Precision Medicine Consortium of the All of Us Research Program. Vanderbilt University Medical Center's BioVU is supported by numerous sources: institutional funding, private agencies, and federal grants, including the NIH funded Shared Instrumentation Grant S10RR025141; and CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975.