Lecture 22: PheWAS
Phenome-wide association studies
<table>
<thead>
<tr>
<th>Project</th>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
<th>Lec Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>Fri, Sep 8</td>
<td>Introduction</td>
<td>L1 Intro: Biology, Algorithms, Machine Learning, Course Overview</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Fri, Sep 16</td>
<td>Module I: Aligning and Modeling Genomes</td>
<td>L2 Alignment I: Dynamic Programming, Global and local alignment</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Thu, Sep 20</td>
<td>Frontiers</td>
<td>L4 Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Thu, Sep 27</td>
<td>Module II: Expression Analysis and Networks</td>
<td>L6 Expression Analysis: Clustering/Classification, K-means, Hierarchical Bayesian, L7 Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Thu, Oct 4</td>
<td>Frontiers</td>
<td>L8 Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Fri, Oct 7</td>
<td>Frontiers</td>
<td>L9 Three-dimensional chromatin interactions: 3C, 5C, HiC, ChiP-Pet</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Fri, Oct 7</td>
<td>Frontiers</td>
<td>L4 Reaction 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Fri, Oct 21</td>
<td>Project Planning: research areas, initial ideas, type of project, mentor matching, finding</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Thu, Nov 1</td>
<td>Frontiers</td>
<td>L10 Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Thu, Nov 3</td>
<td>Frontiers</td>
<td>L7 Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, IMWAS, L8 Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>Fri, Nov 18</td>
<td>Frontiers</td>
<td>L8 Reaction 8: Rare Variants, ExAC</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>Tue, Nov 22</td>
<td>Frontiers</td>
<td>L17 Comparative genomics and evolutionary signatures</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>Tue, Nov 29</td>
<td>Module VI: Current Research Directions</td>
<td>L21 Single-cell genomics: technology, analysis, microfluidics, applications, insights</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>Tue, Dec 13</td>
<td>Frontiers</td>
<td>L24 Genome Engineering with CRISPR/Cas9 and related technologies</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>Tue, Dec 13</td>
<td>Frontiers</td>
<td>L25 Final Presentations - Part 1 (11am), 32-G8 reading room</td>
</tr>
</tbody>
</table>

**Project profile due Tue 9/27**

**Identify previous project proposals, recent papers, and potential partners that match your areas of interest. List initial project ideas and partners.**

**Project team/ideas due 10/4**

**Form teams of two, specify project goals, division of work, milestones, datasets, challenges. Prepare slide presentation for the class and the mentors. Project proposal due 10/18. Presented to mentor on Fri 10/21.**

**Evaluate/discuss three peer proposals, NIH review format.**

**Reviews due Mon 10/31**

**Reviews returned Thu 11/3**

**Address peer evaluations, revise aims, scope, list of final deliverables / goals. Response due 11/10**

**Continue making substantial progress on proposed milestones. Write outline of final report. Midcourse report due Wed 11/23**

**Complete your milestones, finalize results, figures, write-up in conference publication format. As part of report, comment on your overall project experience. Written report due Sun 12/11**

**Conference format slide pres. Talks on Tue 12/13**

**Due Tue 9/27**

**No more pesets! (work on your final project)** (work on your final project)

**No more classes - student holiday**

**Project Intro: about the projects, self introductions, mentor intro, example projects, teamwork 32D-507**

**No Class - Columbus Day Holiday**

**Project Planning: research areas, initial ideas, type of project, mentor matching, finding 32D-507 at 4-5pm**

**No Recitation. Veterans Day**

**No more classes - thanksgiving break - Thu Nov 25, 2015**

**No recitation, thanksgiving break**

**Quiz in-class quiz (the only quiz - the class has no final exam) - covers L1-L20. R1-R9**

**No more recitations - thanksgiving break**

**Conference format slide presentation. Talks on Tue 12/13**

**No More Classes - student holiday**
Module 6: Current research directions

- **L21: Single-cell genomics**
  - Measuring and analyzing biology at the single-cell level

- **L22: PheWAS (Phenome-wide associations studies)**
  - Multi-phenotype analyses, inferences, association, imputation

- **L23: Cancer genomics**
  - Mutational heterogeneity, tumor evolution, immune evasion

- **L24: Genome engineering & high-throughput biology**
  - From reading to writing, CRISPR-Cas9,
PheWAS: Multi-phenotype studies

1. Motivation of phenome-wide association studies
   - PheWAS-informed phenotyping, improved GWAS power, etc
   - Electronic health record (EHR) contain rich personalized information

2. Modeling PheWAS using genetic information
   - inferring pleiotropic variants/pathways
   - Summary-statistics based inference of disease correlation

3. PheWAS without genetic: Rationale and EHR data properties and challenges

   - Tensor decomposition to model multi-dimensional EHR data
   - Deep learning to impute missing data and predict diseases
   - Latent topic model to infer latent disease grouping

5. Model EHR with non-missing at mechanism
   - Concepts of missing mechanism
   - Models leveraging missing information and inferring missing mechanism

6. Methods on modeling longitudinal EHR data
1. Motivation of phenome-wide association studies (PheWAS)
Genetic Variant

Tissue/cell type

Molecular Phenotypes

Epigenetic Changes

Gene Expression Changes

Organismal phenotypes

Environment

Feedback from environment / disease state

Brain

Enhancer

H3K27ac

Gene

Gene expr.

Endo phenotypes

Lipids

Tension

Amyloid β

Metabol.

Drug resp

Disease

Heart

Brain

Cortex

Lung

Blood

Skin

Nerve

Promoter

Insulator

Gene expr.

Gene expr.
Why do PheWAS?

- Draw associations of unknown or underappreciated phenotypes with known phenotypes;
- Identifying disease-causing mechanisms via mediating or easy-to-measure ‘biomarker’ phenotypes
- Improve GWAS power;
- Associated with pleiotropic phenotypes;
- Guide further phenotyping and personalized medicine
EHR contains extremely rich information of a patient:

- **Phenotype**: International Classification of Disease-9 (ICD-9) diagnosis code
- **Lab tests**: Laboratory data LOINC (Logical Observation Identifiers Names and Codes)
- **Pharmaceutical**: Prescription data (RxNorm)
- **Imaging**: Digital Imaging and Communication in Medicine (DICOM) for imaging files
- **Procedures**: ICD-9 procedure code
Directed phenotyping based on initial results

- Raw EHR data are *indirect reflection* of true patient state (e.g. recording errors)
- State of the patient varies, determines both the *output* of the measurement, but also whether the measurement will be carried out at all (*presence*)

Hripcsak & Albers (2012)
PheWAS: Multi-phenotype studies

1. Motivation of phenome-wide association studies
   – PheWAS-informed phenotyping, improved GWAS power, etc
   – Electronic health record (EHR) contain rich personalized information

2. Modeling PheWAS using genetic information
   – inferring pleiotropic variants/pathways
   – Summary-statistics based inference of disease correlation

3. PheWAS without genetic: Rationale and EHR data properties and challenges

   – Tensor decomposition to model multi-dimensional EHR data
   – Deep learning to impute missing data and predict diseases
   – Latent topic model to infer latent disease grouping

5. Model EHR with non-missing at mechanism
   – Concepts of missing mechanism
   – Models leveraging missing information and inferring missing mechanism

6. Methods on modeling longitudinal EHR data
2. Modeling PheWAS using genetic information: inferring pleiotropic variants and pathways implicated in multiple phenotypes

a. Shared risk variants
b. Summary statistics-based methods
c. Shared heritability. Using polygenic risk prediction.
d. Multiple phenotype sharing at the pathway level
Genome-wide PheWAS view

- PheWAS confirms GWAS associations
- PheWAS discovers new associations

Denny et al., (NatBiotech, 2013)
PheWAS Manhattan plots

- Each panel represents 1,358 phenotypes tested for association with a particular SNP.
- Same variant are associated with multiple traits surveyed through National Health and Nutrition Examination Surveys (NHANES).
- Some of traits are known to be highly related (e.g., Type 1 diabetes) implying potential new discoveries of phenotypic associations.

Denny et al., (NatBiotech, 2013)
PheWAS scan using ICD9 code

rs3135388

• Similarly, associating SNP with ICD-9 code reveals modestly co-associated traits with the target disease (i.e., multiple sclerosis).

Nominal P < 0.05

Bonferroni-adj. P < 0.05

Replicated in prior study

Denny et al., (Bioinf 2010)
PheWAS using NHANES data

One “stone” (SNP) many “birds” (traits)

*APoE/APoC1/C1P1/C2/C4, rs4420638, Coded Allele A*

Length of the lines correspond to -logP

rs4420638 near APOC1 associated with LDL and several known and potentially new metabolic-related traits

Pendergrass et al., (PlosGen, 2013)
Leverage multi-trait GWAS co-association

• Modelling multiple traits with multivariate linear mixed model:

\[ Y = BX + G + E \]

\[ G \sim MN(0, V_g, D_k) \]

\[ E \sim MN(0, V_e, I_{n \times n}) \]

• \(Y\): \(D \times N\) matrix of \(N\) individuals and \(D\) traits
• \(B\): \(D \times M\) effect size matrix for \(M\) SNPs
• \(X\): \(M \times N\) genotype matrix
• \(G\): \(D \times N\) random effect matrix and follows matrix normal (MN)
• \(V_g\): \(D \times D\) symmetric matrix of genetic variance component
• \(V_e\): \(D \times D\) symmetric matrix of environmental variance component
• \(D_k\): \(N \times N\) diagonal matrix filled with eigen values of the kinship matrix

Xiang & Stephens, (NatMeth, 2014)
Power gain in modeling multiple phenotypes

- Simulation results illustrate the potential gain in power from four phenotypes versus two-phenotype analyses
- Interestingly, the 4-phenotype model consistently outperforms the 2-phenotype model even when the underlying number of affected phenotypes is lower than 4
- It is expected if the unrelated phenotypes are correlated with the associated phenotypes

Xiang & Stephens, (NatMeth, 2014)
Estimate genetic correlation with LD score regression

\[ E[z_{1j} z_{2j} | l_j] = \frac{\sqrt{N_1 N_2} \rho_g}{M} l_j + \frac{\rho N_S}{\sqrt{N_1 N_2}} \]

\[ l_j = \sum_k r_{jk}^2 \text{ is the LD score of SNP } j \]

- \( N_i \) is the sample size for study \( i \)
- \( \rho_g \) is the genetic covariance, which can be efficiently estimated by fitting linear regression
- \( l_j = \sum_k r_{jk}^2 \text{ is the LD score of SNP } j \)
- \( N_S \) is the number of individuals included in both studies
- \( \rho \) is the phenoetypic correlation among the \( N_S \) overlapping samples

Sullivan et al., (NatGen, 2015)
LDSC-estimated genetic correlation

- Genetic correlation among 24 traits
- Full-squares are sig. at FDR < 0.01
- Known related traits are indeed genetically correlated
- This is remarkable as it suggests that traits are correlated through genome-wide signals beyond GWAS significant SNPs

Sullivan et al., (NatGen, 2015)
Detecting shared genetic influences by regional Bayes factor

Three models test for the same SNP associated with:
1. trait 1 only
2. trait 2 only
3. Both trait 1 and trait 2

Model 4 tests for the same locus but different SNPs

Likelihood in terms of Regional Bayes Factor (RBF)

$$l(\Theta|D) = \sum_{i=1}^{M} \ln \left( \Pi_0 + \sum_{j=1}^{4} \pi_j \text{RBF}_i^{(j)} \right)$$

Pickrell et al., (NatGen 2016)
Proportion of shared risk variants among 42 traits

- Maximum a posteriori (MAP) estimate of the proportion of SNPs shared between any two traits
- Note this is NOT symmetrical due to different # associated SNPs per trait
- Similar to LDSC, some related traits are genetically related at the GWAS loci/regions (besides being related genome-wide as tested in LDSC)

Pickrell et al., (NatGen 2016)
Inferring causal relationships between traits

Correlation between traits based on variants associated with one trait implies causal direction from one trait to the other

Pickrell et al., (NatGen 2016)
Inferring causal relationships between traits

Correlation between traits based on variants associated with one trait implies causal direction from one trait to the other

Pickrell et al., (NatGen 2016)
Probabilistically inferring pleiotropic variants

\[ \pi_{00} = \Pr(Z_{j00} = 1): \quad (P_{j1}|Z_{j00} = 1) \sim U[0,1], \]
\[ (P_{j2}|Z_{j00} = 1) \sim U[0,1], \]

\[ \pi_{10} = \Pr(Z_{j10} = 1): \quad (P_{j1}|Z_{j10} = 1) \sim \text{Beta}(\alpha_1, 1), \]
\[ (P_{j2}|Z_{j10} = 1) \sim U[0,1], \]

\[ \pi_{01} = \Pr(Z_{j01} = 1): \quad (P_{j1}|Z_{j01} = 1) \sim U[0,1], \]
\[ (P_{j2}|Z_{j01} = 1) \sim \text{Beta}(\alpha_2, 1), \]

\[ \pi_{11} = \Pr(Z_{j11} = 1): \quad (P_{j1}|Z_{j11} = 1) \sim \text{Beta}(\alpha_1, 1), \]
\[ (P_{j2}|Z_{j11} = 1) \sim \text{Beta}(\alpha_2, 1), \]

• Need to enumerate all possibilities for every SNP of any given D phenotypes
• An more efficient method is to leverage the epignomic covariance between traits to infer disease co-association (next)

SNP is associated with
1. Neither trait
2. Trait 1 only
3. Trait 2 only
4. Both traits

Different from regional Bayes factor (RBF) method that requires estimates of effect size, GPA (genotype-pleiotropy-association) operates on \( p \)-values, which follow uniform \( U[0,1] \) if no association or \( \text{Beta}[a,1] \) if association

Chung et al., (PlosGen 2014)
Epigenomes of 127 cell types
Multiple GWAS traits exhibit tissue-specific co-enrichments

**Neurologic/Psychiatric traits**

- **Neuroticism**
  - GWAS
  - Prior
  - Posterior

- **Year of Education**
  - GWAS
  - Prior
  - Posterior

- **Schizophrenia**
  - GWAS
  - Prior
  - Posterior

**Immune-related diseases**

- **Crohn Disease**
  - GWAS
  - Prior
  - Posterior

- **Multiple sclerosis**
  - GWAS
  - Prior
  - Posterior

- **Type 1 Diabetes**
  - GWAS
  - Prior
  - Posterior

H3K27ac enrichments

- in immune cells
- in CNS cells

**Genes and Regions**

- chr1: 77605336-77605371
- chr17: 42719303-44122773
- chr17: 42719303-44122773
- chr10: 80110565-8130001

**Top SNPs**
Inferring disease epigenomic-covariance

Li & Kellis (NAR 2016; bioRxiv)
Multi-trait inference improve functional enrichment

(Li & Kellis, NAR 2016)

- Celiac Disease
- Juvenile Idiopathic Arthritis
- Multiple Sclerosis
- Narcolepsy
- Primary Biliary Cirrhosis
- Rheumatoid Arthritis
- Type 1 Diabetes

eQTL enrichment score

- Single-trait mode
- Multi-trait mode
Common polygenic risk variants for schizophrenia and bipolar disorder

- Variants detected in schizophrenia explain substantial phenotypic variance in independent cohorts of patients with not only SCZ but also bipolar
- Indeed, the same set of variants explain little phenotypic variance of unrelated traits (i.e., CAD, CD, HT, etc)

The International Schizophrenia Consortium (Nat, 2009)
Pathway-level co-enrichments

1. Collect pathway information
2. Obtain risk in-cis genes based on GWAS SNPs
3. Pathway enrichment using different methods for the risk genes
4. Rank pathways based on score enrichment scores
5. Test for significance of pathways based on permutation tests
6. Combining p-values via Fisher/Brown’s methods for the same pathway across 3 psychiatric traits (BIP, SCZ, Major depression disorder (MDD)).

Psychiatric Genomics Consortium (NatNeu, 2014)
Combining enrichment improves causal pathway detection

### Table 2: Top results from integrative pathway analysis of three adult disorders

<table>
<thead>
<tr>
<th>Rank</th>
<th>BIP</th>
<th>MDD</th>
<th>SCZ</th>
<th>Combined P</th>
<th>q-value</th>
<th>Pathway ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0000</td>
<td>0.0592</td>
<td>0.0001</td>
<td>$5.75 \times 10^{-8}$</td>
<td>0.0003</td>
<td>GO:51568</td>
<td>Histone H3-K4 methylation</td>
</tr>
<tr>
<td>2</td>
<td>0.0004</td>
<td>0.0500</td>
<td>0.0006</td>
<td>$1.46 \times 10^{-5}$</td>
<td>0.0362</td>
<td>GO:16571</td>
<td>Histone methylation</td>
</tr>
<tr>
<td>3</td>
<td>0.0004</td>
<td>0.1462</td>
<td>0.0011</td>
<td>$4.73 \times 10^{-5}$</td>
<td>0.0414</td>
<td>GO:43414</td>
<td>Macromolecule methylation</td>
</tr>
<tr>
<td>4</td>
<td>0.0008</td>
<td>0.0630</td>
<td>0.0014</td>
<td>$5.10 \times 10^{-5}$</td>
<td>0.0414</td>
<td>GO:34968</td>
<td>Histone lysine methylation</td>
</tr>
<tr>
<td>5</td>
<td>0.4200</td>
<td>0.0001</td>
<td>0.0023</td>
<td>$5.58 \times 10^{-5}$</td>
<td>0.0414</td>
<td>GO:45216</td>
<td>Cell-cell junction organization</td>
</tr>
<tr>
<td>6</td>
<td>0.0001</td>
<td>0.0910</td>
<td>0.0064</td>
<td>$5.69 \times 10^{-5}$</td>
<td>0.0414</td>
<td>P00003</td>
<td>Alzheimer disease–amyloid secretase pathway</td>
</tr>
<tr>
<td>7</td>
<td>0.0007</td>
<td>0.0495</td>
<td>0.0024</td>
<td>$5.86 \times 10^{-5}$</td>
<td>0.0414</td>
<td>P04393</td>
<td>Ras pathway</td>
</tr>
<tr>
<td>8</td>
<td>0.3120</td>
<td>0.0000</td>
<td>0.1286</td>
<td>$7.12 \times 10^{-5}$</td>
<td>0.0422</td>
<td>GO:8601</td>
<td>Protein phosphatase type 2A regulator activity</td>
</tr>
<tr>
<td>9</td>
<td>0.8980</td>
<td>0.0001</td>
<td>0.0017</td>
<td>$7.83 \times 10^{-5}$</td>
<td>0.0422</td>
<td>GO:43297</td>
<td>Apical junction assembly</td>
</tr>
<tr>
<td>10</td>
<td>0.0013</td>
<td>0.0207</td>
<td>0.0055</td>
<td>$9.25 \times 10^{-5}$</td>
<td>0.0422</td>
<td>P00052</td>
<td>TGF-β signaling pathway</td>
</tr>
<tr>
<td>11</td>
<td>0.4890</td>
<td>0.0203</td>
<td>0.0000</td>
<td>$9.53 \times 10^{-5}$</td>
<td>0.0422</td>
<td>GO:14069</td>
<td>Postsynaptic density</td>
</tr>
<tr>
<td>12</td>
<td>0.0085</td>
<td>0.0009</td>
<td>0.0239</td>
<td>0.0001</td>
<td>0.0422</td>
<td>GO:32869</td>
<td>Cellular response to insulin stimulus</td>
</tr>
<tr>
<td>13</td>
<td>0.0188</td>
<td>0.0054</td>
<td>0.0022</td>
<td>0.0001</td>
<td>0.0450</td>
<td>P00010</td>
<td>B cell activation</td>
</tr>
<tr>
<td>14</td>
<td>0.0023</td>
<td>0.2988</td>
<td>0.0003</td>
<td>0.0001</td>
<td>0.0450</td>
<td>GO:8757</td>
<td>S-adenosylmethionine–dependent methyltransferase activity</td>
</tr>
<tr>
<td>15</td>
<td>0.0073</td>
<td>0.0080</td>
<td>0.0044</td>
<td>0.0001</td>
<td>0.0454</td>
<td>GO:23061</td>
<td>Signal release</td>
</tr>
<tr>
<td>16</td>
<td>0.4590</td>
<td>0.0000</td>
<td>0.0168</td>
<td>0.0002</td>
<td>0.0473</td>
<td>GO:34330</td>
<td>Cell junction organization</td>
</tr>
</tbody>
</table>

Psychiatric Genomics Consortium (NatNeu, 2014)
PheWAS: Multi-phenotype studies

1. Motivation of phenome-wide association studies
   – PheWAS-informed phenotyping, improved GWAS power, etc
   – Electronic health record (EHR) contain rich personalized information

2. Modeling PheWAS using genetic information
   – inferring pleiotropic variants/pathways
   – Summary-statistics based inference of disease correlation

3. PheWAS without genetic: Rationale and EHR data properties and challenges

   – Tensor decomposition to model multi-dimensional EHR data
   – Deep learning to impute missing data and predict diseases
   – Latent topic model to infer latent disease grouping

5. Model EHR with non-missing at mechanism
   – Concepts of missing mechanism
   – Models leveraging missing information and inferring missing mechanism

6. Methods on modeling longitudinal EHR data
3. Overview of modeling PheWAS without genotype information using EHR data
Modeling PheWAS without genotype information

Rationale:
- Genotype information over is often not available over large patient cohort.
- However, as eluded at the beginning in the Alzheimer’s example, given the causal mediating phenotypes, diseases of interest are conditionally independent from genotype.
Electronic health record (EHR)

Simplified illustration of EHR data collection:
1. Patients are assigned an admission ID at a hospital visit
2. Patients subject to a select number of lab tests
3. Based on lab tests, patients are diagnosed with certain phenotypes (ICD-9 diagnosis code), recommended with medications (IONIC), and subject to medical treatments (ICD-9 procedure code)
EHR data matrix *with* and *without* missing indicators

<table>
<thead>
<tr>
<th>Patient 1</th>
<th>Patient 2</th>
<th>Patient 3</th>
<th>Patient 4</th>
<th>Patient 5</th>
<th>Patient 6</th>
<th>Patient 7</th>
<th>Patient 8</th>
<th>Patient 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab tests</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>Normal</td>
<td>Low</td>
<td>High</td>
<td>Normal</td>
<td>Low</td>
<td>High</td>
<td>Normal</td>
</tr>
<tr>
<td>Normal</td>
<td>Missing</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
<td>Missing</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Diag./Prescrip./Proc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recorded</td>
</tr>
<tr>
<td>Not recorded</td>
</tr>
</tbody>
</table>
EHR data are highly sparse and non-missing at random

Analysis based on MIMIC data (Johnson et al., SciDat 2016)
Phenome-wide association network modularity

- Thousand of ICD-9 codes form modules based on simple Spearman correlation across patients
- To model the PheWAN network modularity as well as patient clustering, machine-learning methods are used in many recent studies
# PheWAS: Multi-phenotype studies

1. **Motivation of phenome-wide association studies**
   - PheWAS-informed phenotyping, improved GWAS power, etc
   - Electronic health record (EHR) contain rich personalized information

2. **Modeling PheWAS using genetic information**
   - inferring pleiotropic variants/pathways
   - Summary-statistics based inference of disease correlation

3. **PheWAS without genetic: Rationale and EHR data properties and challenges**

4. **Machine learning methods for EHR data (assuming missing-at-random)**
   - Tensor decomposition to model multi-dimensional EHR data
   - Deep learning to impute missing data and predict diseases
   - Latent topic model to infer latent disease grouping

5. **Model EHR with non-missing at mechanism**
   - Concepts of missing mechanism
   - Models leveraging missing information and inferring missing mechanism

6. **Methods on modeling longitudinal EHR data**
4. Machine learning methods in modeling multivariate discrete data assuming missing at random

a. Tensor decomposition
b. Deep learning
c. Latent topic model
Marble: tensor decomposition to model high-dimensional EHR data

- Tensor decomposition is a generalized form of factorization that goes beyond 2D matrix.
- Outer product of D vectors are carried out to reconstruct D-dimensional tensor (illustrated 3D only).
- R tensors are sum by element-wise.
- Each tensor k associated with different data types and suggest their latent clustering.

Ho et al., (KDD 2014)
Imputing new patient EHR data

To impute EHR data, the trained phenotype projection is applied to new patient EHR tensor through the data generative process.

The latent phenotype memberships in lower representations then helps interpreting patients’ symptoms.

One limitation in tensor decomposition is that the method assumes a mapping from one data type to another for the same patient (e.g., lab test -> ICD9), which are often not available.

Ho et al., (KDD 2014)
Tensor-derived phenotypes (red) and procedures (blue)

<table>
<thead>
<tr>
<th>Diabetes Phenotype</th>
<th>Arthritis Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseases of other endocrine glands</td>
<td>Arthropathies and related disorders</td>
</tr>
<tr>
<td>Complications of surgical and medical care</td>
<td></td>
</tr>
<tr>
<td>Chemistry Pathology and Laboratory Tests</td>
<td>Physical Medicine and Rehabilitation Procedures</td>
</tr>
<tr>
<td>Organ or Disease Oriented Panels</td>
<td>Evaluation and Management of Other Outpatient Services</td>
</tr>
<tr>
<td>Hematology and Coagulation Procedures</td>
<td>Surgical Procedures on the Musculoskeletal System</td>
</tr>
<tr>
<td>Surgical Procedures on the Cardiovascular System</td>
<td>Diagnostic Radiology Procedures</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Heart Failure Phenotype</th>
<th>Severe Heart Failure Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other forms of heart disease</td>
<td>Other forms of heart disease</td>
</tr>
<tr>
<td>Complications of surgical and medical care</td>
<td>Pneumoconioses and other lung diseases</td>
</tr>
<tr>
<td>Symptoms</td>
<td>Ill-defined and unknown causes of morbidity and mortality</td>
</tr>
<tr>
<td>Cardiovascular Procedures</td>
<td>Hospital Inpatient Services</td>
</tr>
<tr>
<td>Hematology and Coagulation Procedures</td>
<td>Cardiovascular Procedures</td>
</tr>
<tr>
<td>Evaluation and Management of Other Outpatient Services</td>
<td></td>
</tr>
<tr>
<td>Surgical Procedures on the Cardiovascular System</td>
<td></td>
</tr>
<tr>
<td>Chemistry Pathology and Laboratory Tests</td>
<td></td>
</tr>
</tbody>
</table>

Grouping phenotypes (blue) and procedures (red) based on their shared tensor component membership
Restricted Boltzmann Machines for Collaborative Filtering

- Unsupervised learning of ratings system using collaborative restricted Boltzmann machines (RBM)
- The visible (input) units are the movie ratings by a user (i.e., 1-5 stars for Netflix movies)
- Visible units (movies) are connected to each other via hidden units
- The connection weights (W) are shared among users

Salakhutdinov et al., (ICML 2007)
Deep Poisson Factor analysis

Poisson factor analysis (PFA) module

Deep PFA module

Multi-task PFA module to jointly model medication, lab tests, ICD9 code

- $x_n$ is M-dimensional vector (e.g., M ICD9 codes) and factorized as Poisson model $x_n \sim \text{Poisson}(\Psi(\theta_n \circ h_n))$

- $\Psi$ is the $M \times K$ loading matrix

- $h_{kn}$ is a indicates whether factor $k$ is active

- $y_n$ is the target phenotypes

Henao et al., (JMLR, 2015)
Latent Dirichlet allocation (LDA)

Word frequency for each of $K$ latent topics

Latent topic mixture of the doc

Latent topic assignment to the word

$\alpha$ $\eta$ $\beta$

$\theta$ $z$ $\omega$

$N$ $M$
Grouping words by their topics

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Imputation using LDA

\[ p(w | w_{\text{obs}}) = \int \sum_z p(w | z) p(z | \theta) p(\theta | w_{\text{obs}}) d\theta \]

Very briefly, to impute missing word, we can take the expectation of that word with respect to the latent topics and topic mixture using the variational parameters learned from the training data.

Adapting LDA to model EHR data

Li et al., in prep.
PheWAS: Multi-phenotype studies

1. **Motivation of phenome-wide association studies**
   - PheWAS-informed phenotyping, improved GWAS power, etc
   - Electronic health record (EHR) contain rich personalized information

2. **Modeling PheWAS using genetic information**
   - inferring pleiotropic variants/pathways
   - Summary-statistics based inference of disease correlation

3. **PheWAS without genetic: Rationale and EHR data properties and challenges**

4. **Machine learning methods for EHR data (assuming missing-at-random)**
   - Tensor decomposition to model multi-dimensional EHR data
   - Deep learning to impute missing data and predict diseases
   - Latent topic model to infer latent disease grouping

5. **Model EHR with non-missing at mechanism**
   - Concepts of missing mechanism
   - Models leveraging missing information and inferring missing mechanism

6. **Methods on modeling longitudinal EHR data**
5. Machine learning methods in modeling discrete data with non-missing at random (NMAR)

a. Motivation: data not actually missing at random
b. Impute missing data and missing mechanism
Missing mechanism

• missing completely at random (MCAR): data are randomly missing without specific mechanism
• missing at random (MAR): missing status only depends on observed values
• Non-missing at random (NMAR): missing status depends on the values
Coding missing indicator as part of the data

Lin & Haug, JBI 2008
Using missing information improve medical predictions

Lin & Haug, JBI 2008
Distinct distribution for random and use-selected music ratings

(a) Yahoo! Random
(b) Yahoo! User

Marlin & Zemel, ACM 2009
Jointly model data and NMAR

Effect of rating value \( v \) and item \( d \) on the missing status

Rating frequency for each item in each mixture component

Latent cluster

Marlin & Zemel, ACM 2009
Imputation error on random and user-selected items

(c) Strong Rand NMAE  (d) Strong User NMAE
Adapting the recommendation model to impute missing EHR data

Li et al., in prep.
PheWAS: Multi-phenotype studies

1. Motivation of phenome-wide association studies
   - PheWAS-informed phenotyping, improved GWAS power, etc
   - Electronic health record (EHR) contain rich personalized information

2. Modeling PheWAS using genetic information
   - inferring pleiotropic variants/pathways
   - Summary-statistics based inference of disease correlation

3. PheWAS without genetic: Rationale and EHR data properties and challenges

   - Tensor decomposition to model multi-dimensional EHR data
   - Deep learning to impute missing data and predict diseases
   - Latent topic model to infer latent disease grouping

5. Model EHR with non-missing at mechanism
   - Concepts of missing mechanism
   - Models leveraging missing information and inferring missing mechanism

6. Methods on modeling longitudinal EHR data
6. Modeling longitudinal (time-series) continuous EHR data
Using Bayesian kernel matrix as prior to impute time-series EHR data

\[ \mathcal{K}_{tt'} = b_0 \exp(-a_0(t - t')^2) \]

\[ \Sigma_{0v} = \sigma_{0v} \mathcal{K}_{tt'} \]

\[ p(\mu_{kv}|\mu_{0v}, \sigma_{0v}) = \mathcal{N}(\mu_{kv}; \mu_{0v}, \Sigma_{0v}) \]

\[ p(x_n|r_n, \theta, \mu, \sigma) = \sum_k \theta_k \prod \mathcal{N}(x_{nvt}; \mu_{kvt}, \sigma_{kvt}^2)^{r_{nvt}} \]

(a) ML Estimate

(b) MAP Estimate

Marlin et al, ACM 2012
Deep supervised neural net

- ICD9 codes across times are concatenated together to form the input layer for each patient $X$
- One or more diseases of interests $Y$ are treated as the output units
- Diseases relationships are incorporated as the tree prior (right panel)

Che et al., (KDD 2015)
Summary

1. PheWAS elucidates phenotypic network by linking genetics to co-associated traits

2. Modeling multiple traits reveals genetics correlation between traits, helps elucidate co-morbidities, and improve causal signals detections at SNP and pathway levels

3. EHR data is extremely rich but modeling is challenging because the data are high dimensional, highly sparse, and non-missing at random

4. Machine learning methods especially generative models hold great promise to learn the compressed latent dimensions of the high-dimension EHR data and impute missing EHR data to reveal underlying disease network
PheWAS: Multi-phenotype studies

1. Motivation of phenome-wide association studies
   - PheWAS-informed phenotyping, improved GWAS power, etc
   - Electronic health record (EHR) contain rich personalized information

2. Modeling PheWAS using genetic information
   - inferring pleiotropic variants/pathways
   - Summary-statistics based inference of disease correlation

3. PheWAS without genetic: Rationale and EHR data properties and challenges

   - Tensor decomposition to model multi-dimensional EHR data
   - Deep learning to impute missing data and predict diseases
   - Latent topic model to infer latent disease grouping

5. Model EHR with non-missing at mechanism
   - Concepts of missing mechanism
   - Models leveraging missing information and inferring missing mechanism

6. Methods on modeling longitudinal EHR data