Electronic medical records (EMR) for discovery genomics research in immune-mediated disease

Robert Plenge, M.D., Ph.D.
i2b2 Annual Academic Users’ Group Meeting
June 28, 2011
Cost is dropping
Phenotyping remains expensive
How will we realize the ultimate potential of genomics if phenotyping is rate-limiting?
Can electronic medical records help?
Many risk loci remain “hidden”

~5% disease risk

- HLA studies
- Candidate gene studies
- GWAS and related methods
- RA-celiac overlap

all at $P<5\times10^{-8}$

>50,000 case-control samples

~15% disease risk

- HLA
- DR4
- “shared epitope” hypothesis


Zhernakova et al PLoS Genetics 2011
Clinically relevant subsets of RA

Lung and cardiovascular diseases, response to therapy
What are the options for collecting clinical data and DNA for genetic studies?
# Options for clinical + DNA

<table>
<thead>
<tr>
<th>design</th>
<th>Clinical data</th>
<th>DNA</th>
<th>Sample size</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>clinical trial</td>
<td>+++</td>
<td>+++</td>
<td>+</td>
<td>$$$</td>
</tr>
<tr>
<td>registry</td>
<td>++</td>
<td>+++</td>
<td>++</td>
<td>$$</td>
</tr>
<tr>
<td>claims data</td>
<td>+</td>
<td>n/a</td>
<td>+++</td>
<td>$</td>
</tr>
<tr>
<td>EMR</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>$</td>
</tr>
</tbody>
</table>
...and many others!
Outline of talk today

• **Demonstration**: developing an algorithm to define an RA cohort, proof-of-concept genomic studies

• **Portability**: implementing the EMR classification algorithm at other institutions

• **Application**: defining subsets of patients with clinically-relevant outcomes – and cardiovascular disease in particular
This is not a new idea...

Gabriel (1994) *Arthritis and Rheumatism*
...but EMR data are “dirty”

**Conclusion**: The sole reliance on such databases for the diagnosis of RA can result in substantial misdiagnosis.
Partners HealthCare: 4 million patients
Partners HealthCare: *linked by EMR*
Partners HealthCare: organized by i2b2
Our library of RA phenotypes

- Natural language processing (NLP)
  - disease terms (e.g., RA, lupus)
  - medications (e.g., methotrexate)
  - autoantibodies (e.g., CCP, RF)
  - radiographic erosions

<table>
<thead>
<tr>
<th>Concept/term</th>
<th>Accuracy of concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>presence of erosion</td>
<td>88%</td>
</tr>
<tr>
<td>seropositive</td>
<td>96%</td>
</tr>
<tr>
<td>CCP positive</td>
<td>98.7%</td>
</tr>
<tr>
<td>RF positive</td>
<td>99.3%</td>
</tr>
<tr>
<td>etanercept</td>
<td>100%</td>
</tr>
<tr>
<td>methotrexate</td>
<td>100%</td>
</tr>
</tbody>
</table>
Our library of RA phenotypes

- Natural language processing (NLP)
  - disease terms (e.g., RA, lupus)
  - medications (e.g., methotrexate)
  - autoantibodies (e.g., CCP, RF)
  - radiographic erosions

- Codified data
  - ICD9 disease codes
  - prescription medications
  - laboratory autoantibodies

Shawn Murphy
4 million patients

ICD9 RA and/or CCP checked
(goal = high sensitivity)

31,171 patients

Classification algorithm
(goal = high PPV)

3,585
RA patients

High PPV with adequate sensitivity

<table>
<thead>
<tr>
<th>Model</th>
<th>PPV (SE)</th>
<th>Sensitivity (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Codified + NLP</td>
<td>0.93 (0.02) ✪</td>
<td>0.63 (0.06)</td>
</tr>
<tr>
<td>NLP only</td>
<td>0.89 (0.02)</td>
<td>0.56 (0.05)</td>
</tr>
<tr>
<td>Codified only</td>
<td>0.88 (0.02)</td>
<td>0.51 (0.05)</td>
</tr>
</tbody>
</table>

392 out of 400 (98%) had definite or possible RA!

## Clinical features of patients

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>i2b2 RA</th>
<th>CORRONA</th>
</tr>
</thead>
<tbody>
<tr>
<td>total number</td>
<td>3,585</td>
<td>7,971</td>
</tr>
<tr>
<td>Mean age (SD)</td>
<td>57.5 (17.5)</td>
<td>58.9 (13.4)</td>
</tr>
<tr>
<td>Female (%)</td>
<td>79.9</td>
<td>74.5</td>
</tr>
<tr>
<td>Anti-CCP (%)</td>
<td>63</td>
<td>N/A</td>
</tr>
<tr>
<td>RF (%)</td>
<td>74.4</td>
<td>72.1</td>
</tr>
<tr>
<td>Erosions (%)</td>
<td>59.2</td>
<td>59.7</td>
</tr>
<tr>
<td>MTX (%)</td>
<td>59.5</td>
<td>52.8</td>
</tr>
<tr>
<td>Anti-TNF (%)</td>
<td>32.6</td>
<td>22.6</td>
</tr>
</tbody>
</table>

CCP has an OR = 1.5 for predicting erosions
4 million patients

ICD9 RA and/or CCP checked (goal = high sensitivity)

31,171 patients

Classification algorithm (goal = high PPV)

3,585 RA patients

Discarded blood for DNA
“On demand” biorepository

Within 1 year (at $30/sample):
1,800 RA cases
2,400 matched controls
June 2011: >35 RA risk loci

~5% disease risk

~15% disease risk

HLA studies
Candidate gene studies
GWAS and related methods
RA-celiac overlap

all at P<5x10^-8
>50,000 case-control samples

CD40
CCL21
CD244
IL2RB
TNFRSF14
PRKCQ
TNFAIP3
STAT4
TRAF1-C5
IL2RA
AFF3
CD20
CTLA4
CD81
CD247
UBE2L3
DDX6
UBASH3A
SH2B3
8q24

CD247
UBE2L3
DDX6
UBASH3A
SH2B3
8q24

OR similar in EMR cohort

Odds Ratio
SNP (ordered by chromosome and position)

~1,500 multi-ethnic RA cases and 1,500 matched controls

Kurreeman et al (2011) AJHG
Genetic risk score similar...

1. Assign each risk allele a weight based on OR
2. Sum weights across all risk alleles per person (= “genetic risk score”)
3. Compare distribution of weighted GRS in cases vs controls
4. Compare GWAS GRS vs EMR GRS

European ancestry

- RA case vs control
- GWAS vs EMR (no difference!)

Low GRS | High GRS

- $P_{GWAS} < 10^{-300}$
- $P_{EHR} = 5.55 \times 10^{-46}$
... across all ethnic groups
Outline of talk today

• **Demonstration**: developing an algorithm to define an RA cohort, proof-of-concept genomic studies

• **Portability**: implementing the EMR classification algorithm at other institutions

• **Application**: defining subsets of patients with clinically-relevant outcomes – and cardiovascular disease in particular
Portability to other institutions
Good portability to other institutions

<table>
<thead>
<tr>
<th>Institution</th>
<th>PPV (SE)</th>
<th>Sensitivity (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partners</td>
<td>0.93 (0.02)</td>
<td>0.63 (0.06)</td>
</tr>
<tr>
<td>Northwestern</td>
<td>0.80 (0.02)</td>
<td>0.50 (0.05)</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>0.92 (0.02)</td>
<td>0.54 (0.05)</td>
</tr>
</tbody>
</table>

**Note**: it took us 2 years to develop the algorithm at Partners, but ~2 months to apply it at Northwestern/Vanderbilt. *Still, this needs to be faster (e.g., 2 minutes!)*
Outline of talk today

• **Demonstration**: developing an algorithm to define an RA cohort, proof-of-concept genomic studies

• **Portability**: implementing the EMR classification algorithm at other institutions

• **Application**: defining subsets of patients with clinically-relevant outcomes – and cardiovascular disease in particular
Clinically relevant subsets of RA

Cardiovascular disease

Response to therapy

i2b2
Informatics for Integrating Biology & the Bedside

PGRN
Pharmacogenomics Research Network
Subset patients in clinically meaningful ways: causes of mortality

- cardiovascular disease (CVD) - 50%
- infection - 20%
- malignancy - 15%
- pulmonary - 10%
- other - 5%

There is a 2-fold increased risk of CVD in RA patients...is this due to inflammation?
The immune system in atherosclerosis

Göran K Hansson & Andreas Hermansson

Cardiovascular disease, a leading cause of mortality worldwide, is caused mainly by atherosclerosis, disease of blood vessels. Lesions of atherosclerosis contain macrophages, T cells and other cells of together with cholesterol that infiltrates from the blood. Targeted deletion of genes encoding costimulatory proinflammatory cytokines results in less disease in mouse models, whereas interference with regulatory it. Innate as well as adaptive immune responses have been identified in atherosclerosis, with components carrying low-density lipoprotein triggering inflammation, T cell activation and antibody production of atherosclerosis. Studies are now under way to develop new therapies based on these concepts of the involvement of the immune system in atherosclerosis.
Work in progress: *model of CVD in RA*

**cardiovascular disease**

**genetics + autoAbs**

↓ ???

CVD
Clinical characteristics of CVD in our EMR RA cohort

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>CAD yes, n=335 (7.5%)</th>
<th>CAD no, n=4118 (92.5%)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years, mean (SD)</td>
<td>72.9 (10.2)</td>
<td>60.0 (14.7)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Female gender, n (%)</td>
<td>207 (5.9)</td>
<td>3316 (94.1)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Male gender, n (%)</td>
<td>128 (13.8)</td>
<td>802 (68.2)</td>
<td></td>
</tr>
<tr>
<td>Race- white, n (%)</td>
<td>265 (79.0)</td>
<td>2714 (91.1)</td>
<td></td>
</tr>
<tr>
<td>Seropositive, n (%)</td>
<td>87 (67.4)</td>
<td>1099 (60.2)</td>
<td>0.10</td>
</tr>
<tr>
<td>MTX, n (%)</td>
<td>158 (47.2)</td>
<td>1851 (45.0)</td>
<td>0.45</td>
</tr>
<tr>
<td>TNFi, n (%)</td>
<td>96 (28.7)</td>
<td>1189 (28.8)</td>
<td>1.0</td>
</tr>
<tr>
<td>Plaquenil, n (%)</td>
<td>101 (30.2)</td>
<td>1200 (29.1)</td>
<td>0.71</td>
</tr>
<tr>
<td>CRP mean, median (mg/L)</td>
<td>10.2, 4.2</td>
<td>7.9, 2.0</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ESR_mean (mm/hr)</td>
<td>36.5</td>
<td>26.2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Erosions, n (%)</td>
<td>206 (61.5)</td>
<td>2168 (52.6)</td>
<td>0.0021</td>
</tr>
<tr>
<td>HTN, n (%)</td>
<td>252 (75.2)</td>
<td>1160 (28.2)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>DM, n (%)</td>
<td>108 (32.2)</td>
<td>375 (9.1)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Hyperlipidemia, n (%)</td>
<td>214 (63.9)</td>
<td>817 (19.8)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
Clinical characteristics of CVD in our EMR RA cohort

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1.06 (1.05, 1.08)</td>
</tr>
<tr>
<td>Female gender</td>
<td>0.35 (0.27, 0.46)</td>
</tr>
<tr>
<td>HTN</td>
<td>2.64 (1.88, 3.72)</td>
</tr>
<tr>
<td>DM</td>
<td>1.64 (1.20, 2.23)</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>2.86 (2.10, 3.90)</td>
</tr>
<tr>
<td>Ever smoker</td>
<td>2.30 (1.73, 3.04)</td>
</tr>
</tbody>
</table>
Conclusions
## EMRs for discovery research

<table>
<thead>
<tr>
<th>design</th>
<th>Clinical data</th>
<th>DNA</th>
<th>Sample size</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>clinical trial</td>
<td>+++</td>
<td>+++</td>
<td>+</td>
<td>$$$</td>
</tr>
<tr>
<td>registry</td>
<td>+</td>
<td>n/a</td>
<td>++</td>
<td>$$</td>
</tr>
<tr>
<td>claims data</td>
<td>+</td>
<td>n/a</td>
<td>+++</td>
<td>$</td>
</tr>
<tr>
<td>EMR</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>$</td>
</tr>
</tbody>
</table>

**Conclusion:** Informatics methods can yield accurate clinical data.
### EMRs for discovery research

<table>
<thead>
<tr>
<th>design</th>
<th>Clinical data</th>
<th>DNA</th>
<th>Sample size</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>clinical trial</td>
<td>+++</td>
<td>+++</td>
<td>+</td>
<td>$$$</td>
</tr>
<tr>
<td>registry</td>
<td>+++</td>
<td>n/a</td>
<td>+</td>
<td>$</td>
</tr>
<tr>
<td>claims data</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>$</td>
</tr>
</tbody>
</table>

**Conclusion:** EMR-based biorepositories for genetic studies yield effect sizes similar to traditional cohorts.
## EMRs for discovery research

<table>
<thead>
<tr>
<th>design</th>
<th>Clinical data</th>
<th>DNA</th>
<th>Sample size</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>clinical trial</td>
<td>+++</td>
<td>+++</td>
<td>+</td>
<td>$$$</td>
</tr>
<tr>
<td>registry</td>
<td></td>
<td></td>
<td></td>
<td>$$</td>
</tr>
<tr>
<td>claims data</td>
<td></td>
<td></td>
<td></td>
<td>$</td>
</tr>
<tr>
<td>EMR</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>$</td>
</tr>
</tbody>
</table>

**Conclusion:** It should be possible to extend this same framework to a multitude of other phenotypes across multiple institutions, **but**...
Of course, this is not the only way

- This approach will be good for some applications, and not good for others.

- This may serve as an effective way to generate hypotheses.

- There will always be a role for traditional registries.
i2b2 and PGRN acknowledgments

Zak Kohane
Susanne Churchill
Vivian Gainer
Kat Liao
Tianxi Cai
Shawn Murphy
Beth Karlson
Raul Guzman-Perez
Qing Zing
Pete Szolovits
Lee-Jen Wei
Lynn Bry (Crimson)
Ashwin Ananthakrishnan
Barbara Mawn
Zongqi Xia
Phil De Jager
& many others!

Josh Denny
Abel Kho
Will Thompson
Richard Pope
Anne Eyler
Chad Boomershine
Eric Ruderman
Art Mandelin
Tom Thomas
Robert Carroll
& others!
Funding
rplenge @ partners.org