Computational Phenotyping and Patient Stratification Using Electronic Health Records and Text Mining

Jessica Tenenbaum, PhD, FACMI
@jessiet1023
Requisite doom and gloom slide: Mental illness bad

Fact: 43.8 million adults experience mental illness in a given year.

1 in 5 adults in America experience a mental illness.

Impact

1st
Depression is the leading cause of disability worldwide, and is a major contributor to the global burden of disease.

-$193b
Serious mental illness costs America $193.2 billion in lost earning every year.

Anxiety 41%
Depression 39%

Mental Health Crisis for Graduate Students, C Flaherty, Inside Higher Ed, March 6, 2018.

https://www.nami.org/learn-more/mental-health-by-the-numbers
The DSM

Foundation of clinical diagnosis in mental health
Diagnostic heterogeneity: Schizophrenia

>=2 of the following, each present for a significant portion of time during a 1-month period... At least one must be 1, 2, or 3:

1. Delusions.
2. Hallucinations.
3. Disorganized speech (e.g., frequent derailment or incoherence).
4. Grossly disorganized or catatonic behavior.
5. Negative symptoms (i.e., diminished emotional expression or avolition).
‘-omics’ technologies can help stratify a seemingly homogeneous population.

- Adverse event
- Non-responder

**Diabetes**

- Exercise + Diet A
- A Exercise + Diet B
- No Exercise + Diet + Medication

Molecular Profiling

Courtesy of G Ginsburg, Duke University
If mental health phenotypes look like this,

Under the hood almost certainly looks like this.
Biological insights from 108 schizophrenia-associated genetic loci.

Schizophrenia Working Group of the Psychiatric Genomics Consortium.
Overlapping symptoms among mental health disorders

- MDD
- Bipolar Disorder
- Schizophrenia+

Symptoms:
- Depressed mood
- Mania
- Negative symptoms
- Delusions
- Abnormal motor behavior
- Hallucinations
- Scattered thinking

Conditions:
- Schizophrenia
- Schizopreniform
- Schizoaffective disorder
Common inheritance
# NIMH “RDoC”- Research Domain Criteria

## Negative Valence Systems

<table>
<thead>
<tr>
<th>Construct/Subconstruct</th>
<th>Genes Notice</th>
<th>Molecules</th>
<th>Cells</th>
<th>Circuits</th>
<th>Physiology</th>
<th>Behavior</th>
<th>Self-Report</th>
<th>Paradigms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute Threat (&quot;Fear&quot;)</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
<tr>
<td>Potential Threat (&quot;Anxiety&quot;)</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
<tr>
<td>Sustained Threat</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
<tr>
<td>Loss</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
<tr>
<td>Frustrative Nonreward</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
</tbody>
</table>

## Positive Valence Systems

<table>
<thead>
<tr>
<th>Construct/Subconstruct</th>
<th>Genes Notice</th>
<th>Molecules</th>
<th>Cells</th>
<th>Circuits</th>
<th>Physiology</th>
<th>Behavior</th>
<th>Self-Report</th>
<th>Paradigms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reward Responsiveness</td>
<td>Reward Anticipation</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
<tr>
<td>Initial Response to Reward</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
<tr>
<td>Reward Satiation</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
<tr>
<td>Reward Learning</td>
<td>Probabilistic and Reinforcement Learning</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
<tr>
<td>Reward Prediction Error</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
<tr>
<td>Habit - PVS</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
<tr>
<td>Reward Valuation</td>
<td>Reward (probability)</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
<tr>
<td>Delay</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
<tr>
<td>Effort</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
<td>Elements</td>
</tr>
</tbody>
</table>
Deconstructed, parsed, and diagnosed.

A hypothetical example illustrates how precision medicine might deconstruct traditional symptom-based categories. Patients with a range of mood disorders are studied across several analytical platforms to parse current heterogeneous syndromes into homogeneous clusters.

**Symptom-based categories**
- Major depressive disorder
- Mild depression (dysthymia)
- Bipolar depression

**Integrated data**
- Genetic risk
  - polygenic risk score
- Brain activity
  - insula cortex
- Physiology
  - inflammatory markers
- Behavioral process
  - affective bias
- Life experience
  - social, cultural, and environmental factors

**Data-driven categories**
- Cluster 1
- Cluster 2
- Cluster 3
- Cluster 4

Prospective replication and stratified clinical trials

Insel & Cuthbert, Science 2015
A Framingham for the omics era

$35 \text{ M}

N = 12k
The Learning Healthcare System

Reimagining the research-practice relationship: policy recommendations for informatics-enabled evidence-generation across the US health system

Peter J Embi, Rachel Richesson, Jessica Tenenbaum, Joseph Kannry, Charles Friedman, Indra Neil Sarkar, Jeff Smith,
The members of 2016 AMIA Policy Invitational Planning Committee

JAMIA Open, ooy056, https://doi.org/10.1093/jamiaopen/ooy056
Published: 16 January 2019  Article history

Tenenbaum et al., An informatics research agenda to support precision medicine: Seven key areas, JAMIA 2016
2 cohorts*, overlapping goals

<table>
<thead>
<tr>
<th></th>
<th>K01</th>
<th>Bass Connections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disorder</td>
<td>Mental illness including SCZ subset</td>
<td>Schizophrenia</td>
</tr>
<tr>
<td>Prediction</td>
<td>Response to medication</td>
<td>30 day readmission</td>
</tr>
<tr>
<td>Date range</td>
<td>7/1/96 - 8/18/16</td>
<td>1/1/14 - 4/10/17</td>
</tr>
<tr>
<td>Defined by</td>
<td>ICD, Notes, Problem list</td>
<td>ICD only</td>
</tr>
<tr>
<td>N</td>
<td>10,904</td>
<td>5822</td>
</tr>
</tbody>
</table>

*Due to IRB peculiarities
Bass schizophrenia cohort (Initial N = 5822)
Computational phenotyping

“Simple” example: Diabetes

Richesson et al. JAMIA 2013
Challenges in MH electronic phenotyping

- Diagnoses heterogeneous
- Labs not typically used
- Medications not specific
- Stigma, extra-sensitive PHI
## CMS Chronic Conditions Data Warehouse (CCW)
### Other Chronic or Potentially Disabling Condition Algorithms

(Rev. 01/2016)

<table>
<thead>
<tr>
<th>Algorithms¹</th>
<th>Valid ICD-9 / MS DRG / HCPCS Codes²</th>
<th>Valid ICD-10 Codes²</th>
<th>Number/Type of Claims to Qualify³</th>
</tr>
</thead>
</table>
The utility of notes

Wei WQ, et al 2015 JAMIA
Big News! DEDUCE has text search!
Schizophrenia

**Total**

N = 10,904

**Notes**

N = 4564

**Problem list**

N = 3599

ICD N = 8248

* Not yet validated
Wide spectrum of approaches to NLP

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun. It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!
Text mining: term extraction

• “Elizabeth, a 25-year-old female, CEO of a biotech startup company in Silicon Valley, presented with shortness of breath and delusions. Negative for asthma, CPOE.”
“Elizabeth, a 25-year-old female, CEO of a biotech startup company in Silicon Valley, presented with shortness of breath and delusions. Negative for asthma, CPOE.”

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Shortness of breath</th>
<th>Delusions</th>
<th>Asthma</th>
<th>CPOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>Female</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>
Mining EHR data for patient stratification in MH

• Build feature matrix (80% of effort!)
  – Combining structured data with terms extracted via text mining

• Clustering
  – Future direction- biomarker discovery

• Predictive modeling
  – Bass:
    Who will be readmitted in < 30 days?
  – K01
    Which patients will respond to drug X?
Feature engineering

Additional derived features:

- Visits in past year
- Slope of curve of visit frequency
- Etc.
How to handle time?
It depends...

[Diagram showing a matrix with columns for Persons, Features, Procedures, Drugs, Diseases, Devices, and Encounters. The rows show time with values 1, 0, or -1.]
Feature reduction

- Thousands of drugs, tens of thousands of ICD codes
- Features not useful if they are too frequent or too rare
- Use hierarchical terminologies to aggregate terms
  - SNOMED for diseases
  - ATC for drugs

http://bok.ahima.org/
What level should you aggregate to?

Information content =

# of patients who have the concept or a descendant of the concept in their record

total # of patients
IC defines an abstraction level

Nigam Shah, Stanford University
Recall diagnostic criteria for schizophrenia

2 out of 5:

1. Delusions.
2. Hallucinations.
3. Disorganized speech (e.g., frequent derailment or incoherence).
4. Grossly disorganized or catatonic behavior.
5. Negative symptoms (i.e., diminished emotional expression or avolition).
MH terminology - it’s complicated

3. Disorganized speech
   - Disorganized thinking
   - Tangential thinking
   - Word salad
   - Flight of ideas
   - Loosening of associations
   - Incoherent speech
   - Incoherent thinking
   - Rambling speech
   - Thought block
   - Clang associations
   - ...

4. Grossly disorganized or catatonic behavior
   ...

5. Negative symptoms
   ...

Duke Medicine
NLP tools, like standards- so many to choose from...

BioPortal

Annotator

CLAMP
Clinical Language Annotation, Modeling, and Processing Toolkit

CLAKES

MetaMap - A Tool For Recognizing UMLS Concepts in Text
clinical Text Analysis and Knowledge Extraction System (cTAKES)
Savova et al. JAMIA 2010
Standardized terminology of biomedical concepts: SNOMED

• 340,000+ concepts
• Many (100+) semantic types
• First pass:
  – Mental or Behavioral Dysfunction
  – Mental Process
  – Sign or Symptom
  – Disease or Syndrome
  – Finding
Custom dictionary

• Start with schizophrenia symptoms
• Add common acronyms & abbreviations
• Add synonyms and related terms
Run cTAKES on clinical notes, output JSON

```

{
  "termFoundInDoc": "visual hallucinations",
  "termStartOffset": 2781,
  "termLength": 21,
  "subject": "patient",
  "annotationType": "SignSymptomMention",
  "isNegated": "true",
  "isFamilyHistory": "false",
  "findings": [
    {
      "ontologyName": "SNOMEDCT_US",
      "ontologyTermCode": "64269007",
      "preferredText": "Hallucinations, Visual",
      "umlsCui": "C0233763",
      "umlsTui": "T184"
    }
  ]
}
```
Run cTAKES on clinical notes, output JSON

```
{
  "termFoundInDoc": "visual hallucinations",
  "termStartOffset": 2781,
  "termLength": 21,
  "subject": "patient",
  "annotationType": "SignSymptomMention",
  "isNegated": "true",
  "isFamilyHistory": "false",
  "findings": [
    {
      "ontologyName": "SNOMEDCT_US",
      "ontologyTermCode": "64269007",
      "preferredText": "Hallucinations, Visual",
      "umlsCui": "C0233763",
      "umlsTui": "T184"
    }
  ]
}
```
Run cTAKES on clinical notes, output JSON

```
{
  "termFoundInDoc": "visual hallucinations",
  "termStartOffset": 2781,
  "termLength": 21,
  "subject": "patient",
  "annotationType": "SignSymptomMention",
  "isNegated": "true",
  "isFamilyHistory": "false",
  "findings": [
    {
      "ontologyName": "SNOMEDCT_US",
      "ontologyTermCode": "64269007",
      "preferredText": "Hallucinations, Visual",
      "umlsCui": "C0233763",
      "umlsTui": "T184"
    }
  ]
}
```
JSON converted to feature matrix

```json
{"termFoundInDoc":"visual hallucinations","termStartOffset":2781,"termLength":21,"subject":"patient","annotationType":"SignSymptomMention","isNegated":"true","isFamilyHistory":"false","findings":[{"ontologyName":"SNOMEDCT_US","ontologyTermCode":64269007,"preferredText":"Hallucinations, Visual","umlsCui":"C0233763","umlsTui":"T184"}]
```
Notes vs. encounters

Separate positive vs. negated - may see both in one note

<table>
<thead>
<tr>
<th>Notes</th>
<th>SNO-1</th>
<th>SNO-2</th>
<th>SNO-3</th>
<th>...</th>
<th>SNO-N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Note 1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Note 2</td>
<td></td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Note 3</td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Note 4</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Aggregation variations depend on use case
You say hallucinations, I say...

- hallucinate
- hallucinating
- avh
- a/v/h
- ah
- vh
You say Plan, I say…

• PLAN (PLA2G6-Associated Neurodegeneration)
• aka Infantile Neuroaxonal Dystrophy??
VERY early results
(credit Dylan Liu)

• Topic modeling using LDA
• tSNE for visualization - t-distributed stochastic neighbor embedding

- Topic 1
  0.586*"Anhedonia" + 0.143*"Mutism" + 0.122*"Paranoid delusion" + 0.068*"Stupor" + 0.052*"Hallucinations, Visual" + 0.008*"Hallucinations, Auditory" + 0.007*"Negativism in catatonia" + 0.005*"Catatonia" + 0.004*"Poor eye contact" + 0.001*"Paranoid ideations"
What other terms correspond to symptoms?

3. Disorganized speech
   - Disorganized thinking
   - Tangential thinking
   - Word salad
   - Flight of ideas
   - Loosening of associations
   - Incoherent speech
   - Incoherent thinking
   - Rambling speech
   - Thought block
   - Clang associations
   - ...

Based on DSM, SNOMED, clinical experts’ input
Last but not least:
Informatics approaches to term set expansion

Seeded Pattern Discovery

Seed terms
- Word salad
- Flight of ideas
- Word block

Pattern discovery
- Shows signs of...
- Has developed...
- ...nearly every day

Instance extraction
- Clang associations
- Ideas of reference
- Rambling speech

Word2Vec
- Male-Female
- Verb tense
- king → man
- woman
- queen
- walked
- walking
- swimming
- swam
About those fully identified psychiatric notes...

- IRB approved Pro00081628, Pro00074074- waiver of consent
- Undergraduates do NOT have access to any PHI
- Best attempts to exclude all Duke-affiliated patients
  - “Duke affiliate” flag
  - Free text employer
  - Health insurance plan
#OpenScience and #Reproducibility

- Not feasible to share all data
- May be able to share some de-identified data - Big Duke
- BUT can share code
- Code versioning (GitLab)
Brain disorders? Precisely
Precision medicine comes to psychiatry

By Thomas R. Insel and Bruce N. Cuthbert

STAND UP TO STIGMA
Let’s talk about MENTAL HEALTH

J Kalat, Biological Psychology
In summary

• Mental health is area of dire need, major opportunity for precision medicine approach
• EHRs enable data mining, learning healthcare system
• Text mining facilitates extraction of rich phenotypic data from notes
• New methods are needed to enhance mental health concept representation in standardized terminologies
Acknowledgements

**Clinical Experts:**
- Jane Gagliardi, MD
- Gopalkumar Rakesh, MD
- John Beyer, MD

**Team NLP:**
- Myung Woo, MD
- Dylan Liu
- Scarlett Zhou
- Steve Evans

**Bass Students:**
- Sanya Kochhar
- Abhi Jhadav
- Aakash Thumaty
- Kamyar Yazdani
- Pranav Warman
- Casey Riffel

**Also:**
- Team PACE
- Colette Blach
- Nigam Shah, MD
- W. Ed Hammond, PhD

K01LM012529, Bass Connections
Evidence-based prescribing?
Predict drug responders, but how do we know?

**Responder:** stays on steady dose of Drug X for $\geq 3$ months without additional meds

**Non-responder:** does not achieve stable dose for greater than 3 months, or changes Rx

**Partial responder:** stays on steady dose of Drug X for $\geq 3$ months with additional meds

**Indeterminate:** initial Rx but no subsequent scripts for Rx 1 or other