Bioinformatics based Phenotyping Algorithms from Electronic Medical Records for Rotator Cuff Tears

Kindred K. Harris, BS
RREMMS Research Institution: Vanderbilt University
Home Institution: David Geffen School of Medicine at UCLA
Disclosures

• This research was sponsored in part by The Association of Academic Physiatrists (AAP) and the Foundation for PM&R
Shoulder pain

• **Common**
  – 3rd most common musculoskeletal complaint
    • Yearly prevalence ranges from 14-34%

• **Etiology**
  – ~1% of the population ≥ age 45 present to primary care settings with shoulder pain, usually from atraumatic causes, annually

• **Costly**
  • Outpatient visits: $5,000,000
  • Diagnostic Imaging: $20,000,000
  • Physical Therapy: 13,000,000

**Most Common Etiology: Rotator Cuff Pathology**
Electronic medical record (EMR)

• **Benefit**
  – Valuable resource for large scale clinical outcome and genome-association studies

• **Drawbacks**
  – Individual case review of entire data repositories is cumbersome and frankly unfeasible
  – Inaccurate information (i.e. coding)
    » Most degenerative RTC tears are asymptomatic and do not have a billing code associated with pathology
Objective

• To develop high accuracy bioinformatics algorithms to efficiently identify subjects with and without rotator cuff tears in a robust EMR repository
Synthetic Derivative

- EMR at Vanderbilt University Medical Center
  - de-identified image of EMR system
  - 2.5 million patients
  - longitudinal data collected over decades
  - sex proportionate
  - provides information such as patient’s age, sex, ethnicity, history of trauma, counts of ICD9/10 codes reported, and descriptions of rotator cuff tendons integrity (full-thickness, partial-thickness, tendinopathy)
Selection of potential rotator cuff tear (RCT)

Individuals age 40-75

AND ≥1 of the following

- ICD9, ICD10, or CPT related to rotator cuff pathology
- MRI report matching for rotator cuff pathology
- Clinician note matching for rotator cuff pathology
- Rehabilitation note containing a diagnosis of rotator cuff tear
Selection of potential intact rotator cuff (IRC)

Individuals age 40-75

AND ≥1 of the following

- No variable related to rotator cuff pathology in previous slide
- Explicit mention of “intact” RTC in MRI report
- Explicit mention of “intact” RTC in clinician note
- Explicit mention of “intact” RTC in operative note
SD as of 01/01/2017
(Age 40-75)

Billing codes or radiology report c/w rotator cuff disorder
+
-

Results
Likely RCT = 14,690
Likely IRC = 1,066,053

Training set of 1000 patients
Likely RCT (500) IRC (500)
Gold Standard - Chart Review

- RCT:
  - (+) shoulder MRI report of tear; OR
  - (+) operative report of tear; OR
  - (+) H&P mentioning history of rotator cuff repair

- NRC:
  - (-) shoulder MRI report of tear; AND
  - (-) operative report of tear; AND
  - (-) H&P mentioning history of rotator cuff repair
Algorithm development

Training set of equal numbers of likely tear and intact rotator cuff (1000)

Chart review
(~5 minutes per chart for experienced researcher)

Confirmed RCT (193)

Confirmed NRC (100)

Algorithm development (92 features)
(1) Billing codes: ICD9, ICD10, CPT
(2) Regular expression (RegEx)
  - Radiology report
  - Shoulder MRI report
  - Operative Report
  - H&P
  - Rehabilitation Note

Included: 293
Excluded: 707
• Indeterminate phenotype

Model development by combining informatics variables
High-yield Variables Selected Based on Clinical Acumen

- Age
- Sex
- ICD9 codes related to rotator cuff pathology
- CPT codes related to rotator cuff surgery
- Expression of general radiology report denoting ALL rotator cuff pathology
- Expression of shoulder MRI report denoting ALL rotator cuff pathology
- Expression of shoulder MRI report denoting rotator cuff tear of repair
- Expression of H&P denoting ALL rotator cuff pathology
- Expression of operative report denoting ALL rotator cuff pathology
Logistic regression modelling of RCT

| Variable                                      | Coef  | S.E.   | Wald Z | Pr(>|Z|) |
|-----------------------------------------------|-------|--------|--------|----------|
| Intercept                                     | -3.0833 | 1.2225 | -2.52  | 0.0117   |
| Age                                           | 0.0289 | 0.0180 | 1.60   | 0.1093   |
| Sex = male                                    | 0.3624 | 0.3099 | 1.17   | 0.2422   |
| ICD9_RC = 1 or 2                              | -0.2294 | 0.5810 | -0.39  | 0.6929   |
| ICD9_RC > 2                                   | -0.2184 | 0.6015 | -0.36  | 0.7165   |
| CPT_RC = 1 or 2                               | 1.0462 | 0.5446 | 1.92   | 0.0547   |
| CPT_RC > 2                                    | 1.1988 | 0.4495 | 2.67   | 0.0077   |
| ShoulderMRI_RCT/RCR/RCI ≥ 1                   | 0.0924 | 0.6505 | 0.14   | 0.8871   |
| ShoulderMRI_RCT/RCR ≥ 1                       | 0.1521 | 0.7005 | 0.22   | 0.8281   |
| H&P_RCT/RCR/RCI = 1 or 2                      | 1.4925 | 0.3862 | 3.86   | 0.0001   |
| H&P_RCT/RCR/RCI > 2                           | 3.0824 | 0.5336 | 5.78   | <0.0001  |
| OP_RCT/RCR/RCI ≥ 1                            | 0.1919 | 0.6805 | 0.28   | 0.7780   |
| Rad_RCT/RCR/RCI ≥ 1                           | 0.7266 | 0.4041 | 1.80   | 0.0722   |
Odds ratio for RCT

- Age – 66:53
- Sex – female: male
- ICD9_RC = 1 or 2 : 0
- ICD9_RC >2 : 0
- CPT_RC = 1 or 2 : 0
- CPT_RC > 2 : 0
- ShoulderMRI_RCT/RCR/RCI
- ShoulderMRI_RCT/RCR ≥ 1 :
- H&P_RCT/RCR/RCI = 1 or 2 :
- H&P_RCT/RCR/RCI >2 : 0
- OP_RCT/RCR/RCI ≥ 1 : 0
- Rad_RCT/RCR/RCI ≥ 1 : 0
Nomogram to Predict RCT

- Age
- Sex – female: male
- ICD9_RC
- CPT_RC
- ShoulderMRI_RCT/RCR/RCI
- ShoulderMRI_RCT/RCR
- H&P_RCT/RCR/RCI
- OP_RCT/RCR/RCI
- Rad_RCT/RCR/RCI

Total points
Probability of RCT
Full Model Calibration

![Graph showing model calibration]

- **Actual** line deviates from the **Ideal** line, indicating a lack of perfect model calibration.
Conclusion

• The informatics variables showing highest precision in predicting RCT are:
  – CPT\textsubscript{RC}
  – H\&P\textsubscript{RCT/RCR/RCI}
  – Rad\textsubscript{RCT/RCR/RCI}

• Phenotypic algorithm shows its promise in facilitating large scaled clinical or genome studies by eliminating the need for case-by-case chart review

• Next Step: External validation of the algorithm
Acknowledgements

• Chan Gao, MD, PhD - Resident, Vanderbilt University Medical Center (Role: Co-Author)
• Nitin Jain, MD - Associate Professor, Physical Medicine and Rehabilitation. Director of PM&R Research, Vanderbilt Medical Center (Role: Corresponding Author)
• Department of Biomedical Informatics Vanderbilt University Medical Center
• Additional integral team members: genetic epidemiologists, data analysts, senior biostatisticians, and Phenotyping and PheWAS research core at Vanderbilt University Medical Center
References


