Big Data and Big Crowds: Getting Useful Data from Text Fields Using Large Data Sets

Janna L. Friedly  
*University of Washington - Seattle Campus, friedlyj@uw.edu*

Jeffery G. Jarvik  
*University of Washington - Seattle Campus, jarvikj@uw.edu*

Bryan Comstock  
*University of Washington - Seattle Campus, bac4@uw.edu*

Follow this and additional works at: [http://repository.academyhealth.org/webinars](http://repository.academyhealth.org/webinars)

Part of the [Health Information Technology Commons](http://academyhealth.org/webinars) and the [Health Services Research Commons](http://academyhealth.org/webinars)

**Recommended Citation**


[http://repository.academyhealth.org/webinars/17](http://repository.academyhealth.org/webinars/17)

This Video/Media is brought to you for free and open access by the Events at EDM Forum Community. It has been accepted for inclusion in Webinars by an authorized administrator of EDM Forum Community.
EDM Forum: Big Data and Big Crowds: Getting Useful Data from Text Fields Using Large Data Sets

Janna Friedly, M.D., University of Washington (moderator); Bryan Comstock, M.S., University of Washington; and Jeffrey (Jerry) Jarvik, M.D., M.P.H., University of Washington

Wednesday, October 15, 2014
Welcome

→ Beth Johnson, M.P.H.
  – Senior Manager, Research and Education in HSR
  – Project Manager, Electronic Data Methods (EDM) Forum
  – Managing Editor, eGEMs

Follow the conversation on Twitter!
#EDMForum @edm_ah @academyhealth
AcademyHealth: Improving Health & Health Care

AcademyHealth is a leading national organization serving the fields of health services and policy research and the professionals who produce and use this important work.

Together with our members, we offer programs and services that support the development and use of rigorous, relevant and timely evidence to:

1. Increase the quality, accessibility and value of health care,
2. Reduce disparities, and
3. Improve health.

A trusted broker of information, AcademyHealth brings stakeholders together to address the current and future needs of an evolving health system, inform health policy, and translate evidence into action.
Evidence, Data, and Methods to Build Learning Health Systems (LHS) of the Future

- Advancing learning and collaboration in big data and big science in healthcare
- Working with cutting-edge research and quality improvement networks that reach more than 1 in 10 American across priority populations and conditions
- Sign up for updates: edmforum@academyhealth.org
Check Out eGEMs!

eGEMs (Generating Evidence and Methods to improve patient outcomes)

- Free, peer-reviewed, open access, e-publication
- Focus: generalizable lessons learned within analytic methods, clinical informatics, governance, and the learning health system
- 62 publications with over 26,500 downloads!

We are actively accepting submissions and recruiting qualified peer-reviewers. Learn more at repository.academyhealth.org/egems
The audio and slide presentation will be delivered directly to your computer

- Speakers or headphones are required to hear the audio portion of the webinar.

- If you do not hear any audio now, check your computer’s speaker settings and volume.

- If you need an alternate method of accessing audio, please submit a question through the Q&A pod.
Technical Assistance

→ Live technical assistance:
  – Call Adobe Connect at (800) 422-3623

→ Refer to the ‘Technical Assistance’ box in the bottom left corner for tips to resolve common technical difficulties.

→ Please turn off your pop-up blocker in order to take a survey
Questions may be submitted at any time during the presentation

To submit a question:

1. Click in the Q&A box on the left side of your screen
2. Type your question into the dialog box and click the Send button
Today’s Moderator

Janna Friedly, M.D.
- Assistant Professor, University of Washington Department of Rehabilitation Medicine
- Medical Director, outpatient Rehabilitation Medicine clinics and the Amputee Rehabilitation program, Harborview Medical Center
Learning Objectives

→ At the conclusion of the session, participants will be able to:
  – Review the BOLD/BOLDER registry
  – Review the approaches for abstracting data from imaging reports
  – Describe the accuracy, cost, and time of Amazon mTurk and NLP compared with the trained researcher abstraction
Today’s Faculty

Bryan A. Comstock, M.S.

Jeffrey (Jerry) Jarvik, M.D., M.P.H.
Big Data vs. Big Crowds:
Getting Useful Data from Text Fields Using Large Data Sets

Bryan A. Comstock, MS
Senior Biostatistician
Center for Biomedical Statistics
University of Washington, Seattle

Jeffrey (Jerry) G. Jarvik, MD MPH
Professor of Radiology, Neurological Surgery and Health Services
University of Washington, Seattle
Bryan A. Comstock, MS

Disclosures
Cofounder of www.CSATS.com (June 2014), a company dedicated to efficiently evaluating surgical technical skills via the crowd

Jeffrey (Jerry) G. Jarvik, MD MPH

Acknowledgements
AHRQ: 1R01HS022972-01
NIH: UH2 AT007766-01; UH3 AR066795-03

Disclosures
Physiosonix (ultrasound company): Founder/stockholder
Healthhelp (utilization review): Consultant
GE Healthcare: CER Advisory Board (former); GERRAF Advisory Board
Springer Publishing: Evidence-based Neuroradiology: Co-Editor
Aim:

To examine the time, cost, and accuracy of three approaches for abstracting information from lumbar imaging (MR, CT, X-ray) text reports:

- Trained human abstracters with a gold standard
- Crowdsourcing via Amazon.com’s Mechanical Turk™
- Natural language processing (NLP)
Example X-ray report:

CLINICAL HISTORY: Low back pain LUMBOSACRAL SPINE SERIES Weight-bearing AP, lateral and coned lumbosacral lateral views. COMPARISON: None 

FINDINGS: Mild overall loss of height of L5. Mildly wedged configuration of L4 with minimal wedging of L3 which could be chronic remodeling. There could be minimal retrolisthesis of L3. No other malalignment is demonstrated. Mild to moderate multilevel discogenic degenerative changes with endplate sclerosis, small endplate osteophytes. Lumbosacral junction mild to moderate facet arthrosis. Osteopenia. Symmetric SI joints within normal limits for patient age. Pelvic calcifications are likely vascular. IMPRESSION: Appearance of mild L5 compression. Mildly / minimally wedged configuration of L3-L4; this could reflect chronic remodeling. Mild to moderate multilevel discogenic degenerative disease.
Example X-ray report:

CLINICAL HISTORY: Low back pain LUMBOSACRAL SPINE SERIES Weight-bearing AP, lateral and coned lumbosacral lateral views. COMPARISON: None FINDINGS: Mild overall loss of height of L5. Mildly wedged configuration of L4 with minimal wedging of L3 which could be chronic remodeling. There could be minimal retrolisthesis of L3. No other malalignment is demonstrated. Mild to moderate multilevel discogenic degenerative changes with endplate sclerosis, small endplate osteophytes. Lumbosacral junction mild to moderate facet arthrosis. Osteopenia. Symmetric SI joints within normal limits for patient age. Pelvic calcifications are likely vascular. IMPRESSION: Appearance of mild L5 compression. Mildly / minimally wedged configuration of L3-L4; this could reflect chronic remodeling. Mild to moderate multilevel discogenic degenerative disease.
STUDY NAME: LUMBAR SPINE MRI W/O CONTRAST
Clinical Information: Chronic worsening LBP.
FINDINGS:
- Mild multilevel degenerative disc and spondylitic changes, greatest at L2-L3 and L5-S1 are stable. Conus is normal terminating at L1-L2. L5 is transitional. Postsurgical changes L5-S1 at L4-L5 levels, status post posterior bone fusion.
- L1-L2: Stable mild disc bulge. No significant canal or foraminal stenosis.
- L2-L3: Mild concentric disc bulge without significant canal or foraminal stenosis. Mild bilateral facet arthropathy, similar to prior exam.
- L3-L4: No significant disc abnormality. Moderate bilateral facet arthropathy, similar to prior exam.
- L4-L5: Postsurgical changes. No canal or foraminal stenosis.
- L5-S1: Disc protrusion eccentric to the left, similar to prior exam, without nerve root impingement. Right-sided S2 Tarlov cyst unchanged from prior exam.
## Findings of interest:

<table>
<thead>
<tr>
<th>X-ray</th>
<th>CT</th>
<th>MRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

1. Disc space height loss
2. Central canal stenosis
3. Lateral recess or foraminal stenosis
4. Moderate or severe facet degeneration
5. Listhesis
6. *Severe* scoliosis
7. *Severe* kyphosis
8. Compression fractures
9. *Moderate or severe* disc desiccation
10. *Moderate or severe* disc bulge
11. *Moderate or severe* disc protrusion
12. *Moderate or severe* disc extrusion
13. Disc herniation
14. Endplate edema
15. Nerve root compression
Motivation:

Back Pain Outcomes using Longitudinal Data – Extended Research (BOLDER)
Agency for Healthcare Research & Quality (AHRQ)

=> 6,100 imaging reports

Lumbar Imaging with Reporting of Epidemiology (LIRE)
National Institute of Health Common Fund

=> 270,000+ imaging reports
5,239 patients ≥ 65 with new primary care visits for back pain

3 integrated systems: Kaiser Perm N. CA, Henry Ford Health System, Harvard Vanguard/Harvard Pilgrim

Identify patients using Health Care Information Systems

Contacted at 3, 6, 12, 24 months

PROs: pain, disability, depression, anxiety, HRQoL
Background – BOLD/BOLDER

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions

- Prospective observational cohort
- Compare effectiveness of early imaging to no early imaging in seniors with new visit for LBP
- Propensity-matched comparison

Outcomes
- Disability (RMDQ)
- Pain
- Subsequent resource utilization through EMR data pull
Lumbar spine imaging frequently reveals incidental findings.

These findings may have an adverse effect on:

- Subsequent healthcare utilization
- Patient health related quality of life
Background - LIRE

LIRE

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions
Study the impact of inserting epidemiologic benchmark information into lumbar imaging reports (Xray, CT, MR)

- N=98 clinics randomized across Henry Ford, Kaiser N. California, Group Health, Mayo Clinic Health Systems
- Stepped wedge cluster randomized trial
- Outcomes:
  - Relative value units (RVUs)
  - Rate of new narcotic prescriptions
  - Rate of subsequent imaging
- 270,000 patients with an index lumbar image
Background - LIRE

Intervention

Example intervention text:

“The following findings are so common in normal, pain-free volunteers that while we report their presence, they must be interpreted with caution and in the context of the clinical situation. Among people over the age of 60 who do **not** have back pain, an MRI will find that about:

- 9 in 10 have disc degeneration
- 9 in 10 have disc signal loss (desiccation)
- 8 in 10 have disk height loss
- 8 in 10 have a bulging disc
- 4 in 10 have an annular fissure
- 4 in 10 have a disc protrusion
- 4 in 10 have facet degeneration”
Methods

Background
- Imaging Reports
- BOLD/BOLDER
- LIRE

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions
**Methods – Study Design**

**Background**
- Imaging Reports
- BOLD/BOLDER
- LIRE

**Methods**
- Abstracters
- Crowdsourcing
- NLP
- Study design

**Results**
- Time
- Cost
- Accuracy

**Conclusions**

Select n=400 reports*

180 Xrays
54 CTs
166 MRIs

Crowd

10 ratings per imaging finding

NLP

1 rating per imaging finding

2 trained abstracters + senior Radiologist

2 ratings per imaging finding + adjudication

7 – 10 positive ratings = “probable finding”
- Secondary analysis: 5 – 10 positive ratings

* 400 BOLD imaging reports, one of 10 LIRE CPT codes
Methods - Abstracters

**Multi-rater data abstraction**

1. Select raters

2. Create abstraction rules, form, and database

3. Pilot test and assess inter-rater reliability

4. Distribute records to each rater

5. Raters evaluate assigned records

6. Adjudicate discrepant responses with gold standard
### Example MRI Reporting Form

<table>
<thead>
<tr>
<th>Field</th>
<th>Data Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disc Desiccation/Signal</td>
<td>none/mild</td>
</tr>
<tr>
<td>Disc Space Height</td>
<td>none/small</td>
</tr>
<tr>
<td>Disc Bulge</td>
<td>medium/large</td>
</tr>
<tr>
<td>Disc Protrusion</td>
<td>medium/large</td>
</tr>
<tr>
<td>Disc Extrusion</td>
<td>medium/large</td>
</tr>
<tr>
<td>Disc Herniation</td>
<td>medium/large</td>
</tr>
<tr>
<td>Endplate Edema/Type 1 Modic changes</td>
<td>none/mild</td>
</tr>
<tr>
<td>Central Stenosis</td>
<td>no</td>
</tr>
<tr>
<td>Lateral Recess or Foraminal Stenosis</td>
<td>yes</td>
</tr>
<tr>
<td>Nerve root Compression</td>
<td>yes</td>
</tr>
<tr>
<td>Facet Degeneration/arthropathy/hypertrophy/osteophytes/Fluid</td>
<td>none/mild</td>
</tr>
<tr>
<td>Listhesis: anterolisthesis/retrorlisthesis/blocked</td>
<td>1</td>
</tr>
<tr>
<td><strong>Comments:</strong></td>
<td></td>
</tr>
</tbody>
</table>

Abstraction database
**Wikipedia:**

*Crowdsourcing is…*

“the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community, rather than from traditional employees or suppliers.”
Methods – Crowdsourcing: What it’s not

Not to be confused with crowdfunding:

"Remember—we’re not begging. We’re crowdfunding."
“At a 1906 country fair in Plymouth, eight hundred people participated in a contest to estimate the weight of a slaughtered and dressed ox. Statistician Francis Galton observed that the median guess, 1207 pounds, was accurate within 1% of the true weight of 1198 pounds. This has contributed to the insight that a crowd's individual judgments can be modeled as a probability distribution of responses with the mean centered near the true mean of the quantity to be estimated.”


Amazon Mechanical Turk™, www.mturk.com

– “Artificial artificial intelligence”

– A marketplace for human intelligence tests (HITs)

– 500k+ crowd contributors, 190 countries

– 240k+ HITs available as of 10/6/2014

– HITs range: $0.01 to $56.10 --> Amazon account
“Turker” Demographics (Ross 2010)

- **Location**
  - 56% United States
  - 36% India (steadily increasing)
  - 8% Other

- **Gender** – 52% female

- **Age** – 67% are 18-35 years old

- **Income** – 46% make < $20k/year USD
  - 85% of Indian workers make < $20k/year

- **Education** – Bachelors or graduate degree
  - 55% of US workers
  - 66% of Indian workers
Methods – Crowdsourcing: Demographics

Limited Demographics

- My worker ID: A34ARR5E0Xxxxxxxxx
- Number of HITs submitted: 112
- Number of HITs accepted: 110 (98.2% approval)
- Location: United States
- Task specific: start / stop time
Methods – Crowdsourcing: Examples

Typical crowdsourcing activities:

• Fill out a survey
  • Opinions
  • Research (often in psychology)

• Translate a passage into a different language

• Search web for images

• Write a passage about a topic

• Tasks that a computer *could* do, e.g. natural language processing
Methods – Crowdsourcing: Examples

HITs: 137 hits I’m qualified for that pay ≥ $0.50

HITs that pay at least $0.50 for which you are qualified
1-10 of 137 Results

<table>
<thead>
<tr>
<th>Requester: Kris Brower</th>
<th>HIT Expiration Date: Feb 20, 2014 (14 weeks 1 day)</th>
<th>Reward: $0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Allotted: 4 hours</td>
<td>HITs Available: 678</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Requester: CrowdFlower</th>
<th>HIT Expiration Date: Nov 16, 2013 (3 days 22 hours)</th>
<th>Reward: $1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Allotted: 30 minutes</td>
<td>HITs Available: 580</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Requester: CrowdFlower</th>
<th>HIT Expiration Date: Nov 15, 2013 (3 days 2 hours)</th>
<th>Reward: $1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Allotted: 2 hours</td>
<td>HITs Available: 437</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Requester: CrowdFlower</th>
<th>HIT Expiration Date: Nov 19, 2013 (6 days 22 hours)</th>
<th>Reward: $0.55</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Allotted: 30 minutes</td>
<td>HITs Available: 387</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Requester: CrowdFlower</th>
<th>HIT Expiration Date: Nov 17, 2013 (4 days 14 hours)</th>
<th>Reward: $0.68</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Allotted: 30 minutes</td>
<td>HITs Available: 368</td>
<td></td>
</tr>
</tbody>
</table>

Background - Imaging Reports - BOLD/BOLDER - LIRE

Methods - Abstracters - Crowdsourcing - NLP - Study design

Results - Time - Cost - Accuracy

Conclusions
Methods – Crowdsourcing: Example

Example: Call customer service / rate ($1.50)

Call and rate our customer service department (Insurance). This HIT is periodically reposted. Please make sure you only complete this once. When in doubt, message us BEFORE submitting the HIT.

Call the number below and remain on the line for at least three minutes AFTER getting in touch with a representative. This is very important. Calls under three minutes cannot be paid. Callers from USA only.

Please show valid interest in the offer. We cannot rate the quality of our service if you show no interest in the product at all.

Again, minimum 3 minutes on the call. You cannot use VoIP, we will have no way of knowing you called.

Bonuses are available to those who take the call seriously.

Phone number: (888) 350-9334

Provide the city/state you called from below. Nothing else is needed, as calls are recorded for quality.

You must ACCEPT the HIT before you can submit the results.
Developed and pilot tested survey

- Stripped out identifiers (by hand) from each report
- Used Amazon’s built in survey function
- Based upon abstracter form
  - Split into three HITs: Xray, CT, MRI
- Added key words and synonyms
- Pilot tested with 10 imaging reports
  - Group examination of errors
  - Made wording refinements after pilot #1
  - Saw marginal improvements in error rate in pilot #2
Methods – Crowdsourcing: Screenshot

Created separate HITs for each imaging modality: 10-14 Y/N questions

1. Disc space height loss (or decreased, narrowing, degenerative disc disease):
   - Yes
   - No
2. Moderate or severe stenosis (or central canal stenosis):
   - Yes
   - No
3. Lateral recess or foraminal stenosis:
   - Yes
   - No
4. Moderate or severe facet degeneration (or facet arthropathy, hypertrophy, osteophytes, fluid):
   - Yes
   - No
5. Listhesis (or spondylolisthesis or anterolisthesis or retrolisthesis):
   - Yes
   - No
6. Severe scoliosis (or kyphoscoliosis):
   - Yes
   - No
7. Severe kyphosis (or kyphoscoliosis):
   - Yes
   - No
8. Moderate or severe hydrocephalus:
   - Yes
   - No
9. Compression fracture (or compression deformity, or vertebral body height loss):
   - Yes
   - No
10. Did you understand the instructions for this HIT?
    - Yes
    - No

Thanks for your help!

**HISTORY**: Left lower back pain **FINDINGS**: AP lateral coned-down lateral and lateral flexion and extension views. Diffuse demineralization. Mild to moderate vertebral compression fracture at T11; age and etiology undetermined. Mild vertebral compression fracture of L2; age and etiology undetermined. Grade 1 spondylolisthesis L4-L5 in flexion; partially reducing in extension. Disc space narrowing T11-T12; T12-L1, L1-L2, L4-L5, and L5-S1. Vacuum phenomena at L1-L2 and L5-S1. Mild to moderate thoracolumbar spondylosis. Prominent facet degeneration at multiple lumbar levels; most pronounced L4-5. SI Joints unremarkable. Calcification of the abdominal aorta. **IMPRESSION**: 1. Diffuse demineralization. 2. Vertebral compression fractures at T11 and L2; age and etiology undetermined. 3. Grade 1 spondylolisthesis L4-L5; most likely degenerative type. Evidence of instability L4-L5; see details above. 4. Definitive disc disease; spondylosis; and multilevel facet degeneration as described. 5. Aortic atherosclerosis.
Methods – Crowdsourcing: Screenshot

Background
- Imaging Reports
- BOLD/BOLDER
- LIRe

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions

** Created separate HITs for each imaging modality: 10-14 Y/N questions **

<table>
<thead>
<tr>
<th>Are any of the following findings noted for lumbar spine levels (L1, L2, L3, L4, or L5) in the imaging report?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Disc space height loss (or decreased, narrowing, degenerative disc disease):</td>
</tr>
<tr>
<td>- Yes</td>
</tr>
<tr>
<td>- No</td>
</tr>
<tr>
<td>2. Moderate or severe stenosis (or central canal stenosis):</td>
</tr>
<tr>
<td>- Yes</td>
</tr>
<tr>
<td>- No</td>
</tr>
<tr>
<td>3. Lateral recess or foraminal stenosis:</td>
</tr>
<tr>
<td>- Yes</td>
</tr>
<tr>
<td>- No</td>
</tr>
<tr>
<td>4. Moderate or severe facet degeneration (or facet arthropathy, hypertrophy, osteophytes, fluid):</td>
</tr>
<tr>
<td>- Yes</td>
</tr>
<tr>
<td>- No</td>
</tr>
<tr>
<td>5. Listhesis (or spondylolisthesis or anterolisthesis or retrolisthesis):</td>
</tr>
<tr>
<td>- Yes</td>
</tr>
<tr>
<td>- No</td>
</tr>
<tr>
<td>6. Severe scoliosis (or kyphoscoliosis):</td>
</tr>
<tr>
<td>- Yes</td>
</tr>
<tr>
<td>- No</td>
</tr>
<tr>
<td>7. Severe kyphosis (or kyphoscoliosis):</td>
</tr>
<tr>
<td>- Yes</td>
</tr>
<tr>
<td>- No</td>
</tr>
<tr>
<td>8. Moderate or severe hydrocephalus:</td>
</tr>
<tr>
<td>- Yes</td>
</tr>
<tr>
<td>- No</td>
</tr>
<tr>
<td>9. Compression fracture (or compression deformity, or vertebral body height loss):</td>
</tr>
<tr>
<td>- Yes</td>
</tr>
<tr>
<td>- No</td>
</tr>
</tbody>
</table>

** HISTORY **: Left lower back pain ** FINDINGS **: AP lateral coned-down lateral and lateral flexion and extension views. Diffuse demineralization. Mild to moderate vertebral compression fracture at T11; age and etiology undetermined. Mild vertebral compression fracture of L2; age and etiology undetermined. Grade 1 spondylolisthesis L4-L5 in flexion; partially reducing in extension. Disc space narrowing T11-T12; T12-L1; L1-L2; L4-L5; and L5-S1. Vacuum phenomena at L1-L2 and L5-S1. Mild to moderate thoracolumbar spondylosis. Prominent facet degeneration at multiple lumbar levels; most pronounced L4-5. SI Joints unremarkable. Calcification of the abdominal aorta. ** IMPRESSION **: 1. Diffuse demineralization. 2. Vertebral compression fractures at T11 and L2; age and etiology undetermined. 3. Grade 1 spondylolisthesis L4-L5; most likely degenerative type. Evidence of instability L4-L5; see details above. 4. Definitive disc disease; spondylosis; and multilevel facet degeneration as described. 5. Aortic atherosclerosis.

10. Did you understand the instructions for this HIT? |
| - Yes |
| - No |

Thanks for your help!
Methods – NLP: Approaches

Background
- Imaging Reports
- BOLD/BOLDER
- LIRE

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Rules-based NLP
*Def*: Set of deterministic rules by experts to map NLP components

- Advantages
  - Good solution if the test has simple rules
  - Relatively simple to code

- Disadvantages
  - May not generalize to new data (or new radiologist, system)
  - Negation is difficult

Statistical NLP
*Def*: Statistical relationships among NLP components

- Advantages
  - Good solution rules are complicated
  - Generalizable to different datasets

- Disadvantages
  - Need a lot of data (findings) to build models

Conclusions

---

Slide 42 of 58

Big Data vs. Big Crowds:
Getting Useful Data from Text Fields Using Large Data Sets

Bryan Comstock, bac4@uw.edu
Jerry Jarvik, jarvikj@uw.edu
## Methods – NLP: Tools

### Background
- Imaging Reports
- BOLD/BOLDER
- LIRE

### Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

### Results
- Time
- Cost
- Accuracy

### Conclusions

#### Software

- **Python**: PyEnchant spellcheck, text matching code
- **Natural Language ToolKit (NLTK)**: tagging, chunking
- **Stanford NLP**: Sentence segmenting
- **NegEx**: Negation detection
- **Lucene**: Reverse indexing system to recognize concepts
- **MALLET**: Statistical Machine Learning application to Natural Language

*A lot of software out there (mostly by computer scientists) is available for different goals*
Example text-based rules we looked for:

**Disc Space Height Loss:**
- “disc space height loss”
- “decreased disc space”
- “decreased disk space”
- “disc space narrowing”
- “degenerative disc”

**Compression Fracture:**
- “compression fracture”
- “compression deformity”
- “vertebral body height loss”
- “vertebral height loss”

**Moderate or Severe Disc Bulge:**
- “moderate disc bulge”
- “medium disc bulge”
- “severe disc bulge”
- “large disc bulge”
  (or bulging disc)

**Nerve Root Compression:**
- “nerve root compression”

Notes:
- Negation was not included.
- Conjoined words were split and text spellchecked
- Spelling variations of disc / disk
Other Study Design Parameters

Abstracters (thank you!)
- Reviewer 1: Sean Rundell, PT, DPT, PhD
- Reviewer 2: Nathan Dettori, MD
- Discrepancy adjudicator: Jerry Jarvik, MD MPH

Crowdsourcing
- Conducted by Bryan Comstock
- $0.40 for Xray, $0.60 for CT and MRI
- >18 years old, fluent in reading English
- From one of 10 countries with highest proportion with English as the primary language

NLP (thank you!)
- Katherine Tan, UW Biostatistics PhD student
- Patrick Heagerty, UW Biostatistics Professor/Chair
### Statistical Analyses

<table>
<thead>
<tr>
<th>Test</th>
<th>Gold Standard</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>A+C</td>
<td>B+D</td>
</tr>
</tbody>
</table>

**Accuracy** \[ \frac{(A + D)}{(A+B+C+D)} \]

- Fraction of text reports where NLP or crowd rating aligns with gold standard
Methods – Study Design

---

**Statistical Analyses**

<table>
<thead>
<tr>
<th>Test</th>
<th>Gold Standard</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>-</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>A+C</td>
<td>B+D</td>
</tr>
</tbody>
</table>

**Accuracy** \( \frac{(A + D)}{(A+B+C+D)} \)
- Fraction of text reports where NLP or crowd rating aligns with gold standard

**Sensitivity** \( \frac{A}{(A+C)} \)
- proportion with finding with (+) test
Methods – Study Design

Statistical Analyses

<table>
<thead>
<tr>
<th>Test</th>
<th>Gold Standard</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>A</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A+C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B+D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A+B+C+D</td>
</tr>
</tbody>
</table>

Accuracy \[ \frac{(A + D)}{(A+B+C+D)} \]
- Fraction of text reports where NLP or crowd rating aligns with gold standard

Sensitivity \[ \frac{A}{(A+C)} \]
= proportion with finding with (+) test

Specificity \[ \frac{D}{(B+D)} \]
= proportion without finding with (-) test

Bryan Comstock, bac4@uw.edu
Jerry Jarvik, jarvikj@uw.edu
Results

Background
- Imaging Reports
- BOLD/BOLDER
- LIRE

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions
Results

Study flow diagram

- 400 reports viewed by abstracters
- 386 reports evaluated by crowd and NLP*
- 361 reports of actual lumbar imaging

*used in analysis sample

- 14 reports insufficient or “comparison to older exam”

- 25 reports non-lumbar:
  - 8 pelvis
  - 6 head / neck
  - 4 breast
  - 3 abdomen
  - 2 lower extremity
  - 2 other

*used in analysis sample
Results – Time: Set-up

Abstracters: 5+ hours
- Form set-up and modification: 3+ hours
- Cross-training with senior reviewer: 2 hours
  Note: started with existing forms

Crowd: 10 hours
- Account set-up: 1 hour
- Text report de-identification: 3 hours
- Web-form set-up: 6 hours
  Note: started with moderate knowledge of mTurk

NLP: 9+ hours
- Coding: 5.5 hours
- Meetings with NLP expert: 3.5 hours
  Note: 9 months of part-time learning methods in R, Python, and NLP with faculty mentorship
Results – Time: Analysis-ready data set

Background
- Imaging Reports
- BOLD/BOLDER
- LIRE

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions

Abstracters: 69 hours (over weeks)
- Reviewers 1 and 2: 60 hours
- Data cleaning: 5 hours
- Gold standard adjudication: 4 hours

Crowd: 6h, 13min for 3,860 ratings (over a weekend)
- Xray: 2h 18m for 1,770 ratings (Friday afternoon)
- CT: 1h 24m for 500 ratings (Saturday morning)
- MRI: 2h 26m for 1,590 ratings (Sunday mid-day)
- 5 minutes to upload/download data from mTurk

NLP: 2 hours and 5 seconds
- Imaging text report cleaning: 2 hours
- Run code: 5 seconds
Results – Costs

Abstracters: $3,969.00 (person time)
- No costs beyond person-time

Crowd: $3,052.30 (person time + mTurk costs)
- Two pilot tests: $132.00
- Xray: $778.80 for 1,770 ratings @ $0.40/rating
- CT: $330.00 for 500 ratings @ $0.60/rating
- MRI: $1,049.50 for 1,590 ratings @ $0.60/rating
  => Median $/hr: MRI: $7.10, CT: $7.70, Xray: $7.66

NLP: $1,311.28 (person time)
- Using freeware, no costs beyond person-time
- Does not include 9+ months of part-time learning
Results – Accuracy: Prevalence

Prevalence of Findings

- Disc Height Loss
- Lischism
- Disc Desiccation
- Nerve Root Compression
- Facet Degeneration
- Spinal Stenosis
- Disc Protrusion
- Disc Herniation
- Compression Fracture
- Edema
- Scoliosis
- Disc Extrusion
- Kyphosis
### Example: Disc Height Loss

<table>
<thead>
<tr>
<th></th>
<th>Gold Standard</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Crowd</td>
<td>+</td>
<td>-</td>
<td>Row total</td>
</tr>
<tr>
<td></td>
<td>157</td>
<td>17</td>
<td>174</td>
</tr>
<tr>
<td></td>
<td>54</td>
<td>158</td>
<td>212</td>
</tr>
<tr>
<td></td>
<td>211</td>
<td>175</td>
<td>386</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Gold Standard</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP</td>
<td>+</td>
<td>-</td>
<td>Row total</td>
</tr>
<tr>
<td></td>
<td>132</td>
<td>25</td>
<td>157</td>
</tr>
<tr>
<td></td>
<td>79</td>
<td>150</td>
<td>229</td>
</tr>
<tr>
<td></td>
<td>211</td>
<td>175</td>
<td>386</td>
</tr>
</tbody>
</table>

*Crowd positive = 7-10 positive ratings*
## Results – Accuracy: Example

### Example: Disc Height Loss

<table>
<thead>
<tr>
<th></th>
<th>Gold Standard</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Crowd</td>
<td>+</td>
<td>-</td>
<td>Row total</td>
</tr>
<tr>
<td>+</td>
<td>157</td>
<td>17</td>
<td>174</td>
</tr>
<tr>
<td>-</td>
<td>54</td>
<td>158</td>
<td>212</td>
</tr>
<tr>
<td></td>
<td>211</td>
<td>175</td>
<td>386</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Gold Standard</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP</td>
<td>+</td>
<td>-</td>
<td>Row total</td>
</tr>
<tr>
<td>+</td>
<td>132</td>
<td>25</td>
<td>157</td>
</tr>
<tr>
<td>-</td>
<td>79</td>
<td>150</td>
<td>229</td>
</tr>
<tr>
<td></td>
<td>211</td>
<td>175</td>
<td>386</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>81.6%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>73.1%</td>
</tr>
</tbody>
</table>

*Crowd positive = 7-10 positive ratings*
## Results – Accuracy: Example

### Example: Disc Height Loss

<table>
<thead>
<tr>
<th>Crowd</th>
<th>Gold Standard</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>157</td>
<td>17</td>
</tr>
<tr>
<td>-</td>
<td>54</td>
<td>158</td>
</tr>
<tr>
<td></td>
<td>211</td>
<td>175</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NLP</th>
<th>Gold Standard</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>132</td>
<td>25</td>
</tr>
<tr>
<td>-</td>
<td>79</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>211</td>
<td>175</td>
</tr>
</tbody>
</table>

- **Accuracy**: 81.6%
- **Sensitivity**: 74.4%
- **Accuracy**: 73.1%
- **Sensitivity**: 62.6%

*Crowd positive = 7-10 positive ratings*
### Results – Accuracy: Example

#### Example: Disc Height Loss

<table>
<thead>
<tr>
<th></th>
<th>Gold Standard</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Crowd</td>
<td>+</td>
<td>-</td>
<td>Row total</td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>157</td>
<td>17</td>
<td>174</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>54</td>
<td>158</td>
<td>212</td>
<td></td>
</tr>
<tr>
<td></td>
<td>211</td>
<td>175</td>
<td>386</td>
<td></td>
</tr>
</tbody>
</table>

|                  | Gold Standard |                  |                  |                  |
| NLP              | +             | -                | Row total        |                  |
| +                | 132           | 25               | 157              |                  |
| -                | 79            | 150              | 229              |                  |
|                  | 211           | 175              | 386              |                  |

- **Accuracy**: 81.6%
- **Sensitivity**: 74.4%
- **Specificity**: 90.3%

#### Conclusions

- *Crowd positive = 7-10 positive ratings*
Results – Accuracy

Background
- Imaging Reports
- BOLD/BOLDER
- LIRE

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions

Overall Accuracy:
Crowd = 86.6%
NLP = 82.8%

Accuracy

Disc Height Loss
Liathesis
Disc Desiccation
Nerve Root Compression
Spinal Stenosis
Facet Degeneration
Disc Bulge
Spinal Stenosis
Disc Protrusion
Disc Herniation
Compression Fracture
Edema
Scoliosis
Disc Extrusion
Kyphosis

Big Data vs. Big Crowds:
Getting Useful Data from Text Fields Using Large Data Sets

Bryan Comstock, bac4@uw.edu
Jerry Jarvik, jarvikj@uw.edu
Results – Accuracy: Sensitivity (Crowd 7+)

Background
- Imaging Reports
- BOLD/BOLDER
- LIRE

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions

Big Data vs. Big Crowds:
Getting Useful Data from Text Fields Using Large Data Sets

Bryan Comstock, bac4@uw.edu
Jerry Jarvik, jarvikj@uw.edu
Results – Accuracy: Sensitivity (Crowd 5+)

Background
- Imaging Reports
- BOLD/BOLDER
- LIRE

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions

Big Data vs. Big Crowds:
Getting Useful Data from Text Fields Using Large Data Sets

Bryan Comstock, bac4@uw.edu
Jerry Jarvik, jarvikj@uw.edu
Results – Accuracy: Specificity (Crowd 7+)

Background
- Imaging Reports
- BOLD/BOLDER
- LIRE

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions

Slide 62 of 58
Results – Accuracy: Specificity (Crowd 5+)

Specificity

- Disk Height Loss
- Listhesis
- Disc Desiccation
- Nerve Root Compression
- Facet Degeneration
- Disc Bulge
- Spinal Stenosis
- Disc Protrusion
- Disc Herniation
- Compression Fracture
- Edema
- Scoliosis
- Disc Extrusion
- Kyphosis

Crowd vs. NLP Specificity

- 100%
- 90%
- 80%
- 70%
- 60%
- 50%
- 40%
- 30%
- 20%
- 10%
- 0%

Background
- Imaging Reports
- BOLD/BOLDER
- LIRE

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions
Conclusions

Background
- Imaging Reports
- BOLD/BOLDER
- LIRE

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions
Conclusions: Time

Background
- Imaging Reports
- BOLD/BOLDER
- LIRE

Methods
- Abstracters
- Crowdsourcing
- NLP
- Study design

Results
- Time
- Cost
- Accuracy

Conclusions

Time

- Abstracters:
  - Set up time is low
  - Time to analysis ready data set
    - subject to availability of abstracters
    - time increases with # of ratings

- Crowdsourcing:
  - Set up time is low to moderate
  - Time to analysis ready data set
    - can turn on and let ‘run’
    - time increases with # of ratings

- NLP:
  - Initial set up time is high
  - Time to analysis ready data set: instant
Conclusions: Costs

Costs

- Abstracters:
  - Costs function of
    - Number of abstracters (and $/hr)
    - Number of ratings

- Crowdsourcing:
  - Costs function of
    - Number of HITs
    - Number of ratings per HIT (and $/HIT)

- NLP:
  - Costs function of
    - Person time, knowledge
Conclusions: Accuracy

Accuracy

- Abstracters:
  - 9% of findings had discordant ratings on more than half of the imaging reports
  - Note: concordant ratings were not reviewed

- NLP:
  - 82.8% accurate against gold standard
  - Improved with negation, statistical NLP

- Crowdsourcing:
  - 86.6% accurate against gold standard
  - At least as accurate as NLP for 13 / 15 findings
  - Improved with more instruction
    - “broad-based”, “diffuse”
Concluding Thoughts

- This is a challenging problem
- Accuracy of crowdsourcing and NLP depends on input
- The crowd performed at least as well as a rules-based NLP algorithm in classifying against a gold standard, but both need improvement in sensitivity
- Projects considering human abstracters might consider using crowdsourcing as an alternative
- Large scale projects (like LIRE) need to use NLP due to cost and time constraints
Thank you!

Questions / comments / suggestions
Submitting Questions

To submit a question:

1. Click in the Q&A box on the left side of your screen
2. Type your question into the dialog box and click the Send button
Upcoming EDM Forum Webinar

→ Improving Understanding of Stress Exposure in Minority Communities with Asthma: Using Geographic Information Systems (GIS) Data Collection Methods
  – Tuesday, October 21, 2014
  – 1:00 - 2:30 p.m. EDT
  – Email
    nina.bhattacharyya@academyhealth.org if interested
Thank You

Please take a moment to fill out the brief evaluation which will appear in your browser.