Health Informatics

Lecture 8

Samantha Kleinberg
samantha.kleinberg@stevens.edu

Go over midterm

- How do we know which EHR interface is better?
- Does an NLP method get all mentions of uncertainty or diabetes in a dataset?

Evaluation methods

- Quantitative
  - Compare to Gold standard
  - Classification/prediction
  - Evaluate cost, time
- Qualitative
  - Compare to existing literature
  - Usability

- Replication

Gold standard

- Create labels for data
- Evaluate based on how many items found
- Example: have MDs mark Framingham symptoms in text, evaluate what percentage NLP method can find

- Challenges?
  - Inter-rater reliability
  - Cost
  - Evaluating precision

Inter-rater reliability

How well do annotators agree?

E.g. do clinicians agree on which records mention diabetes?

What's difference delay between two people annotating activities? Does it differ significantly?

How to annotate?

- Domain experts
  - + Expertise
  - + Can help develop annotation scheme
  - – $$$, time
  - – Smaller pool of annotators
- Crowdsourcing
  - + Large pool of potential annotators
  - + Cheaper
  - – Need training
  - – May try to game system
  - – Privacy issues

Crowdsourcing

Journal List › J Med Internet Res › v.15(4); Apr 2013 › PMC3636329



Journal of Medical Internet Research
The leading peer-reviewed eHealth journal
Current Issue Submit Membership Editorial Board

J Med Internet Res. 2013 April; 15(4): e73.
Published online 2013 April 2. doi: 10.2196/jmir.2426
PMCID: PMC3636329

Web 2.0-Based Crowdsourcing for High-Quality Gold Standard Development in Clinical Natural Language Processing

Monitoring Editor: Gunther Eysenbach

Reviewed by Miguel Luengo-Oroz, Hua Xu, and Lynette Hirschman

Haijun Zhai, PhD,[#1] Todd Lingren, MA,[#1] Louise Deleger, PhD,[1] Qi Li, PhD,[1] Megan Kaiser, BA,[1] Laura Stoutenborough, BSN,[1] and Imre Solti, MD, PhD[✉1]

Author information ▶ Article notes ▶ Copyright and License information ▶

This article has been cited by other articles in PMC.

J Med Internet Res

Exclusion Criteria:

- treatment with NSAIDs, clopidogrel, ticlopidine, dipyridamole, warfarin or any other drugs known to affect platelet function.

- ischemic vascular event within the previous 12 months

- revascularization (angioplasty or coronary by-pass graft surgery) within the previous 12 months

- intake of NSAIDs within 1 week of myocardial infarction (group: "Previous myocardial infarction").

- platelet count

- previous myocardial infarction (group: "CAD").

- not able to give informed consent

gender: Both

minimum_age: 18 Years

maximum_age: N/A

healthy_volunteers: Accepts Healthy Volunteers
mesh_term: Coronary Artery Disease,Diabetes Mellitus,Aspirin

Extend Highlight | Shrink Highlight

Annotated entity list:	
Medication Name	**Medication Type**
clopidogrel	NSAIDs
ticlopidine	drugs
dipyridamole	----
warfarin	----
Aspirin	----

Medication Name
Medication Type

lication types are
chema of the

ing medications
the linking task:
sentence, “Advair
1, tablet, and
tributes of Advair,
igure 1.

w, topical)
t (eg, active,
ge)
conditional
eg,

Results

- Agreement between crowd/expert

  (1) annotations (0.87, F-measure for medication names; 0.73, medication types)

  (2) correction of previous annotations (0.90, medication names; 0.76, medication types)

  (3) linking medications with their attributes (0.96).

  No statistically significant difference between the crowd and traditionally-generated corpora.

Classification, prediction

- Classification: train model to determine who has CHF. Apply to new data to separate CHF/non-CHF.
- Prediction: Use past data, predict who will develop CHF, and follow cohort.

- Challenges?
  - Labels aren't certain
  - Getting labels can be difficult, expensive
  - May have to follow cohort over a long time
  - Predictions are indirect measure, doesn't ensure causality

Cost/time

- Does new method reduce readmissions to hospital?
- Did expenditures go down?
- Are orders slowed down by new decision support?

Replication

- Compare against prior method
- Compare findings in two populations

- Recall earlier discussion paper!

Literature review

- Do findings conflict with prior knowledge?
- Are there known mechanisms that support a hypothesis?

- Problems:
  - Can find supporting results for almost any finding
  - What if it's truly new?

Usability

- Metrics from HCI
  - Surveys, questionnaires, interviews
  - Lab vs contextual
  - Focus groups?
  - Screen capture
  - Eye tracking

Evaluation strategy depends on..

- What you're evaluating
  - NLP, decision support, new interface?
- Goal
  - Evaluate usability vs. annotation accuracy
- Constraints
  - Time, money, people
  - In-depth study of small group vs. superficial test with large population

Table 1 – The STARE-HI principles: items recommended to be included in Health Informatics evaluation reports.

Item #	Item
1	Title
2	Abstract
3	Keywords
4	Introduction
4.1	Scientific background
4.2	Rationale for the study
4.3	Objectives of study
5	Study context
5.1	Organizational setting
5.2	System details and system in use
6	Methods
6.1	Study design
6.2	Theoretical background
6.3	Participants
6.4	Study flow
6.5	Outcome measures or ev
6.6	Methods for data acquisi
6.7	Methods for data analysi
7	Results
7.1	Demographic and other s
7.2	Unexpected events durin
7.3	Study findings and outco
7.4	Unexpected observations
8	Discussion
8.1	Answers to study questions
8.2	Strengths and weaknesses of the study
8.3	Results in relation to other studies
8.4	Meaning and generalisability of the study
8.5	Unanswered and new questions
9	Conclusion
10	Authors' contribution
11	Competing interests
12	Acknowledgement
13	References
14	Appendices

Case study

A large hospital wanted to find out how many of their patients with hypertension have their blood pressure (BP) under control.

To do this, they searched their EHR for patients with ICD9 codes for hypertension, and then did a search of text and database fields for words confirming the diagnosis and the BP values. Patients where this data could not be found were reviewed manually.

To validate their results, they examined a random sample of patients whose BP was found to be in control to see if this was true.

Study methods

- Observational
  - Cross-sectional
  - Longitudinal cohort study
  - Case-control
- Intervention
  - Clinical trials

Is Health IT improving outcomes

Say we observe…

- Reduced readmissions
- Fewer errors
- Less time spent on notes

among hospitals using electronic records

How do we know if it's a result of EHR use?

What causes intervention vs what causes effect?

- More skilled doctors may choose a particular procedure
- Healthier patients may use an app

Comparing two fitness trackers.

Can assign half participants to one and half to the other, but is there a better approach?

- Subjective assessment
- Delayed effects (side effects)
- Intervention indirectly causing effect

RCTs

- Two groups identical in all respects except for treatment
- Differences between them should then be due to treatment

James Lind and Scurvy

- Divided 12 sailors into groups of two
- Each pair received an addition to their diet
- Citrus led to improved symptoms

But didn’t randomly assign the treatments

- Individual
- Group-level/cluster
  - Consider true sample size

Who to randomize?

Testing heartburn medication…

Should population be…?

-Everyone

-People with history of heartburn

-People w/heartburn and not taking drugs that may interact w/proposed one

-People w/o heartburn due to another condition

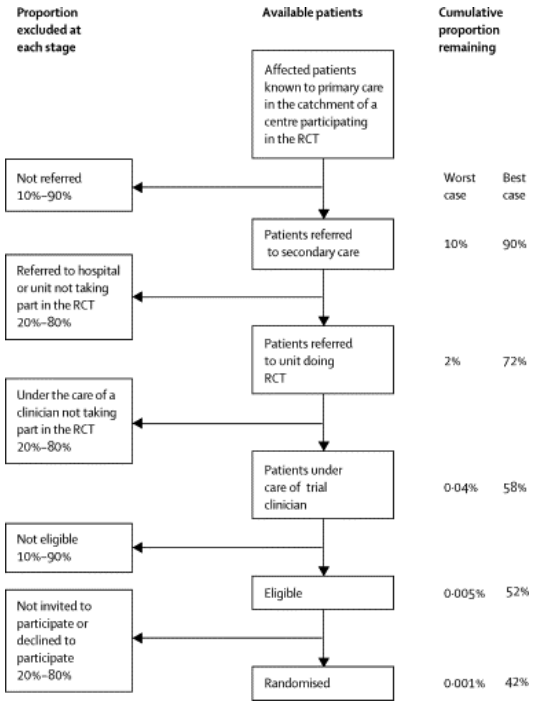


Figure 4. Schematic diagram illustrating effect of multiple stages of selection inherent in clinical practice on proportion of patients in catchment of participating centre entered into an RCT done in secondary careWorst case assumes proportion of patients exc...

Peter M Rothwell

External validity of randomised controlled trials: "To whom do the results of this trial apply?"

null, Volume 365, Issue 9453, 2005, 82–93

http://dx.doi.org/10.1016/S0140-6736(04)17670-8

- Randomizing to avoid selection bias but…

Who to analyze?

- People not available for all data collection on outcomes = lost to follow-up

- Ignore subjects with missing data?
- Use all data?

Survival bias

Example: analyze habits of people who've lost weight and kept it off for 2 years.

What's wrong with this?

To assign treatments we could…

-Have researcher alternate between treatment/control for each person

-Flip a coin each time person joins study

-Enclose assignment in set of envelopes that researcher only opens after each subject is enrolled

What are we controlling?

1st RCT compared bed rest (then current standard of care) and streptomycin for treating tuberculosis

Why not compare streptomycin against nothing?

Why placebo

Act of treatment (even with no active ingredient) can lead to real change in outcomes

Just because something changes after treatment, doesn't mean it's because of the treatment!

Decisions aren't treat/don't treat, but WHICH treatment

Control ethics

- If effective treatment exists, should compare against that, not placebo
- Principle: we should in theory not know whether one treatment is better than the other.
- If we know a treatment exists and we withhold it, that's unethical

Control and bias

- Ever see commercials for diet programs or supplements?
  - Just shows A diet better than none, not that this particular one is better than any other
- Remember, choosing population and HOW the other treatment is given (dosage, etc)

Choosing a control

- Mimic the process as much as possible

Examples

-sugar pill

-sham acupuncture

-fake surgery

-active placebos

Placebo effect

- Dosage
- Injection vs pill
- Color of pills
- Packaging/branding
- Surgeries

THE LANCET

Volume 299, Issue 7763, 10 June 1972, Pages 1279–1282

Originally published as Volume 1, Issue 7763



Medical Education

DEMONSTRATION TO MEDICAL STUDENTS OF PLACEBO RESPONSES AND NON-DRUG FACTORS

Barry Blackwell, SaulS. Bloomfield, C.Ralph Buncher

Show more

Abstract

A class experiment for medical students ***Summary*** was devised to demonstrate the influence of the placebo effect and non-drug factors on response to drugs. The subjects were conditioned to expect sedative or stimulant effects, but all received placebo in one or two blue or pink capsules. Predictions about the size and nature of the placebo response and influence of the non-drug factors were made before the experiment and discussed afterwards. Four of six predictions were fully confirmed. Drug- associated changes were reported by 30% of the subjects and were severe in 1 or 2 individuals. Two capsules produced more noticeable changes than one, and blue capsules were associated with more sedative effects than pink capsules. Students rated the experiment highly both as a learning experience and for its relevance to their future practice of medicine.

Placebos without Deception: A Randomized Controlled Trial in Irritable Bowel Syndrome

Ted J. Kaptchuk[1,2]*, Elizabeth Friedlander[1], John M. Kelley[3,4], M. Norma Sanchez[1], Efi Kokkotou[1], Joyce P. Singer[2], Magda Kowalczykowski[1], Franklin G. Miller[5], Irving Kirsch[6], Anthony J. Lembo[1]

1 Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, United States of America, **2** Osher Research Center, Harvard Medical School, Boston, Massachusetts, United States of America, **3** Psychology Department, Endicott College, Beverly, Massachusetts, United States of America, **4** Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, United States of America, **5** Department of Bioethics, National Institutes of Health, Bethesda, Maryland, United States of America, **6** Department of Psychology, University of Hull, Hull, United Kingdom

Abstract

Background: Placebo treatment can significantly influence subjective symptoms. However, it is widely believed that response to placebo requires concealment or deception. We tested whether open-label placebo (non-deceptive and non-concealed administration) is superior to a no-treatment control with matched patient-provider interactions in the treatment of irritable bowel syndrome (IBS).

Methods: Two-group, randomized, controlled three week trial (August 2009-April 2010) conducted at a single academic center, involving 80 primarily female (70%) patients, mean age 47±18 with IBS diagnosed by Rome III criteria and with a score ≥150 on the IBS Symptom Severity Scale (IBS-SSS). Patients were randomized to either open-label placebo pills presented as "placebo pills made of an inert substance, like sugar pills, that have been shown in clinical studies to produce significant improvement in IBS symptoms through mind-body self-healing processes" or no-treatment controls with the same quality of interaction with providers. The primary outcome was IBS Global Improvement Scale (IBS-GIS). Secondary measures were IBS Symptom Severity Scale (IBS-SSS), IBS Adequate Relief (IBS-AR) and IBS Quality of Life (IBS-QoL).

Findings: Open-label placebo produced significantly higher mean (±SD) global improvement scores (IBS-GIS) at both 11-day midpoint (5.2±1.0 vs. 4.0±1.1, $p<.001$) and at 21-day endpoint (5.0±1.5 vs. 3.9±1.3, $p=.002$). Significant results were also observed at both time points for reduced symptom severity (IBS-SSS, $p=.008$ and $p=.03$) and adequate relief (IBS-AR, $p=.02$ and $p=.03$); and a trend favoring open-label placebo was observed for quality of life (IBS-QoL) at the 21-day endpoint ($p=.08$).

Blinding

- Single: Patient unaware of what treatment being received
- Double: Patient + clinician unaware of what treatment being administered/received
- Triple: Double + person analyzing results not aware of which group is which

Example

RCT of two treatments + placebo for multiple sclerosis

Examination by blinded and unblinded neurologists

Unblinded Neurologists showed one treatment beneficial at 6, 12, 24 months, p-value<0.05

Noseworthy, J. H., Ebers, G. C., Vandervoort, M. K., Farquhar, R. E., Yetisir, E., & Roberts, R. (1994). The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology*, *44*(1), 16-16.

Using RCT results

- Tension between study adequately testing hypothesis… and broader applicability

- Study tests one intervention, but in reality do we only change one thing?
- Homogeneous study population isolates cause, but what happens when we move to more varied setting?
- RCT tells us treatment CAN cause something, not that it WILL when we try to use it

Key questions

- Context
  - Are needed features present?
- Will treatment work the same way in another place?
  - Ex: incentives, mosquito nets/fishing
- Side effects
- Interactions
  - Factors that will undermine the cause?

- Intervention: office chairs
- Outcome: weight loss



Weight loss in first test!

Failure in new population. How?

Why Olanzapine Beats Risperidone, Risperidone Beats Quetiapine, and Quetiapine Beats Olanzapine: An Exploratory Analysis of Head-to-Head Comparison Studies of Second-Generation Antipsychotics

Stephan Heres, M.D.

John Davis, M.D.

Katja Maino, M.D.

Elisabeth Jetzinger, M.D.

Werner Kissling, M.D.

Stefan Leucht, M.D.

Objective: In many parts of the world, second-generation antipsychotics have largely replaced typical antipsychotics as the treatment of choice for schizophrenia. Consequently, trials comparing two drugs of this class—so-called head-to-head studies—are gaining in relevance. The authors reviewed results of head-to-head studies of second-generation antipsychotics funded by pharmaceutical companies to determine if a relationship existed between the sponsor of the trial and the drug favored in the study's overall outcome.

Method: The authors identified head-to-head comparison studies of second-generation antipsychotics through a MEDLINE

sources of bias that could have affected the results in favor of the sponsor's drug.

Results: Of the 42 reports identified by the authors, 33 were sponsored by a pharmaceutical company. In 90.0% of the studies, the reported overall outcome was in favor of the sponsor's drug. This pattern resulted in contradictory conclusions across studies when the findings of studies of the same drugs but with different sponsors were compared. Potential sources of bias occurred in the areas of doses and dose escalation, study entry criteria and study populations, statistics and methods, and reporting of results and wording of findings.

n=1

How can we figure out which of two interventions is best for one individual?

Instead of randomizing people, randomize order of treatment

- A-B?
  - No statistical significance
- A-B-A-B?
  - What if condition improves always over time?
- Random?
  - Could get unlucky and can't guarantee half/half
- Randomize pairs
  - Again, what if ABAB?

Considerations

- How many treatment periods?
- How to randomize order?
- Washout period?

- Inappropriate surrogate outcome measures, so that what is said to have been measured is in fact not (see Gøtzsche et al. 1996; Jaeschke and Sackett 198
- Incomplete curate statistical analysis (Fei Medicine Working Gr
- Problems w ulting in inade- quate contr on 1998);
- Inability to biguous objec- tives or failu aeschke and Sackett 1989);
- False negatives, resulting from small numbers or insensitive outcome measures (Jaeschke and Sackett 1989).

JASON GROSSMAN AND FIONA J. MACKENZIE

To argue from the fact that RCTs have certain advantages, other things being equal, to the claim that the RCT is a gold standard, is like arguing that since being tall makes for a good high-jumper, it follows that a 6' elderly drunkard with a spinal injury is bound to be a better high-jumper than a 5'11" Olympic athlete.

What do we need to know before we can use the results of an RCT?

- Study is internally valid
  - i.e. it can answer the question it aims to answer
- Factors affecting external validity
  - Characteristics of setting
  - Selection of patients
  - Characteristics of patients
  - Follow-up
- Control?
- Blinding – single, double, triple

See also: Rothwell, P. M. (2006). Factors that can affect the external validity of randomised controlled trials. *PLoS Clin Trials*, *1*(1), e9.

Next week

- See syllabus for reading!