Using machine learning and crowdsourcing for the identification of studies of diagnostic test accuracy

AH Noel-Storr, J Thomas, I Marshall, B Wallace
Information specialist for Cochrane Dementia and Cognitive Improvement Group

Co-lead on Cochrane Crowd (a sub-project of Cochrane’s Project Transform)
Identifying studies

Why so challenging?
Identifying studies

- Poor reporting
- Poor indexing
- Increasing volume
Identifying studies: current model
Identifying studies: current model

2 step process
Identifying studies: current model

2 step process

Step 1
Design and run complex, sensitive, filter-free searches across multiple databases (aka the search)
Identifying studies: current model

2 step process

Step 1
Design and run complex, sensitive, filter-free searches across multiple databases

Step 2
Download, de-duplicate and screen what’s left at title and abstract

(aka the search)

(aka the screen)
Identifying studies: current model

2 step process

Step 1
Design and run complex, sensitive, filter-free searches across multiple databases

Step 2
Download, de-duplicate and screen what’s left at title and abstract

(aka the search)

(aka the screen)
Identifying studies: current model

DTA reviews in 2016

44% identified more than 20,000 citations to screen
How long does it take to screen 20,000 search results for a DTA review?
Identifying studies: current model

How long does it take to screen 20,000 search results for a DTA review?

20,000 x 2 = 40,000 to screen
How long does it take to screen 20,000 search results for a DTA review?

20,000 x 2 = 40,000 to screen
40,000 x 30 = 1,200,000 seconds
Identifying studies: current model

How long does it take to screen 20,000 search results for a DTA review?

20,000 x 2 = 40,000 to screen

40,000 x 30 = 1,200,000 seconds

1,200,000/60 = 20,000 minutes
How long does it take to screen 20,000 search results for a DTA review?

20,000 \times 2 = 40,000 \text{ to screen}

40,000 \times 30 = 1,200,000 \text{ seconds}

1,200,000/60 = 20,000 \text{ minutes}

20,000/60 = \textbf{333 hours!!}
How long does it take to screen 20,000 search results for a DTA review?

\[
20,000 \times 2 = 40,000 \text{ to screen}
\]
\[
40,000 \times 30 = 1,200,000 \text{ seconds}
\]
\[
1,200,000/60 = 20,000 \text{ minutes}
\]
\[
20,000/60 = 333 \text{ hours}!!
\]
Current challenges: what can we do?

Both approaches showing big potential in other domains. Could they work well for DTA?
Current challenges: what can we do?

Machine learning a method of data analysis where the machine builds ‘models’ based on training data.
Current challenges: what can we do?

Crowdsourcing is the process of “obtaining needed services, ideas or content by soliciting contributions from a large group of people, especially an online community”
Embase project and Cochrane Crowd

A large batch of records identified from a very sensitive search for RCTs

“the distribution of small parts of a problem”
Cochrane Crowd: progress so far

Snapshot: 17 July 2016

851,430 classifications
324,185 citations
26,258 RCTs
2,985 Signed up

This includes all the data since the start of the Embase project
Evidence Pipeline + Cochrane Crowd

Bringing the machine into the record workflow

How?

RCT  RCT  RCT
Current challenges: what can we do?

Pilot study 1 & 2
Pilot study 3

Could these methods work well for DTA?
Current challenges: what can we do?

For both studies we needed a gold standard dataset
Current challenges: what can we do?

A set of ‘known’ records

For both studies we needed a gold standard dataset
Current challenges: what can we do?

For both studies we needed a gold standard dataset

A set of ‘known’ records

- Known ‘good’ records (DTA studies)
Current challenges: what can we do?

For both studies we needed a gold standard dataset

- Known ‘good’ records (DTA studies)
- Known ‘reject’ records (not DTA studies)
Current challenges: what can we do?

A set of ‘known’ records

- Known ‘good’ records (DTA studies)
- Known ‘reject’ records (not DTA studies)
- Decent size

Wanted the dataset to be ‘generic’
Current challenges: what can we do?

In Medline (Ovid SP)

1. exp "sensitivity and specificity"
2. Diagnos*.ti,ab.
3. di.fs.
4. du.fs.
5. predict*.ti,ab.
6. accura*.ti,ab.
7. or/1-6
8. 201501*.ed.
9. 7 and 8

Yes, I used a filter
Current challenges: what can we do?

In Medline (Ovid SP)

1. exp "sensitivity and specificity"/
2. Diagnos*.ti,ab.
3. di.fs.
4. du.fs.
5. predict*.ti,ab.
6. accura*.ti,ab.
7. or/1-6
8. 201501*.ed.
9. 7 and 8

Hits: 20,623
Current challenges: what can we do?

Out of 1000 records:
- 47 were reporting diagnostic test accuracy studies
- 953 were ‘realistic’ reject records

Took a random sample of 1000 and screened them
Current challenges: what can we do?

Yes, I enriched the batch!
And ended up with a test dataset of 167 DTAs and 953 Rejects
How did the machine get on?

Pilot study 1

- Using the Gold standard dataset and doing a five-fold cross-validation, the binary classifier achieved 95% sensitivity, 40% specificity.

- 100% sensitivity could be achieved but specificity falls to 13%.
How did the machine get on?

**Pilot study 2**

- Trained on 70% and tested on 30% of the gold standard dataset

- An active learning simulation where abstracts were prioritised as a function of their predicted probability of relevance
How did the machine get on?

Pilot study 2

- N = 336 (30% of batch)
- 50 reports of DTAs
- Machine gave 41 of them a greater than 50% likelihood of being a DTA (FN)
- Machine gave 8 Rejects a greater than 50% likelihood of being a DTA (FP)
How did the Crowd get on?

Who were the Crowd?
- **5 known members** of the current Cochrane Crowd community with an interest in DTA

- **45 randomly chosen members** of the Crowd who may or may not be familiar with DTA but who had screened over 500 records for the RCT ID task (so familiar with the principle and likely to be up for giving this beta task a go)
How did the Crowd get on?

Comparison of Laser Scanning Diagnostic Devices for Early Glaucoma Detection. [201535]

Background: To compare the diagnostic accuracy and to evaluate the correlation of optic nerve head and retinal nerve fiber layer thickness values between Fourier-Domain optical coherence tomography (FD-OCT), confocal scanning laser ophthalmoscopy (CSLO), and scanning laser polarimetry (SLP) for early glaucoma detection. Patients and Methods: Ninety-three patients with early open-angle glaucoma, 58 patients with ocular hypertension, and 60 healthy control subjects were included in this observational, cross-sectional study. All study participants underwent FD-OCT (RTVue-100), CSLO (HRT3), and SLP (GDx VCC) imaging of the optic nerve head and the retinal nerve fiber layer. Area under the receiver operating characteristic curves (AUROC) and Bland-Altman analysis were performed. Results: The parameters with the highest diagnostic accuracy were found for FD-OCT cup-to-disc ratio (AUROC=0.841), for SLP NFI (AUROC=0.835), and for CSLO cup-to-disc ratio (AUROC=0.789). Diagnostic accuracy of the best CSLO and SLP parameter was similar (P=0.259). There was a small statistically significant difference between the best CSLO and FD-OCT parameters for differentiating between glaucoma and healthy eyes (P=0.047). Conclusions: FD-OCT and SLP have a similarly good diagnostic ability to distinguish between early glaucoma and healthy subjects. The diagnostic accuracy of CSLO was comparable with SLP and marginally lower compared with FD-OCT.

We agree!

This is a diagnostic test accuracy study. Unlike the previous record the study design is not an RCT. It is described as an observational, cross-sectional study. This study involved selecting a group of 'cases' - those already known to have the target condition (glaucoma) and those known to be without the condition - 'controls'. Lots of DTA studies adopt this case-control design, usually in the early stages of an index test's development. Now, what do you make of the next record...

Mandatory training: 20 practice records describing the main features to look out for when spotting DTAs
Cochrane Crowd: algorithm

DTA → DTA → DTA
Reject → Reject → Reject
Reject → DTA
Unsure → RESOLVER

It’s IN
‘BIN’
RESOLVER
How did the Crowd get on?

20 out of 50 invitees have completed training and done some screening.

890 individual assessments made = 237 records ‘completed’

After 2 weeks
How did the Crowd get on?

What classifications are people making?

If everyone says **Unsure**, we have a problem
(outcome: system efficiency)

If everyone **disagrees**, we have a problem
(outcome: system efficiency)

If everyone gets it **wrong**, we have a problem
(outcome: system accuracy)

After 2 weeks
How did the Crowd get on?

890 classifications

- Agreement classifications: 79.9% (711)
- Disagreement classifications: 9.9% (88)
- Unsure classifications: 10.2% (91)

What classifications are people making?
How did the Crowd get on?

Interim results: two weeks since launched; 20% complete
92% sensitivity, 97% specificity

<table>
<thead>
<tr>
<th></th>
<th>True positives</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crowd</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>Info specialist with extensive experience of screening for DTA (me)</td>
<td>True negatives</td>
<td>False negatives</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>120</td>
</tr>
</tbody>
</table>
What next?

It’s a doable task; we need to:

Complete pilot
Evaluate
Iterate
Scale
What next?

- We have a technical infrastructure to support this
  - We have a successful model set up
    - We have a Crowd
  - We have an expert DTA community (you!)
In summary

We set out to consider the **feasibility** of using **machine learning** or **crowdsourcing** to successfully identify studies of **diagnostic test accuracy**.
In summary

We set out to consider the **feasibility** of using **machine learning** or **crowdsourcing** to successfully identify studies of **diagnostic test accuracy**

Cochrane’s Embase project had shown that it is feasible to crowdsource the identification of RCTs and that the Crowd data could then be used to train the machine to help perform the task
In summary

We set out to consider the **feasibility** of using **machine learning** or **crowdsourcing** to successfully identify studies of **diagnostic test accuracy**.

Cochrane’s Embase project had shown that it is feasible to crowdsource the identification of RCTs and that the Crowd data could then be used to train the machine to help perform the task.

This work shows that **both the machine and the crowd have a role to play in the identification of DTA studies**. The challenge will be in working out just what that partnership should look like.
In summary

“Great things are done by a series of small things brought together.”

Vincent van Gogh
Thank you to all our contributors

crowd.cochrane.org
crowd@cochrane.org