Not just a tick-box exercise:

Three computational techniques and one technology to launch patient-centred assessments into the big data era

Chris Gibbons
NIHR Research Fellow | Cambridge School of Clinical Medicine
Director of Health Assessment and Innovation | Psychometrics Centre
• Introduce patient-reported data

• Strengths, weaknesses, opportunities and threats of big data and machine learning

• Machine learning, patient-reported outcomes, and health services research

• Three computational techniques to transform patient-centred big data
  - Computer adaptive testing
  - Psychometric predictions of answers to unseen items
  - Automated open-text analysis

• One technology to assist development and implementation of these tools
  - Concerto

• Where might we be heading?
PATIENT-REPORTED DATA

• Patient-reported outcome and experience measures

• Thousands, suitable for use in diverse conditions

• Many applications (trials, clinical practice, epidemiology, audit, quality, improvement)

• Other forms of patient-reported data – open text comments, social media (?)
PATIENT-REPORTED DATA

The good…

• *Can be an effective clinical intervention* to improve processes and outcomes of care (Valderas, 2008)

• *Patients favour it* and see it as evidence that their doctor is taking their issues seriously (Dowrick, 2008)

• Can be used for *comparative research and clinical trials*

• “Measure once, cut twice”
Routine provision of information on patient-reported outcome measures to healthcare providers and patients in clinical practice (Protocol)

PATIENT-REPORTED DATA

The bad…

• Ad hoc use (motivated clinicians and providers)
• Take a long time to administer, score and interpret
• Scores can be difficult to understand
• Unclear/low incentive for stakeholders
PATIENT-REPORTED DATA

The ugly…

• Risk of gamification by patients or providers

• Harming patients by exposure to worrying symptoms

• Response bias which protects underperforming providers

• Selection of politically convenient measures

• What is measured improves (but what about what is not?)
What works for clinicians?

• Specific PROM training/supported implementation

• Timely feedback
• Well aligned with practice
• Information integrated into available systems
• Information that is linked to specific clinical action

• Interventions with a formal, structured, feedback process perform better (Krageloh, 2014)
IMPROVING PATIENT-REPORTED DATA

• Tackle the barriers!
• Make measurement more efficient
• Engage patients and clinicians with instant feedback
• Align feedback more closely with care
• Address response biases
Hypothesis

Computational techniques could make patient-reported outcome and experience measures less time-consuming while making them more accurate, relevant, useful, and interesting.
(SELECTIVE) HISTORY

- 2200 BC
- 1886
- 1950
- 1960
- 2007
- 2011
Cattell sets up the first psychometric laboratory in the Cavendish at Cambridge University
(SELECTIVE) HISTORY

- 2200 BC
- 1886 Cavendish Lab
- 1950 First PROMS for measuring psychiatric diagnosis
- 1960
- 2007
- 2011
(SELECTIVE) HISTORY

- 2200 BC
- 1886: Cavendish Lab
- 1950: PROMS
- 1960
- 2007
- 2011

Paradigm shift in psychometric science with item response theory (IRT)
(SELECTIVE) HISTORY

- 2200 BC
- 1886: Cavendish Lab
- 1950: PROMS
- 1960: IRT
- 2007: PROMIS study begins in the USA
- 2011
(SELECTIVE) HISTORY

- **2200 BC**
- **1886**: Cavendish Lab
- **1950**: PROMS
- **1960**: IRT
- **2007**: PROMIS
- **2011**: Cambridge releases Concerto and conducts the myPersonality study
(SELECTIVE) HISTORY

- 2200 BC
- 1886 Cavendish Lab
- 1950 PROMS
- 1960 IRT
- 2007 PROMIS
- 2011 CONCERTO

myPersonality
Pencil-and-paper mindset was formed and persisted for millennia
Pencil-and-paper mindset was formed and persisted for millennia

Modern technologies can disrupt the paper-based mindset

- 2200 BC: Postal service, Pencil-and-paper
- 1886: Turing machine
- 1950: EHRs
- 1960: PCs
- 2007: Cloud computing, APIs, Smartphones
- 2011: Deep learning, Social media, Statistical computing languages
“[D]ata of a very large size, typically to the extent that its manipulation and management present significant logistical challenges.” OED
If it crashes Excel or SPSS, it is **BIG** data
BIG DATA ERA

• Every computer-based interaction generates data
• Every day 2.5 billion gigabytes of data are created (2.5 exabytes)

A = 1 Byte
BIG DATA ERA

1992
100gb | day

1997
100gb | hour

2013
25,000gb | second

2018
50,000gb | second
Regression (continuous prediction)

- Given multiple X values (inputs), what is the output (Y) value?
- Analogous to linear regression

Classification (binary prediction)

- What class does this data belong to?
- Analogous to logistic regression

Clustering (marking similarities)

- What is this data most similar to?
- Analogous to component/factor analysis
Human-level performance in automated –

• Object and image recognition
• Written text comprehension
• Video game mastery
• Inferring sounds from videos
• Board games (Chess, Go)
• Language translation

• Personality assessment (2013)
• Depression assessment (2016)
• Classifying doctors (2017)
• Classification of skin cancers (2017)
ML IN PRACTICE
Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun
STRENGTHS

- Task automation
- Near-unlimited workload capacity
- Optimise decision making
- New forms of data (images, text)
- Create insights instantly
- Constant monitoring
- Easily/cheaply scalable
WEAKNESSES

• Requires available and ‘clean’ data
• Requires highly skilled workers
• Requires well-organised infrastructure
• Big data can require linking
• Identification (care.data)
• Black-box problem
• Probabilistic, not deterministic

“All models are wrong, but some are useful” – Box (1987)
OPPORTUNITIES

- New forms of patient-reported assessments
- Improve quality and safety via constant monitoring
- Learning health systems
- Automation of tasks that were at risk of human error
- Event forecasting (relapse/safety)
- Foundation of continuous improvement (IoM, 2011)
OPPORTUNITIES

- New forms of patient-reported assessments
- Improve quality and safety via constant monitoring
- Learning health systems
- Automation of tasks that were at risk of human error
- Event forecasting (relapse/safety)
- Foundation of continuous improvement (IoM, 2011)
THREATS

• Encoding biases from poorly-selected data
• Missing important sources of information
• False sense of security
• Focussing on common events, missing big ones (scarce events in the long tail)
• Making patient care less patient centred
A PATIENT-CENTRED BIG DATA FUTURE

Patient-reported data

Which could be:
- Questionnaires
- Open-text
- Social media (?)

Patient-centred big data analytics to improve quality, efficiency and safety in the NHS
THREE TECHNIQUES & ONE TECHNOLOGY

- Computer adaptive testing
  (Efficiency and precision)
- Prediction and feedback
  (Validation and engagement)
- Machine learning
  (New forms of data)
Study One:
Developing an electronic assessment system for the World Health Organisation Quality of Life 100 assessment
Goal: Improve the accuracy and efficiency of quality of life assessment.
COMPUTER ADAPTIVE TESTING

• Referred to as CAT

• “Computer system which iteratively ‘learns’ about the test taker and selects the best item from a large ‘bank’ of items”

• Interfaces with item response theory

• Could be efficient, more reliable and better targeted

• Used in international high stakes educational assessments (USA, UK and Australia)
COMPUTER ADAPTIVE TESTING

Item Bank

1  2
3  4
6  5
7  8

Computer adaptive test

Participant
COMPUTER ADAPTIVE TESTING

Item Bank

1  2
3  4
6  5
7  8

Computer adaptive test

Best item

Participant
COMPUTER ADAPTIVE TESTING

Item Bank

- 2
- 3
- 4
- 6
- 5
- 7
- 8

Computer adaptive test

Participant

Item
COMPUTER ADAPTIVE TESTING

Item bank

1. 2
2. 3
3. 4
4. 5
5. 6
6. 7
7. 8

Computer adaptive test

Participant

Response
COMPUTER ADAPTIVE TESTING

Item bank

2 3 4 5 6 7 8

Computer adaptive test

\[ P_{ni} = \frac{e^{(\theta_n - b_i)}}{1 + e^{(\theta_n - b_i)}} \]

Participant

Response

Quality of life estimate = 22

Reliability

| 70 | .80 | .90 |

Quality of life estimate = 22
Computer Adaptive Testing

Item Bank

Computer adaptive test

Participant

Best item

Quality of life estimate = 22

Reliability

70  .80  .90
COMPUTER ADAPTIVE TESTING

Quality of life estimate = 47

Reliability

<table>
<thead>
<tr>
<th></th>
<th>70</th>
<th>.80</th>
<th>.90</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
COMPUTER ADAPTIVE TESTING

Item Bank

2
3
4
5
7

Computer adaptive test

Quality of life estimate = 55

Participant

1
6
8

Reliability

70          .80           .90
Quality of life estimate = 60

Item Bank

Computer adaptive test

Participant

70
Reliability
.80
.90
Select appropriate items based on a patient’s level of functional impairment

http://www.pittmed.health.pitt.edu/story/next-guy
The WHOQOL

- Multi-dimensional measure of global quality of life
- Physical, psychological, social and environmental domains
- Fifteen field centers
Item bank development

- Psychometric analysis of the WHOQOL-100 on UK sample \( (n = 320) \)
- Four domains fit the Rasch Model \( (p > 0.5) \)
- 52 items removed
- Mean 11 items per bank
- Banks were suitable for patients with long-term conditions
WHOQOL adaptive test

reliability .90  reliability .80  reliability .70
9 Items        4 Items        2 Items
82% shorter    46% shorter    
than WHOQOL-100 than WHOQOL-BREF

WHOQOL can be 82% shorter and more reliable
(Gibbons et al., 2016. J Med Int Res)

WHOQOL-CAT assessments are accurate and comparable between countries (Gibbons et al., 2015. Qual Life Res)
Computer adaptive test

Personalised Quality of Life (QoL) feedback
Unique Identifier - NPOZN

<table>
<thead>
<tr>
<th>Scale</th>
<th>Score out of 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical QoL</td>
<td>50</td>
</tr>
<tr>
<td>Psychological QoL</td>
<td>26</td>
</tr>
<tr>
<td>Social QoL</td>
<td>52</td>
</tr>
<tr>
<td>Environmental QoL</td>
<td>79</td>
</tr>
</tbody>
</table>

The scores above are worked out from the answers that you gave to the questions you have just completed. Your results are given on the chart above. A higher score means that you have a higher quality of life. The table below gives you a little bit more information about what each of the scores mean.

Physical Quality of Life
Your physical quality of life includes things like how well you are able to move around, how much energy you have, or how much you are in pain.

Your score of 50 on this scale indicates that your physical quality of life is normal. This means that the scores you gave on the questionnaire are similar to most people in the United Kingdom. It means that you are likely to be satisfied with your physical health.

The score suggests that you have no real issues with your mobility or doing the things which you enjoy and are satisfied with your energy, sleep and ability to work. It may be that you have some physical issues that you would like to discuss, and if that is the case, we have provided some links to services below.

If you are worried about your physical quality of life then you should contact your doctor.

You can click here to find a local doctor or click here to access local emergency services.

eHealthTools.co.uk
Adaptiveqol.com
Study Two:
Developing a predictive module for the WHOQOL-CAT
Goal:
Develop a new technique to assess the accuracy and acceptability of computer adaptive tests.
Feedback is crucial (e.g., Krageloh et al 2015, Kosinski et al, 2013)

Clinicians uncertain of CAT evidence

‘Real world’ evidence is lacking

CAT causes a loss of item-level information
Information from items answered in CAT (Person location/θ)

Predict response to unanswered questions

Feedback predictions to patients

Elicit reports on prediction accuracy
PREDICATE Study

1238 participants
311 UK
45 years old
97 countries
70 chronic conditions
56% male
Results

Entire assessment takes about **Two minutes**

about **10 minutes** quicker than the WHOQOL-100

(Mean assessment time 120.60± 35 seconds)
1244 predictions
Results

79%

average expected accuracy of predictions if real-world data fits the model perfectly
Participants from the UK rated the predictions as “Exactly right”
25% of participants from the UK rated the predictions as "Close, but not quite right".
participants from the UK rated the predictions as “Wrong”
Current feedback

**Personalised Quality of Life (QoL) feedback**

Unique Identifier - CQH2H

<table>
<thead>
<tr>
<th>Scale</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical QoL</td>
<td>50</td>
</tr>
<tr>
<td>Psychological QoL</td>
<td>54</td>
</tr>
<tr>
<td>Social QoL</td>
<td>52</td>
</tr>
<tr>
<td>Environmental QoL</td>
<td>79</td>
</tr>
</tbody>
</table>

**Psychological Quality of Life**

Your psychological quality of life includes things like how much confidence you have in yourself, negative feelings you experience or how much you value yourself.

Your score of 54 on this scale indicates that your psychological quality of life is normal. It looks like your psychological health is similar to most people in the United Kingdom. You can manage any negative feelings and often feel happy with your life. If this doesn’t sound right, and you’re feeling low then you could discuss these feelings with your doctor, who can help. You might also like to look at some of the links below.

If you have any concerns about your psychological quality of life you can [click here to find help locally](#) or [click here to get in touch with the Samaritans](#).


<table>
<thead>
<tr>
<th>Question 1</th>
<th>How satisfied are you with your ability to make decisions?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer 1</td>
<td>Satisfied</td>
</tr>
<tr>
<td>Question 2</td>
<td>How much confidence do you have in yourself?</td>
</tr>
<tr>
<td>Answer 2</td>
<td>A moderate amount</td>
</tr>
<tr>
<td>Question 3</td>
<td>How satisfied are you with your abilities?</td>
</tr>
<tr>
<td>Answer 3</td>
<td>Neither satisfied nor dissatisfied</td>
</tr>
<tr>
<td>Question 4</td>
<td>How worried do you feel?</td>
</tr>
<tr>
<td>Answer 4</td>
<td>Very</td>
</tr>
</tbody>
</table>
Enhanced feedback

How much do you enjoy life?

<table>
<thead>
<tr>
<th></th>
<th>Not at all</th>
<th>A little</th>
<th>A moderate amount</th>
<th>Very much</th>
<th>An extreme amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent</td>
<td>1%</td>
<td>22%</td>
<td>42%</td>
<td>29%</td>
<td>4%</td>
</tr>
</tbody>
</table>

Do you generally feel content?

<table>
<thead>
<tr>
<th></th>
<th>Never</th>
<th>Seldom</th>
<th>Quite often</th>
<th>Very often</th>
<th>Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent</td>
<td>1%</td>
<td>1%</td>
<td>65%</td>
<td>12%</td>
<td>15%</td>
</tr>
</tbody>
</table>

How positive do you feel about the future?

<table>
<thead>
<tr>
<th></th>
<th>Not at all</th>
<th>Slightly</th>
<th>Moderately</th>
<th>Very</th>
<th>Extremely</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent</td>
<td>1%</td>
<td>6%</td>
<td>70%</td>
<td>22%</td>
<td>1%</td>
</tr>
</tbody>
</table>

How well are you able to concentrate?

<table>
<thead>
<tr>
<th></th>
<th>Not at all</th>
<th>Slightly</th>
<th>Moderately</th>
<th>Very</th>
<th>Extremely</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent</td>
<td>1%</td>
<td>10%</td>
<td>30%</td>
<td>56%</td>
<td>1%</td>
</tr>
</tbody>
</table>

Prediction certainty

<table>
<thead>
<tr>
<th></th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
</table>
ATLanTiC conclusions

• The item bank is brief and can be administered in 120 seconds

• As much as a 82% increase in efficiency compared to paper-based versions

• Simulations translated well to ‘real world’ assessments

• Prediction study supports the validity of the measurement model and ‘backfilling’ unanswered item bank items
Study Three:
Using machine learning to make sense of GP performance data
Goal:
Develop an algorithm to make sense of open-text accounts of doctors’ performance
OPEN TEXT FEEDBACK

• Many questionnaires include open-text elements to add further information

• May contain important information missed by questionnaires

• Typically underused / ignored (Wagland, 2015)

• Time-consuming to use human analysts
How good is your GP at taking your problems seriously?

<table>
<thead>
<tr>
<th></th>
<th>Very poor</th>
<th>Poor</th>
<th>Neither poor nor good</th>
<th>Good</th>
<th>Very good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate</td>
<td>9.1%</td>
<td>21.2%</td>
<td>69.7%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How good is your GP at taking your problems seriously?

<table>
<thead>
<tr>
<th></th>
<th>Very poor</th>
<th>Poor</th>
<th>Neither poor nor good</th>
<th>Good</th>
<th>Very good</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9.1%</td>
<td>21.2%</td>
<td>69.7%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Burt et al., 2017
People who said “Good” or “Very good” verbally asked – Did you GP take your problems seriously?

“Taking your problems seriously? Well no he didn’t.”

“No, he didn’t listen to me.”

“Well no, he didn’t really ask about symptoms.”

Burt et al., 2017
Investigating the meaning of ‘good’ or ‘very good’ patient evaluations of care in English general practice: a mixed methods study
GMC Colleague Questionnaire

360 Degree Feedback
1636 Comments
548 Doctors

20-item questionnaire
Free-text information
Comment coding

• All comments expressed positive or neutral sentiment

• Two raters (Kappa=.80) classified comments into five categories:
  ‘Professional’
  ‘Innovative’
  ‘Respected’
  ‘Interpersonal’
  ‘Popular’
TERM DOCUMENT MATRIX (TDM)

- 616 Unique words
- Weighted terms
- 94.4% sparse

<table>
<thead>
<tr>
<th>Texts</th>
<th>Terms</th>
<th>and</th>
<th>colleague</th>
<th>doctor</th>
<th>great</th>
<th>is</th>
<th>patients</th>
<th>respected</th>
<th>this</th>
<th>troublesome</th>
<th>well</th>
<th>with</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Text 2&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Text 3&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<sup>a</sup>Text 1: “A great colleague.”

<sup>b</sup>Text 2: “A troublesome colleague.”

<sup>c</sup>Text 3: “This doctor is well respected, and great with patients.”
Colleague reports of doctors’ performance (TDM)

“The algorithm”

Machine classifications

‘Professional’
‘Innovative’
‘Respected’
‘Interpersonal’
‘Popular’
THE ALGORITHM

- Machine learning ensemble consisting of:
  - Generalised linear model with LASSO regularisation
  - Support Vector Machines (radial basis kernel)
  - Random Forests (500 trees)
  - Bootstrapped boosting
  - Scaled linear discriminant analysis

- Weighted ‘voting’ to ascertain classification
- 10-fold cross validation of model performance
<table>
<thead>
<tr>
<th>Code</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Innovative</td>
<td>.98</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>.80</td>
</tr>
<tr>
<td>Popular</td>
<td>.97</td>
</tr>
<tr>
<td>Professional</td>
<td>.82</td>
</tr>
<tr>
<td>Respected</td>
<td>.87</td>
</tr>
</tbody>
</table>
Comparison of doctors’ scores on the GMC questionnaire with a rating in a category vs those with no ratings

<table>
<thead>
<tr>
<th>Code</th>
<th>Performance</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>T-test</td>
<td>Sig.</td>
</tr>
<tr>
<td>Innovative</td>
<td>48</td>
<td>.00</td>
<td>.99</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>435</td>
<td>1.98</td>
<td>.04</td>
</tr>
<tr>
<td>Popular</td>
<td>107</td>
<td>-.88</td>
<td>.38</td>
</tr>
<tr>
<td>Professional</td>
<td>643</td>
<td>2.51</td>
<td>.01</td>
</tr>
<tr>
<td>Respected</td>
<td>264</td>
<td>3.75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Any category</td>
<td>1069</td>
<td>.77</td>
<td>.001</td>
</tr>
</tbody>
</table>
PREDICTING PERFORMANCE

T-tests comparing doctors’ scores on the GMC questionnaire with a rating in a category vs those with no ratings

<table>
<thead>
<tr>
<th>Code</th>
<th>Performance</th>
<th>N</th>
<th>T-test</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Innovative</td>
<td></td>
<td>48</td>
<td>.00</td>
<td>.99</td>
</tr>
<tr>
<td>Interpersonal</td>
<td></td>
<td>435</td>
<td>1.98</td>
<td>.04</td>
</tr>
<tr>
<td>Popular</td>
<td></td>
<td>107</td>
<td>-.88</td>
<td>.38</td>
</tr>
<tr>
<td>Professional</td>
<td></td>
<td>643</td>
<td>2.51</td>
<td>.01</td>
</tr>
<tr>
<td>Respected</td>
<td></td>
<td>264</td>
<td>3.75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Any category</td>
<td></td>
<td>1069</td>
<td>.77</td>
<td>.001</td>
</tr>
</tbody>
</table>
PREDICTING PERFORMANCE

T-tests comparing doctors’ scores on the GMC questionnaire with a rating in a category vs those with no ratings

<table>
<thead>
<tr>
<th>Code</th>
<th>Performance</th>
<th>Code</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>T-test</td>
<td>Sig.</td>
</tr>
<tr>
<td>Innovative</td>
<td>48</td>
<td>.00</td>
<td>.99</td>
</tr>
<tr>
<td>Interpersonal</td>
<td>435</td>
<td>1.98</td>
<td>.04</td>
</tr>
<tr>
<td>Popular</td>
<td>107</td>
<td>-.88</td>
<td>.38</td>
</tr>
<tr>
<td>Professional</td>
<td>643</td>
<td>2.51</td>
<td>.01</td>
</tr>
<tr>
<td>Respected</td>
<td>264</td>
<td>3.75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Any category</td>
<td>1069</td>
<td>.77</td>
<td>.001</td>
</tr>
</tbody>
</table>
MACHINE LEARNING

• Machine learning algorithms can classify open-text reports of doctors’ performance with human-level accuracy.

• Machine-made classifications can signal significant differences in doctors’ performance even when all comments were positive.
**COLLEAGUE FEEDBACK TOOL**

**Open-text evaluation for NHS doctors**

The way you talk about the doctors you work with can tell us important things about how they perform.

If you write a description of a doctor or other colleague in the box below our system will code the text into five themes which are important markers of a doctor’s performance. The system uses a machine learning algorithm which has been taught to recognise patterns in open-text reports on thousands of doctors from the UK.

Enter your description of your colleague here and then press submit:

Enter text here...

If you need some inspiration then click here to see some examples.

Submit

**Predictions made from your text**

You wrote: “This doctor is a highly motivated and dedicated professional. His care and treatment of patients and his team is highly commendable. In times of difficulty, he is endlessly patient and painstakingly meticulous in his professional aspects with patients and his team.”

The system processes the text that have written and predicts, based on the text you submitted, whether the doctor could be classified into one of the five categories shown in the graph below. These predictions are expressed as a percentage and a score over 50 means that the doctor is likely to belong in that category (e.g., “Respected”).

This system has been designed to demonstrate the ability of machine learning algorithms to automatically make sense of open-text data and to provide instant feedback. The system complements our research paper which demonstrates its accuracy. You can read the paper here.

The system was designed and developed by Chris Gibbons using C1010 with assistance from colleagues at the University of Cambridge Psychometrics Centre and The Primary Care Research Group, University of Essex.

Try the text prediction again
SUMMARY SO FAR

- Computer adaptive testing is an acceptable way to make assessments shorter and more reliable
- Machine learning may be used to make sense of open-text data in questionnaires
The problem
Concerto

Open source software for the development of online assessments utilising computer adaptive testing, machine learning and tailored feedback
Mission

• Adaptive and electronic measurement should be widely available to non-experts

• It should always be free and open-source

• Assessment experience should be made as enjoyable as possible

• Data security protocol should be dictated by the test-developer (your servers, not ours)

• Where expertise is needed, it be should be readily available as consultancy, training, or free guides online
In practice
NEW FORMS OF DATA

PREDICTING PERSONALITY FROM DIGITAL FOOTPRINT
• Example of Big Data analytics in social science, using personality questionnaires
• What might big data analytics based on social media using questionnaires or PROMs look like?
Psychometric tests (BIG-5) hosted on Facebook

Developed and managed by psychometrics centre members

More than 7,000,000 people have completed the Big-5 questionnaire

Large proportion of users shared their Facebook ‘Like’ information
Research question:

Can you build a predictive model of a psychometric test (BIG-5) using Facebook ‘likes’
Do you like machine learning? Yes | No
MY PERSONALITY
### Private traits and attributes are predictable from digital records of human behavior

**Michal Kosinski***, David Stillwell, and Thore Graepel

*From Schull, L.L., The Psychometric Centre, University of Cambridge, Cambridge CB2 8QZ United Kingdom; and Microsoft Research, Cambridge CB3 0FB, United Kingdom

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 26, 2012)

We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of personality traits. Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites [14].

<table>
<thead>
<tr>
<th>Trait</th>
<th>Selected most predictive Likes</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Godfather</td>
<td>Jason Aldean</td>
</tr>
<tr>
<td>Mozart</td>
<td>Tyler Perry</td>
</tr>
<tr>
<td>Thunderstorms</td>
<td>Sephora</td>
</tr>
<tr>
<td>The Colbert Report</td>
<td>Chiq</td>
</tr>
<tr>
<td>Morgan Freemans Voice</td>
<td>Bret Michaels</td>
</tr>
<tr>
<td>The Daily Show</td>
<td>Clark Griswold</td>
</tr>
<tr>
<td>Lord Of The Rings</td>
<td>Bebe</td>
</tr>
<tr>
<td>To Kill A Mockingbird</td>
<td>I Love Being A Mom</td>
</tr>
<tr>
<td>Science</td>
<td>Harley Davidson</td>
</tr>
<tr>
<td><strong>Curly Fries</strong></td>
<td>Lady Antebellum</td>
</tr>
<tr>
<td><strong>Sarah Palin</strong></td>
<td>Hawthorne Heights</td>
</tr>
<tr>
<td>Glenn Beck</td>
<td>Kickass</td>
</tr>
<tr>
<td>Proud To Be Christian</td>
<td>Atreyu (Metal Band)</td>
</tr>
<tr>
<td>Indiana Jones</td>
<td>Lamb Of God</td>
</tr>
<tr>
<td>Swimming</td>
<td>Gorillaz</td>
</tr>
<tr>
<td>Jesus Christ</td>
<td>Science</td>
</tr>
<tr>
<td><strong>Bible</strong></td>
<td>Quote Portal</td>
</tr>
<tr>
<td>Jesus</td>
<td>Stewie Griffin</td>
</tr>
<tr>
<td>Being Conservative</td>
<td>Killswitch Engage</td>
</tr>
<tr>
<td>Pride And Prejudice</td>
<td>Ipod</td>
</tr>
</tbody>
</table>
Computer-based personality judgments are more accurate than those made by humans

Wu Youyou1,2, Michal Kodinska1,3, and David Stillwell4

1Department of Psychology, University of Cambridge, Cambridge CB2 3EB, United Kingdom; and 4Department of Computer Science, Stanford University, Stanford, CA 94305

Edited by David Funder, University of California, Riverside, CA, and accepted by the Editorial Board December 2, 2014 (received for review September 28, 2014)

DOES IT WORK?
• API which allows personality predictions to be made at scale

• Hilton, Wrigley's, Barclays

• Using psychological information to tailor information improves conversion rate and satisfaction (Matz, 2015)
WHY MIGHT THIS ‘INFERENTIAL’ APPROACH BE USEFUL FOR HEALTH SCIENCES

• May be less biased than questionnaire items
• Real-time monitoring/assessment/feedback
• No issues with recall (can be done anywhere)
A decade into Facebook: where is psychiatry in the digital age?

Becky Inkster, David Stillwell, Michal Kosinski, Peter Jones
GOOGLE AND THOMAS INSEL

Former head of NIMH in the USA

Moved to Google in late 2015.

“Technology can have greater impact on mental healthcare than on the care for heart disease, diabetes, cancer or other diseases... It could transform this area in the next five years.” (Insel, 2015)
Computational analysis of Instagram photos provides a more accurate estimation of depression than in-person GP interviews.

Reece & Danforth *arxiv.org pre-publication* 2016
Facebook ‘Likes’ do not accurately predict reports of symptom severity: a machine learning study
Beierl, E., Stillwell, D., Gibbons, CJ

Can Facebook ‘Likes’ be used to predict depressive symptomology?
Beierl, E., Stillwell, D., Gibbons, CJ
FACEBOOK LIKES, DEPRESSION, AND SYMPTOM REPORTING

Facebook ‘Likes’ do not accurately predict reports of symptom severity: a machine learning study
Beierl, E., Stillwell, D., Gibbons, C.J.

Can Facebook ‘Likes’ be used to predict depressive symptomatology?
Beierl, E., Stillwell, D., Gibbons, C.J.
IS USING DATA LIKE THIS A BIT CREEPY*?

*OFF-PUTTING TO PATIENTS
71% of patients were happy to share social media information with their doctor.
Assessment

- 27 Yes/No Questions on aspects of Big Data
- Measuring attitudes towards Personalisation, Ethics, Wearable Tech, Internet of Things, Finance, Policy-Making, Cloud Storage, etc.
- 20-item BIG5 Personality test

Feedback

- Participants given real-time feedback on the similarity of their answers to all previous participants
Sample (by Oct '15)

No. of test sessions in total (Sept)
33,937

Answered At Least One Question
19,126

All Big Data Questions
10,411

All Personality Questions
8,871

49% Female
54% Male

Average age
30

Male
Female
Intersex

43% Europe
15% South America
9% Asia
27% North America
### Results: Audit of Global Public Opinion

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are those with access to your personal data able to accurately predict your future behaviour? (n=19,100)</td>
<td>62% NO</td>
</tr>
<tr>
<td>Should predictive technologies be used to improve the quality of healthcare, for example by helping doctors to recommend personalised nutrition and exercise plans? (n=14,187)</td>
<td>16% NO</td>
</tr>
<tr>
<td>Should predictive technologies be used assess your eligibility for a mortgage? (n=13,725)</td>
<td>62% NO</td>
</tr>
<tr>
<td>Do you think your organisation ought to invest in predictive technologies? (n=2,830)</td>
<td>62% NO</td>
</tr>
<tr>
<td>Is it important for you to understand the psychological attributes of your customers ? (n=2,489)</td>
<td>8% NO</td>
</tr>
</tbody>
</table>
Should predictive technologies be used to improve the quality of healthcare, for example by helping doctors to recommend personalised nutrition and exercise plans? (n=14,187)

16% NO

84% YES
CONCLUSION

• CATs are shorter and more efficient than paper-based measures

• CAT accuracy and efficiency translates from ‘the lab’ to ‘the real world’

• Machine learning can ‘unlock’ unstructured data and identify important indicators of doctors’ performance

• We have to tools to make patient-centred data collection feasible at scale in the NHS and health services worldwide

• More research is needed to honestly evaluate the potential of inferential psychometrics to benefit medical research and care
NEXT STEPS

• Implementation into NHS patient portal  
  (with Queen Mary Hospital Birmingham)

• Clinical trials of CATs in different disease groups  
  (with University of Exeter and Memorial Sloan Kettering Cancer Centre)

• National CATs for quality of life and symptom reporting  
  (with NHS England and Swedish Quality Registers)

• Identifying ‘risk events’ from open-text patient experience reports  
  (with Queen Mary Hospital Birmingham)

• Identification of depression and symptoms using Facebook images  
  (with University of Oxford)
$p_{ni} = \frac{e^{(\theta_i - b_i)}}{1 + e^{(\theta_i - b_i)}}$

The “dark age” of paper-based assessments

Research in 2017

Clinical practice in 2017

Research in 2020

Clinical practice in 2020

The "dark age" of paper-based assessments
THANK YOU!

cg598@cam.ac.uk

ehealthtools.co.uk

chrisgibbons.io

Papers and slides from today: cg598@cam.ac.uk

@DrCGBibbons