Making Algorithms Trustworthy: What Can Statistical Science Contribute to Transparency, Explanation and Validation?

David Spiegelhalter

Chairman of the Winton Centre for Risk & Evidence Communication, University of Cambridge

President, Royal Statistical Society

@d_spiegel

david@statslab.cam.ac.uk

NeurIPS 2018

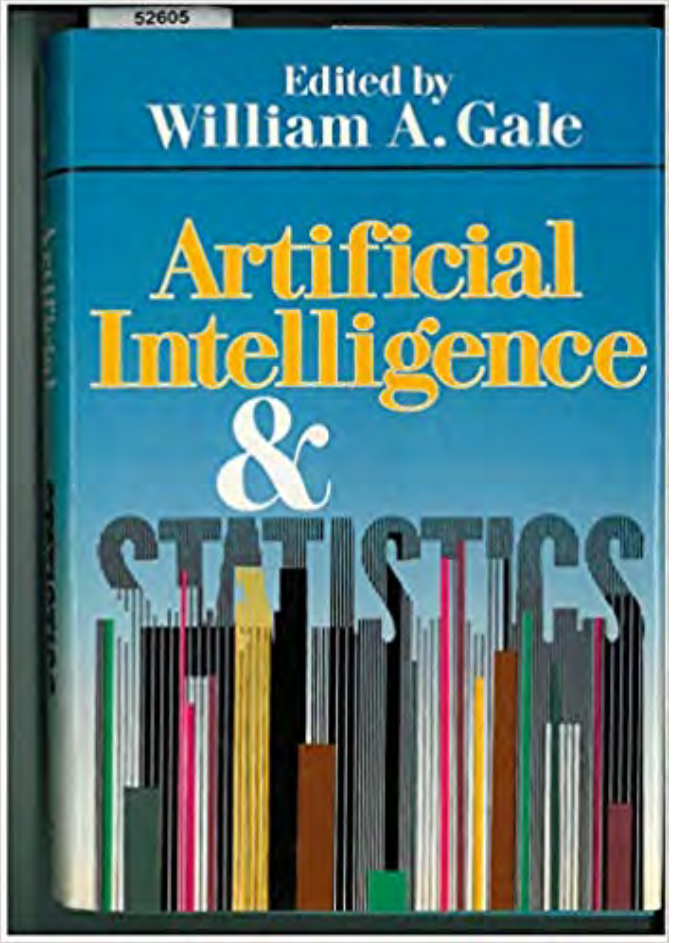
Leo Breiman 1928-2005

FOLLOW

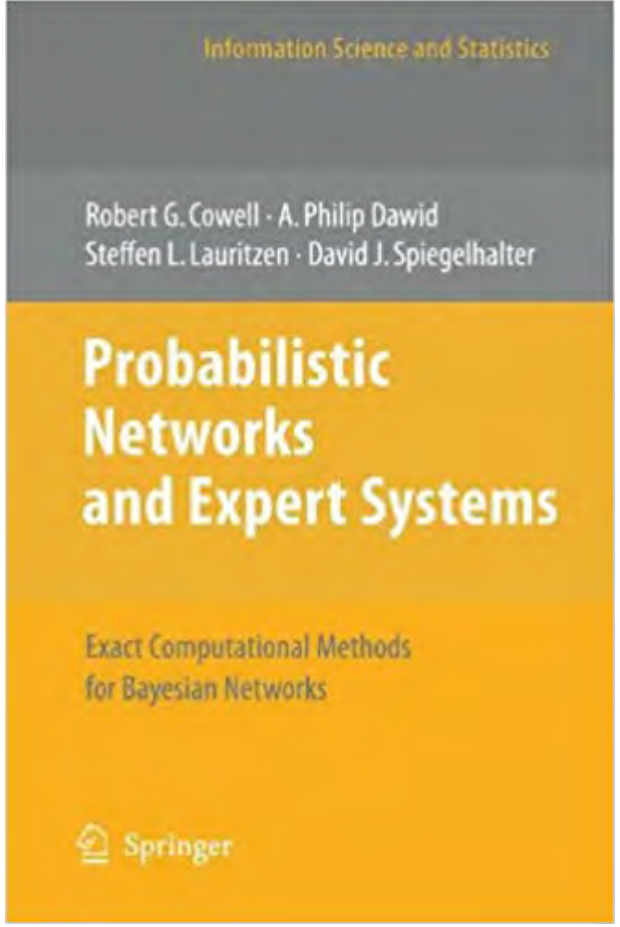
Professor of Statistics, UC Berkeley
Verified email at stat.berkeley.edu - Homepage

Data Analysis Statistics Machine Learning

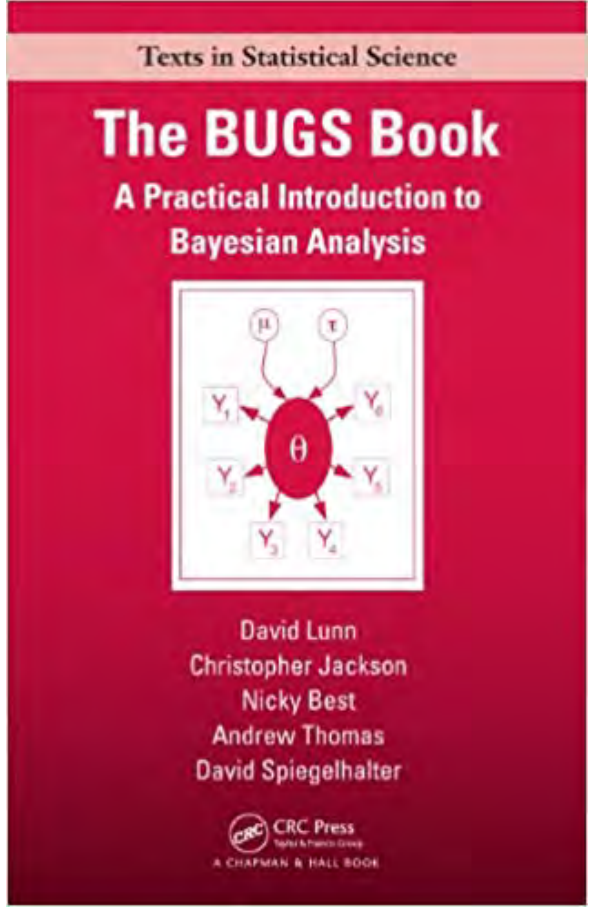
TITLE	CITED BY	YEAR
Random forests L Breiman Machine learning 45 (1), 5-32	41246	2001
Classification and Regression Trees L Breiman, JH Friedman, RA Olshen, CJ Stone CRC Press, New York	~~39082~~ *	1999
Classification and regression trees L Breiman Chapman & Hall/CRC	39082 *	1984
Bagging predictors L Breiman Machine learning 24 (2), 123-140	19724	1996



1979- 1986



1986-1990



1990-2007

BBC | Sign in | News | Sport | Weather | Shop | Earth | Travel | More | Search

FOUR

Climate Change by Numbers

Home Clips

This programme is not currently available on BBC iPlayer

At the heart of the climate change debate is a paradox - we've never had more information about our changing climate, yet surveys show that the public are, if anything, getting less sure they understand what's…

1 hour, 15 minutes

Last on

BBC FOUR

Thu 5 Mar 2015

22:00

BBC FOUR

BBC | Sign in | News | Sport | Weather | iPlayer | TV | Radio | More... | Search

FOUR

Tails You Win: The Science of Chance

Home Clips

DURATION: 1 HOUR

Smart and witty, jam-packed with augmented-reality graphics and fascinating history, this film, presented by Professor David Spiegelhalter, tries to pin down what chance is and how it works in the real world. For…

> SHOW MORE

78 Share

Next on

BBC FOUR

Next Thursday

21:00

BBC Four

See all upcoming broadcasts of Tails You Win: The Science of Chance (3)

Summary

- Trust
- A structure for evaluation
- Ranking a set of algorithms
- Layered explanations
- Explaining regression models
- Communicating uncertainty
- How some (fairly basic) statistical science might help!

(Primary focus on medical systems – only scrape surface)

Onora-O'Neill and trust

- Organisations should not be aiming to 'increase trust'

- Rather, aim to demonstrate *trustworthiness*

Artificial Intelligence

A hospital in London wants to replace doctors with AI to cut A&E waiting times

The hospital will work alongside The Alan Turing Institute to look at ways to make NHS services quicker, safer and more efficient

Bobby Hellard
23 May 2018

CISION News Products Contact Keyword, Company, Stocks

Babylon AI Achieves Equivalent Accuracy With Human Doctors in Global Healthcare First

SHARE THIS ARTICLE

Artificial Intelligence

Now DeepMind's AI can spot eye disease just as well as your doctor

The AI from Google's DeepMind made correct diagnoses 94.5 per cent of the time in a trial with Moorfields Eye Hospital

We should expect trustworthy claims

- **by** the system
- **about** the system

A structure for evaluation?

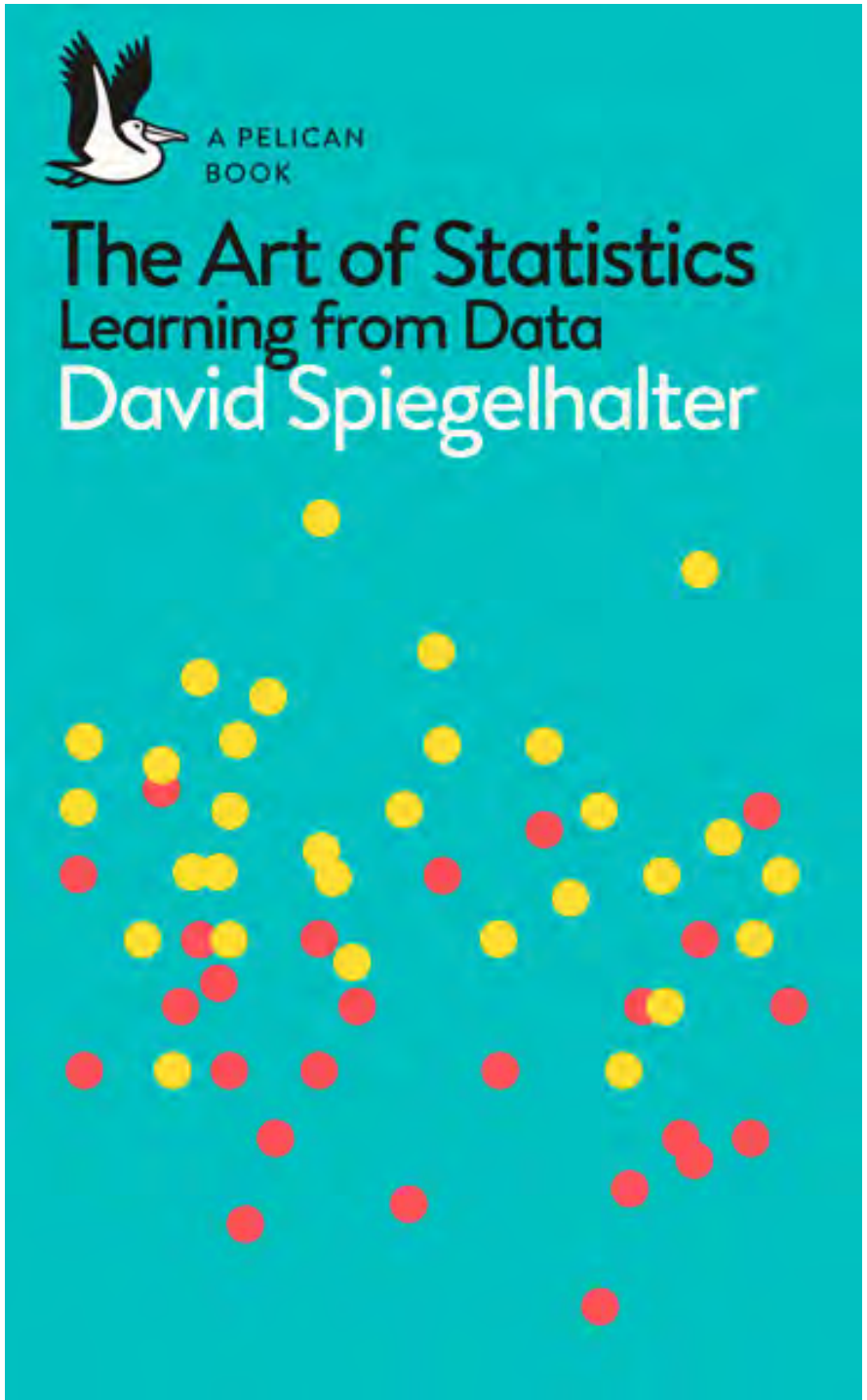
	Pharmaceuticals	Algorithms
Phase 1	*Safety*: Initial testing on human subjects	*Digital testing*: Performance on test cases
Phase 2	*Proof-of-concept*: Estimating efficacy and optimal use on selected subjects	*Laboratory testing*: Comparison with humans, user testing
Phase 3	*Randomised Controlled Trials*: Comparison against existing treatment in clinical setting	*Field testing*: Controlled trials of impact
Phase 4	*Post-marketing surveillance*: For long-term side-effects	*Routine use*: Monitoring for problems

Stead et al, J Med Inform Assoc 1994

Phase 1: digital testing

- A statistical perspective on algorithm competitions

Ilfracombe, North Devon


A PELICAN
BOOK
The Art of Statistics
Learning from Data
David Spiegelhalter

William Somerton's entry in a public database of 1309 passengers (39% survive)

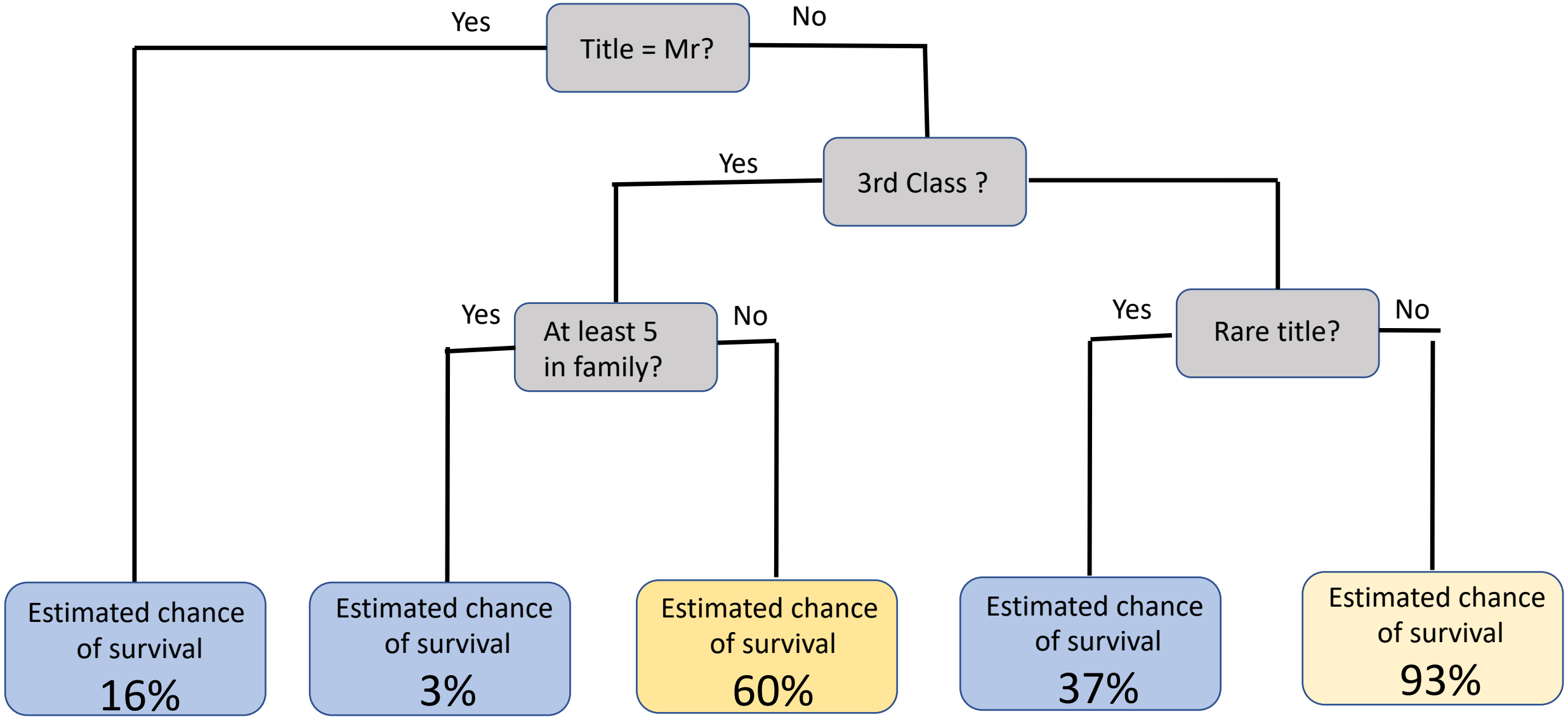
pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body
3	0	Somerton, Mr. Francis William	male	30	0	0	A.5. 18509	8.0500		S		
3	0	Spector, Mr. Woolf	male		0	0	A.5. 3236	8.0500		S		
3	0	Spinner, Mr. Henry John	male	32	0	0	STON/OQ. 369943	8.0500		S		
3	0	Staneff, Mr. Ivan	male		0	0	349208	7.8958		S		
3	0	Stankovic, Mr. Ivan	male	33	0	0	349239	8.6625		C		
3	1	Stanley, Miss. Amy Zillah Elsie	female	23	0	0	CA. 2314	7.5500		S	C	
3	0	Stanley, Mr. Edward Roland	male	21	0	0	A/4 45380	8.0500		S		

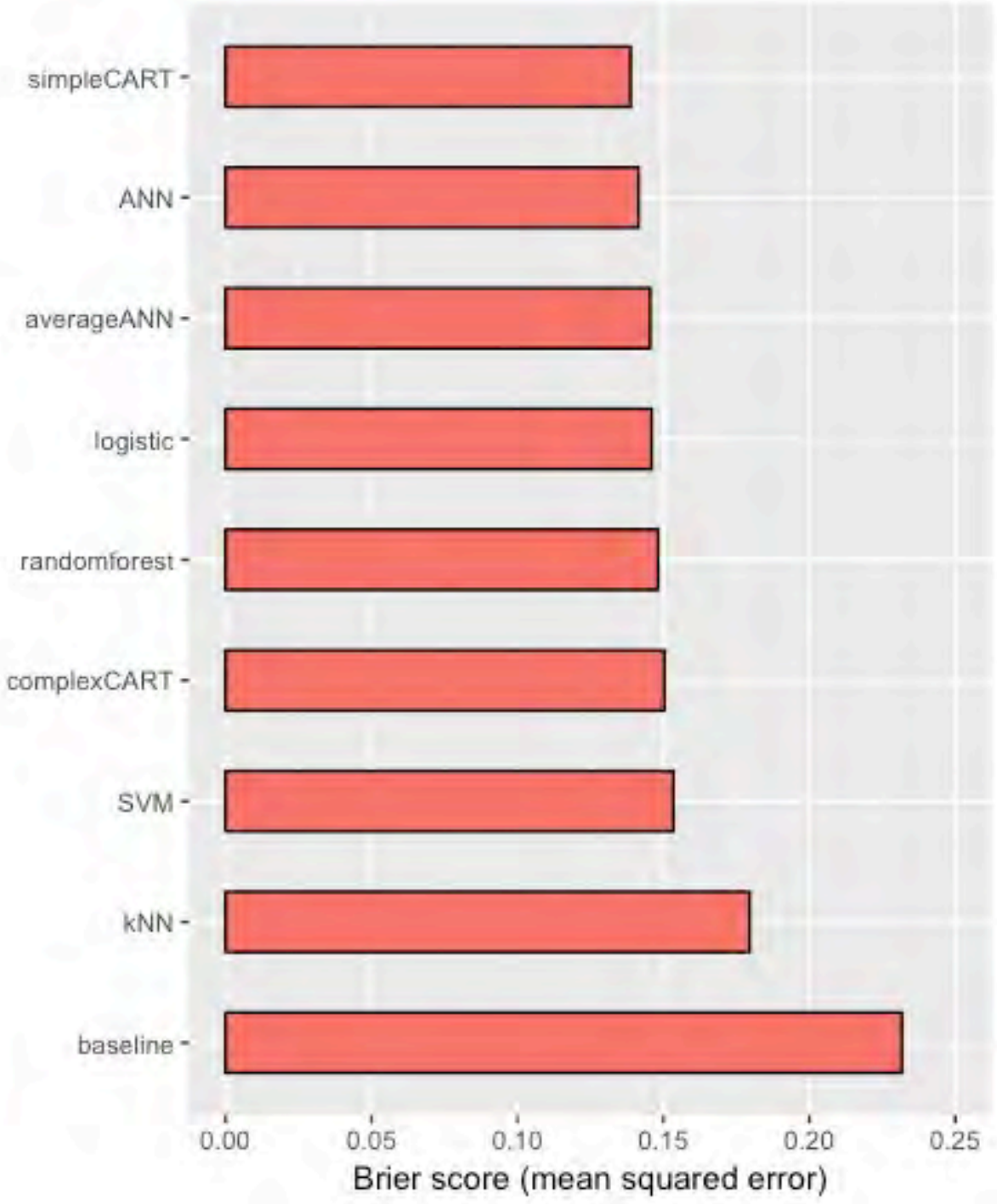
- Copy structure of Kaggle competition (currently over 59,000 entries)
- Split data-base of 1309 passengers at random into
  - **training set** (70%)
  - **test set** (30%)
- Which is the best algorithm to predict who survives?

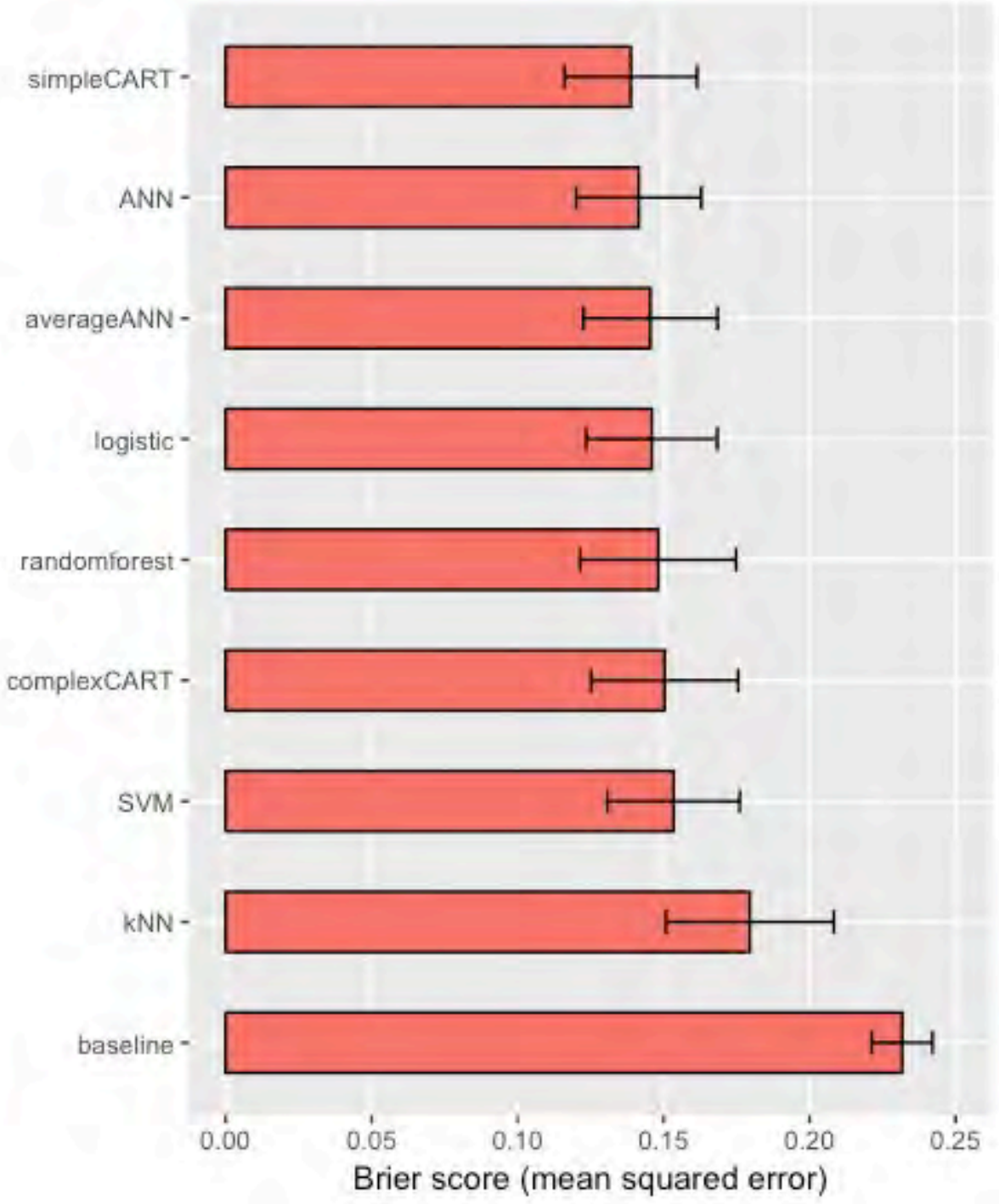
Performance of a range of (**non-optimised**) methods on test set

Method	Accuracy (high is good)	Brier score (MSE) (low is good)
Simple classification tree	**0.806**	**0.139**
Averaged neural network	0.794	0.142
Neural network	0.794	0.146
Logistic regression	0.789	0.146
Random forest	0.799	0.148
Classification tree (over-fitted)	0.806	0.150
Support Vector Machine (SVM)	0.782	0.153
K-nearest-neighbour	0.774	0.180
Everyone has a 39% chance of surviving	0.639	0.232

Simple classification tree for Titanic data

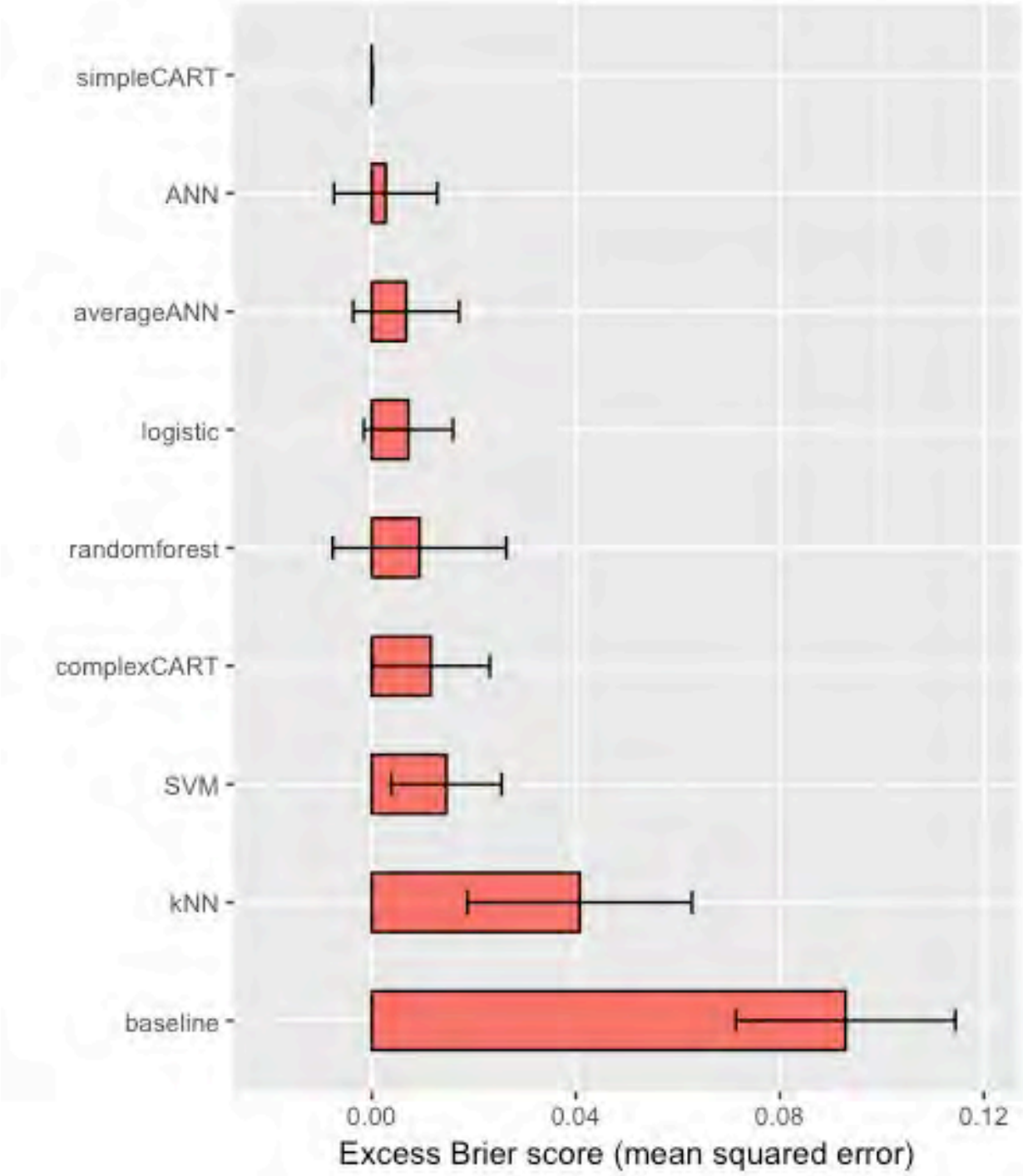



simpleCART
ANN
averageANN
logistic
randomforest
complexCART
SVM
kNN
baseline
0.00
0.05
0.10
0.15
0.20
0.25
Brier score (mean squared error)


simpleCART
ANN
averageANN
logistic
randomforest
complexCART
SVM
kNN
baseline
0.00
0.05
0.10
0.15
0.20
0.25
Brier score (mean squared error)

- Potentially a very misleading graphic!
- When comparing, need to acknowledge that tested on same cases
- Calculate differences and their standard error

- How confident can we be that simple CART is best algorithm?

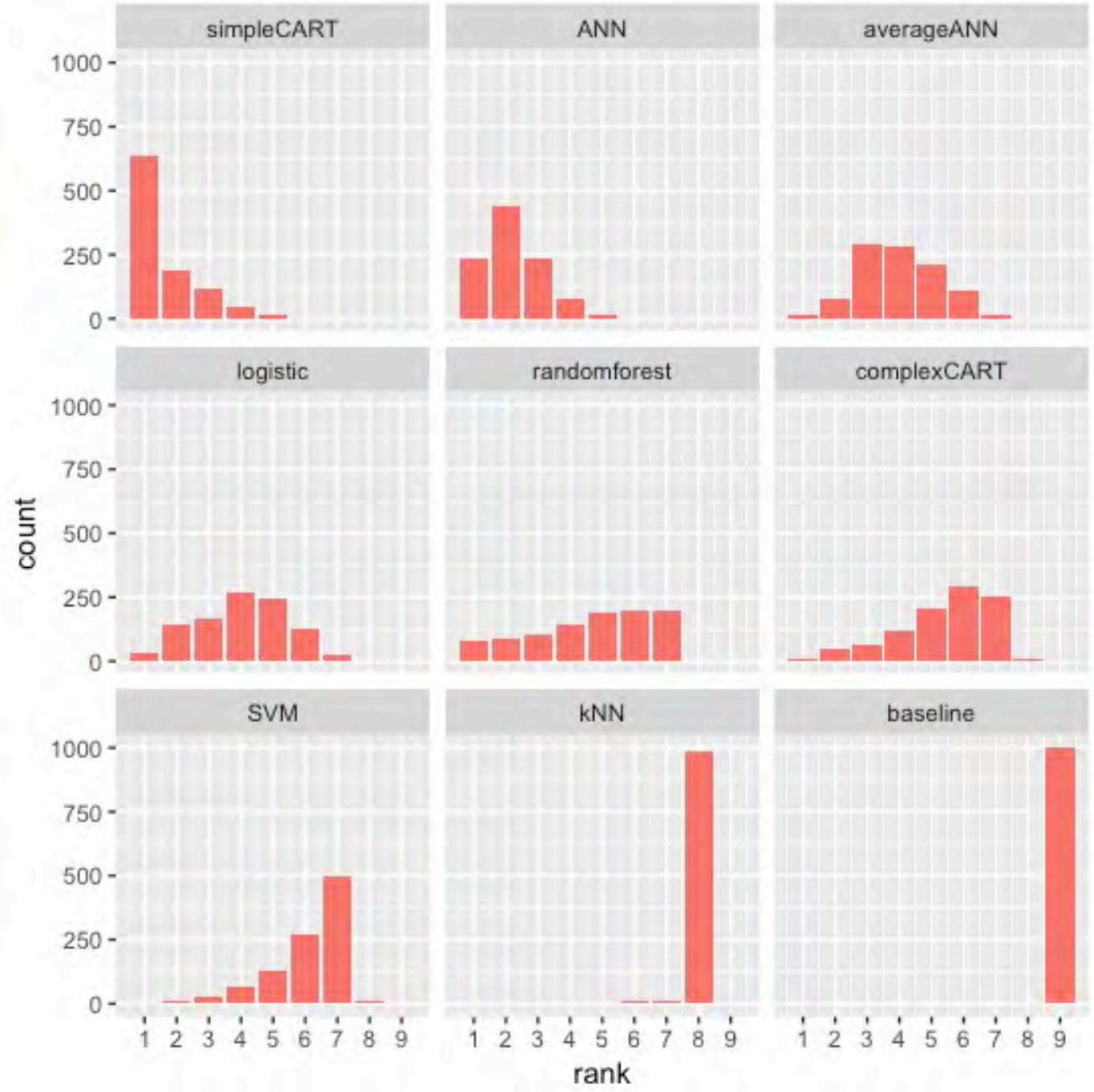


Ranking of algorithms

- Bootstrap sample from test set (ie sample of same size, drawn with replacement)
- Rank algorithms by performance on the bootstrap sample
- Repeat ‘000s of times

- (ranks actual *algorithm* – if want to rank *methods*, need to bootstrap training data too, and reconstruct algorithm each time)

Distribution of true rank of each algorithm



Probability of ‘best’:

63% simpleCART
23% ANN
8% randomforest

Who was the luckiest person on the Titanic?

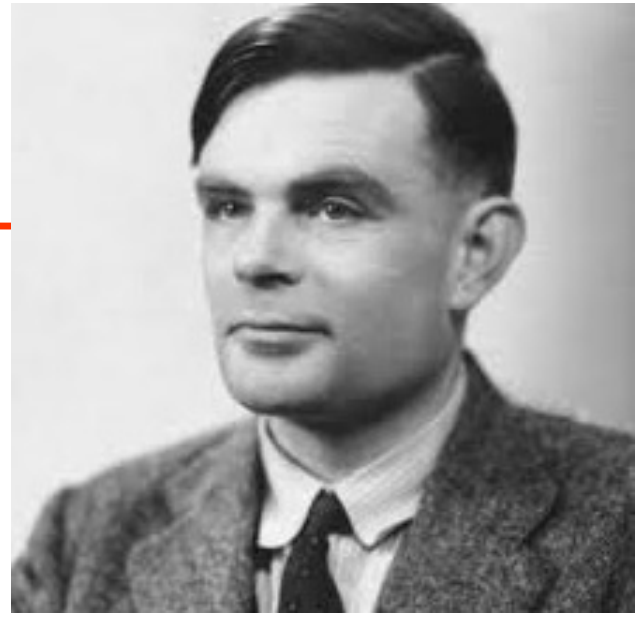
- Karl Dahl, a 45-year-old Norwegian/Australian joiner travelling on his own in third class, paid the same fare as Francis Somerton
- Had the lowest average Brier score among survivors – a very surprising survivor
- He apparently dived into the freezing water and clambered into Lifeboat 15, in spite of some on the lifeboat trying to push him back.

- Hannah Somerton was left just £5, less than Francis spent on his ticket.

Phase 2: laboratory testing

Phase 2: laboratory testing

Judgements on test cases



Turing Test

VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

Phase 2: laboratory testing

- Can reveal expert disagreement: evaluation of Mycin in 1970s found > 30% judgements considered ‘unacceptable’ for both computer **and** clinicians
- June 2018: Babylon AI published studies of their diagnostic system, rating against ‘correct’ answers and external judge

AI's health advice is as good as a doctor's, startup says

- Critique in November 2018 Lancet
  - Selected cases
  - Influenced by one poor doctor
  - No statistical testing
  - Babylon commended for carrying out studies and quality of software
  - **Need further phased evaluation**

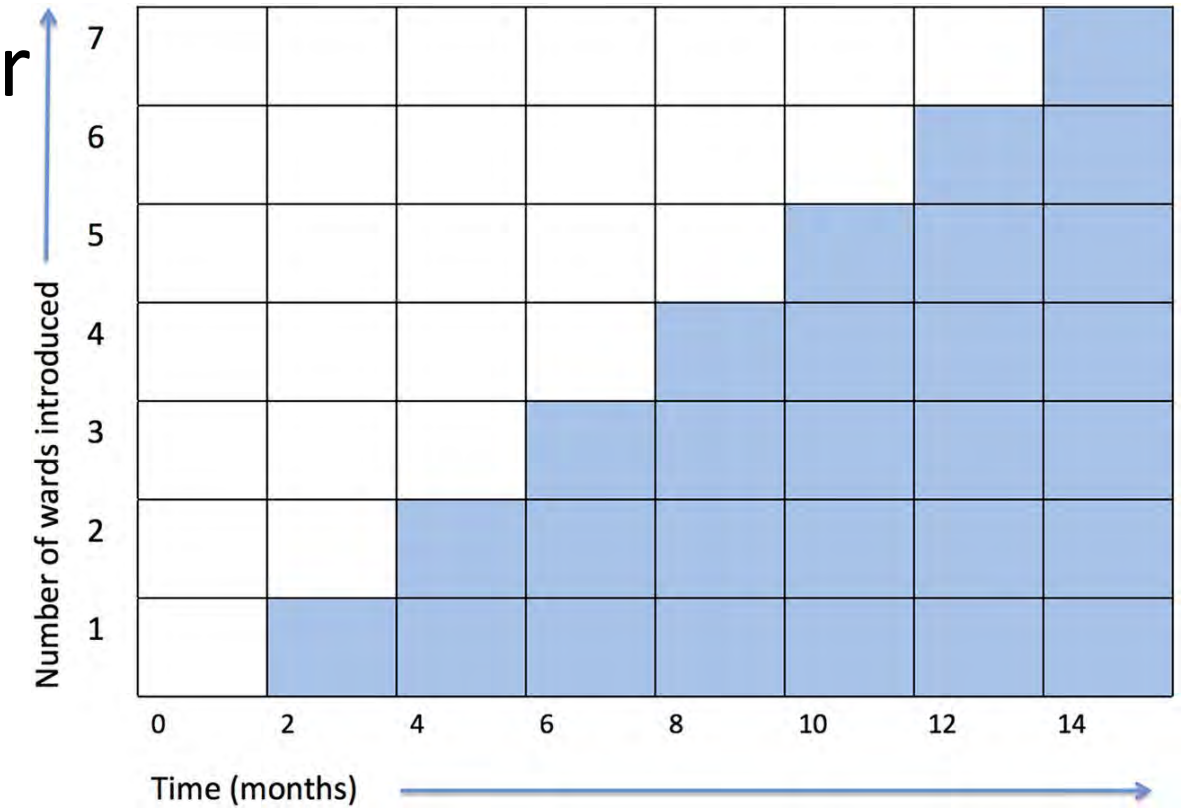
Safety of patient-facing digital symptom checkers

Yu et al, JAMA, 1979; Shortliffe, JAMA, 2018; Fraser et al, Lancet, 2018; Razzaki et al, 2018

Phase 3: field testing

Phase 3: field testing – alternative designs for Randomised Controlled Trials

- **Simple randomised:** A/B trial (but contamination….)
- **Cluster randomised:** by team/user (when expect strong group effect, need to allow for this in analysis)
- **Stepped wedge**: randomised roll-out, when expect temporal changes



Phase 3: a cluster-randomised trial of an algorithm for diagnosing acute abdominal pain

- Design: over 29 months, 40 junior doctors in Accident and Em
  cluster-randomised to
  - Control (12)
  - Forms (12) (had to give initial diagnosis)
  - Forms + computer (8)
  - Forms + computer + performance feedback (8)
- Algorithm: naïve Bayes
- > 5000 patients, but
  - Very clumsy to use
  - Only 64% accuracy
  - Over-confident: < 50% right when claiming appendicitis (but 82% when claiming 'non-specific abdominal pain')
  - Limited usage: forms (65%), computer (50%, only 39% was the result available in time)
  - Very rarely corrected an incorrect initial diagnosis.
- But, for 'non-specific' cases, admissions and surgery fell by > 45%!

So why did this fairly useless system have a positive impact?

- Reduction in operations explained by reduction in admission of ‘non-specific abdominal pain’ (NSAP)
- More correct initial diagnoses of NSAP made by junior doctors
- **Cultural change** from forms and computer, encouraging junior doctors to make a diagnosis

Wellwood et al, JRC Surgeons 1992

Phase 4: surveillance in routine use

- Ted Shortliffe on clinical decision support systems (CDSS):

- *Maintain currency of knowledge base*
- *Identify near-misses or other problems so as to inform product improvement*
- *A CDSS must be designed to be fail-safe and to do no harm*

Shortliffe, JAMA, 2018

Onora-O'Neill on transparency

- Transparency (disclosure) is not enough
- Need 'intelligent openness'
  - *accessible*
  - *intelligible*
  - *useable*
  - *assessable*

FAT/ML 2018 2017 2016 2015 2014 Organization Resources Mailing list

Scholarship Events Projects Principles and Best Practices

Principles for Accountable Algorithms and a Social Impact Statement for Algorithms

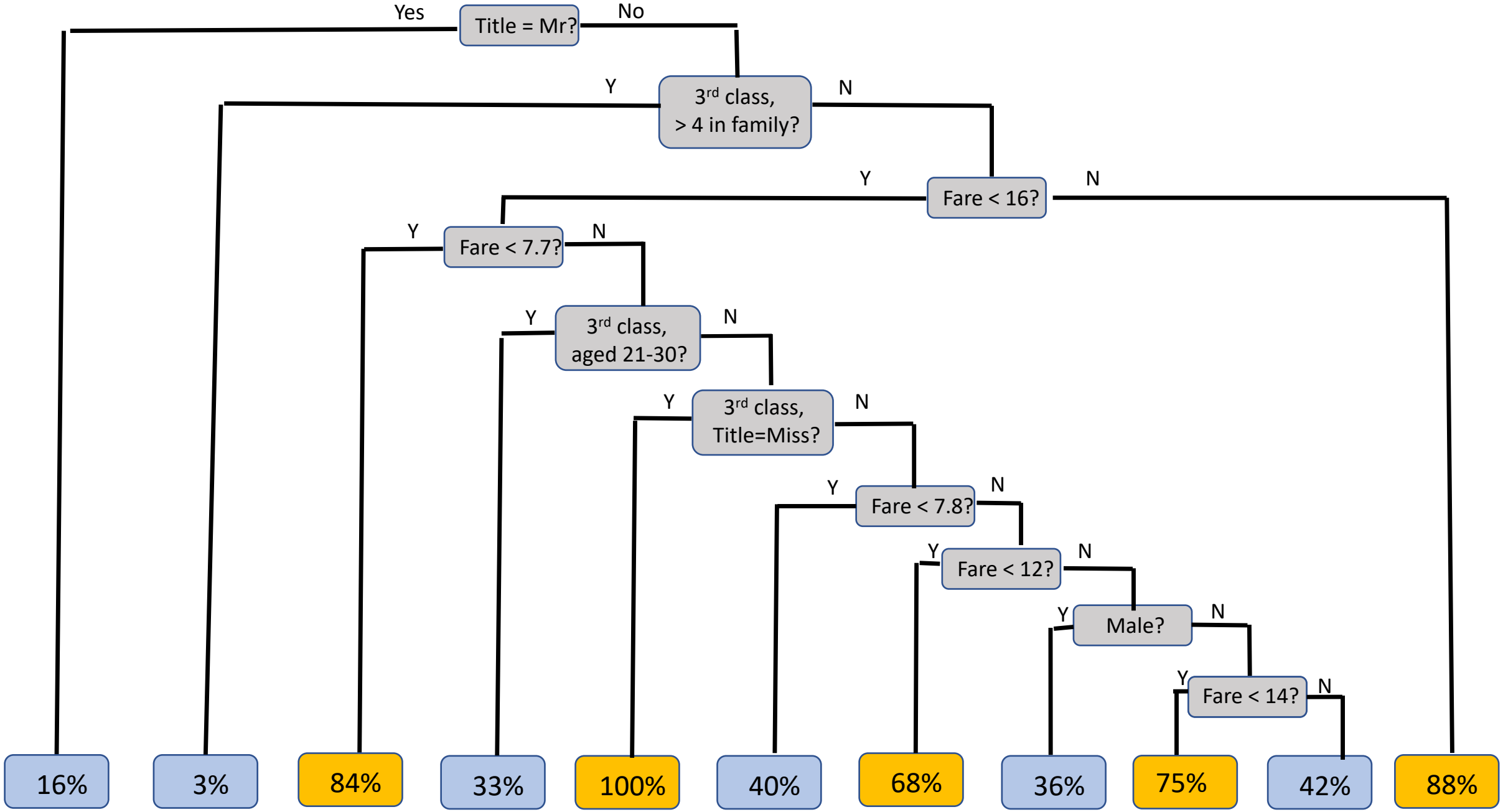
Principles for Accountable Algorithms

- **Responsibility:** whose is it?
- **Auditability:** enable understanding and checking
- **Accuracy:** how good is it? error and uncertainty
- **Explainability:** to stakeholders in non-technical terms
- **Fairness:** to different groups

But what about…

- ***Impact: what are the benefits (and harms) in actual use?***

Transparency does not necessarily imply interpretability…


Title = Mr?
Yes
No
3rd class, > 4 in family?
Y
N
Fare < 16?
Y
N
Fare < 7.7?
Y
N
3rd class, aged 21-30?
Y
N
3rd class, Title=Miss?
Y
N
Fare < 7.8?
Y
N
Fare < 12?
Y
N
Male?
Y
N
Fare < 14?
Y
N
16%
3%
84%
33%
100%
40%
68%
36%
75%
42%
88%

Explainability / Interpretability

Global explainability

About the algorithm in general:

- Empirical basis for the algorithm, pedigree, representativeness of training set etc
- Can see/understand working at different levels?
- What are, in general, the most influential items of information?
- Results of digital, laboratory and field evaluations

many checklists for reporting informatics evaluations: SUNDAE, ECONSORT etc

Local explainability

About the current claim:

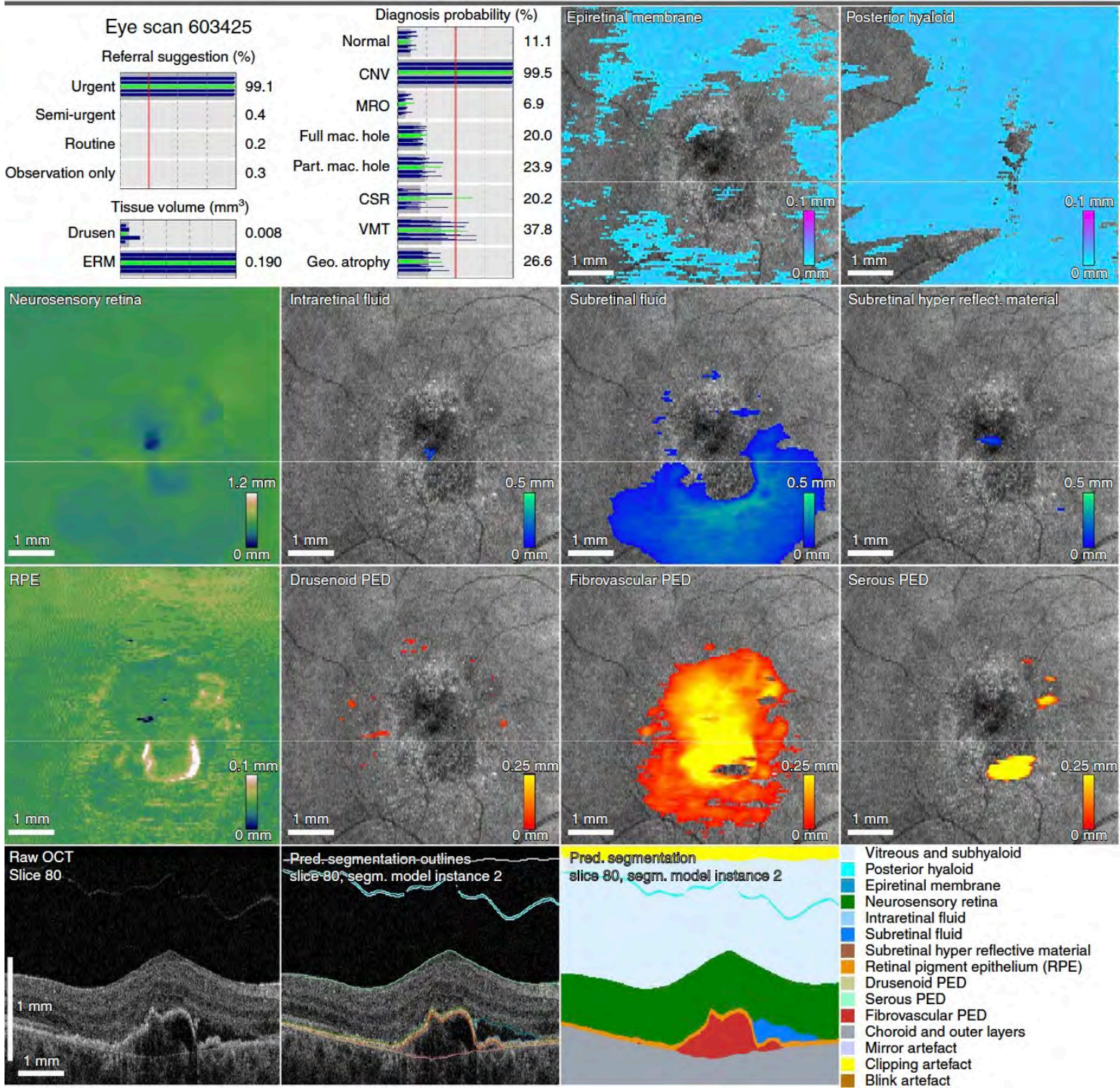
- What drove this conclusion? eg LIME
- What if the inputs had been different? Counterfactuals
- What was the chain of reasoning?
- What tipped the balance?
- Is the current situation within its competence?
- How confident is the conclusion?

Ribiero, 2016; Wachter et al, Harvard JLT, 2018;

ARTICLES
https://doi.org/10.1038/s41591-018-0107-6
nature medicine

Clinically applicable deep learning for diagnosis and referral in retinal disease

- Image from Google Deepmind / Moorfields Hospital collaboration
- Tries to explain intermediate steps between image and diagnosis/triage recommendation



predict

Clear data on breast cancer treatment outcomes

NHS

Home | About | Patient Information | Clinician Information | Predict Tool

What is Predict?

Predict is a tool that helps doctors and patients decide on treatments to have after surgery for breast cancer.

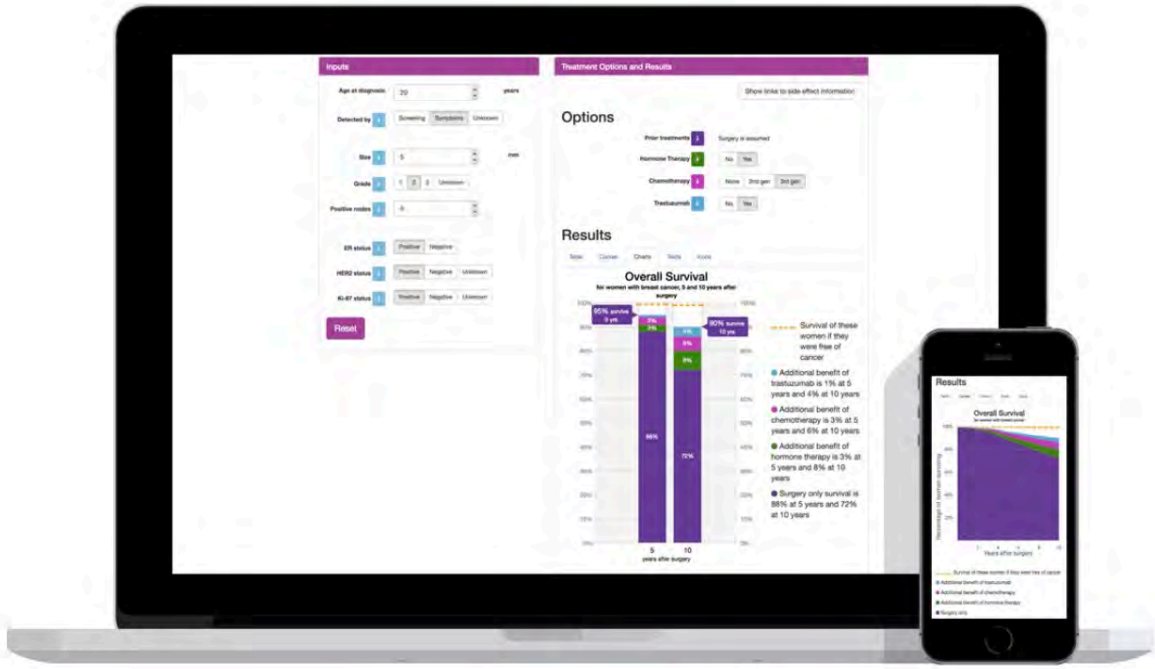
We recommend patients read the patient information section before using the tool.

What will Predict tell me?

The predict tool shows you how different treatments affect the percentage of women that survive over ten years following surgery.

How do I use Predict?

You enter details about the cancer, and then select different



Example outputs

Predict

- Common interface for professionals and patients after surgery for breast cancer
- Provides personalised survival estimates out to 15 years, with possible adjuvant treatments
- Based on competing-risk regression analysis of 3,700 women, validated in three independent data-sets
- Extensive iterative testing of interface – user-centred design
- ~ 30,000 users a month, worldwide
- Starting Phase 3 trial of supplying side-effect information
- Launching version for prostate cancer, and kidney, heart, lung transplants

Levels of explanation in Predict

1. Verbal gist.
2. Multiple graphical and numerical representations, with instant ‘what-ifs’
3. Text and tables showing methods
4. Mathematics, competing risk Cox model
5. Code.

For very different audiences!

Part of mathematical description

The form of the Predict V2.1 algorithm

The estimated baseline cumulative hazard for breast cancer mortality H_c at t years post-surgery has the form

$$H_c(t) = \exp[a_c' f(t)]$$

where a_c is a vector of estimated coefficients, and f a (column) vector of fractional polynomial functions of time post-operation (different models are built for ER+ and ER-).

In Predict 2.1,

- if ER+

$$H_c(t) = \exp[0.7424402 - 7.527762/\sqrt{t} - 1.812513 * \log(t)/\sqrt{t}]$$

- if ER-

$$H_c(t) = \exp[-1.156036 + 0.4707332/t^2 - 3.51355/t].$$

gi

The estimated survival function for breast cancer mortality S_c given risk factors x_R and the ith treatment combination x_T is given by

$$S_c^i(t) = \exp\left[-H_c(t)\exp[b_c' x_R + c' x_T]\right] = \exp\left[-\exp[a_c' f(t) + b_c' x_R + c' x_T]\right]$$

where b, c are vectors of estimated coefficients. This is the chance of living beyond t years after surgery under treatment regime i, assuming only breast cancer mortality.

Explainability / Interpretability

- Variety of audiences and purposes - developer, user, external expert etc
- GDPR demands – not sure how this is to be interpreted
- Need to properly evaluate explanations as part of impact (they may confuse or mislead)
- All sorts of clever technical things going on with black boxes: surrogates, layers
- Or build an interpretable model in the first place?

Doshi-Velez and Kim, 2017; Weller, 2017

Interpretability of regression models?

PREDICT ARREST FOR ANY OFFENSE IF SCORE > 1

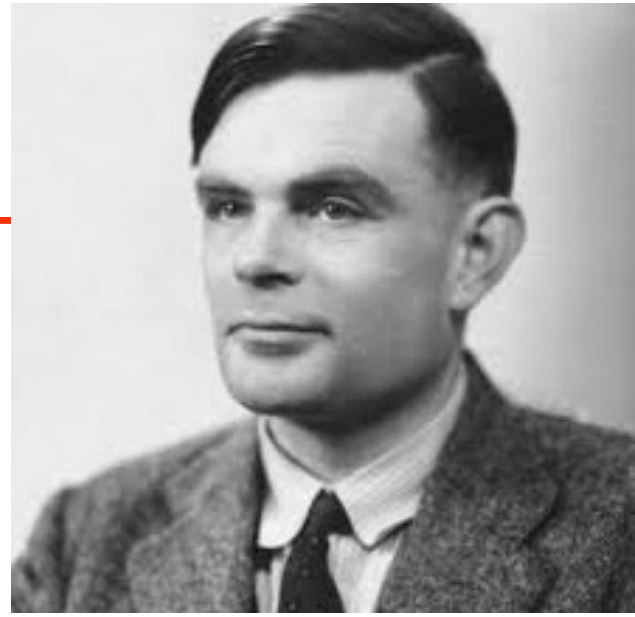
1.	Prior Arrests ≥ 2	1 point		$\cdots$
2.	Prior Arrests ≥ 5	1 point	$+$	$\cdots$
3.	Prior Arrests for Local Ordinance	1 point	$+$	$\cdots$
4.	Age at Release between 18 to 24	1 point	$+$	$\cdots$
5.	Age at Release ≥ 40	-1 points	$+$	$\cdots$
		SCORE	$=$	$\cdots$

SCORE	-1	0	1	2	3	4
RISK	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

- Scoring is interpretable (global and local)
- eg risk scoring using GAMs for pneumonia risk (Caruana)
- Rudin optimising integer scores
- Claim: don't need to trade off performance against interpretability (but in which contexts?)

Caruana et al, KDD, 2015; Rudin and Ustin, Interfaces, 2018

Alan Turing's approach to explanation



- 'Naive Bayes' classifier:

$$\frac{p(H_0|s_1,\ldots,s_p)}{p(H_1|s_1,\ldots,s_p)} = \prod_i \frac{p(s_i|H_0)}{p(s_i|H_1)} \times \frac{p(H_0)}{p(H_1)}.$$

- So

$$\log \frac{p(H_0|s_1,\ldots,s_p)}{p(H_1|s_1,\ldots,s_p)} = \sum_i w_i + \log \frac{p(H_0)}{p(H_1)},$$

where w_i is the 'weight of evidence' $\log \frac{p(s_i|H_0)}{p(s_i|H_1)}$ (Turing and Good)

- Weights of evidence are positive/negative if evidence s_i is for/against H_0
- (Can use w_i's as predictors in a logistic regression to improve their calibration)
- Multiplying by 10 and rounding helps explanation

GLADYS: diagnosis of gastrointestinal pain using input from computer-interviewing

Evidence **for** peptic ulcer		Evidence **against** peptic ulcer	
Abdominal pain	1	History less than 1 year	-8
Episodic	2	No seasonal effect	-1
Relieved by food	4	No waterbrash	-3
Woken at night	3		
Epigastric	3		
Can point at sight of pain	2		
Family history of ulcer	4		
Smoker	4		
Vomits, then eats within 3 hours	5		
Total evidence for	**28**	**Total evidence against**	**-12**
Balance of evidence	16		
Starting score	-8	(based on prevalence of 30%)	
Final score	**8**	**= 68% probability of peptic ulcer**	

Communicating uncertainty

- *"Determine how to communicate the uncertainty / margin of error for each decision".*

- Part of being trustworthy
- But will acknowledging uncertainty lose trust and credibility?

Uncertainty about statistics

Home | UK | World | Business | Politics | Tech | Science | Health | Family & Education

Business | Your Money | Market Data | Markets | Companies | Economy

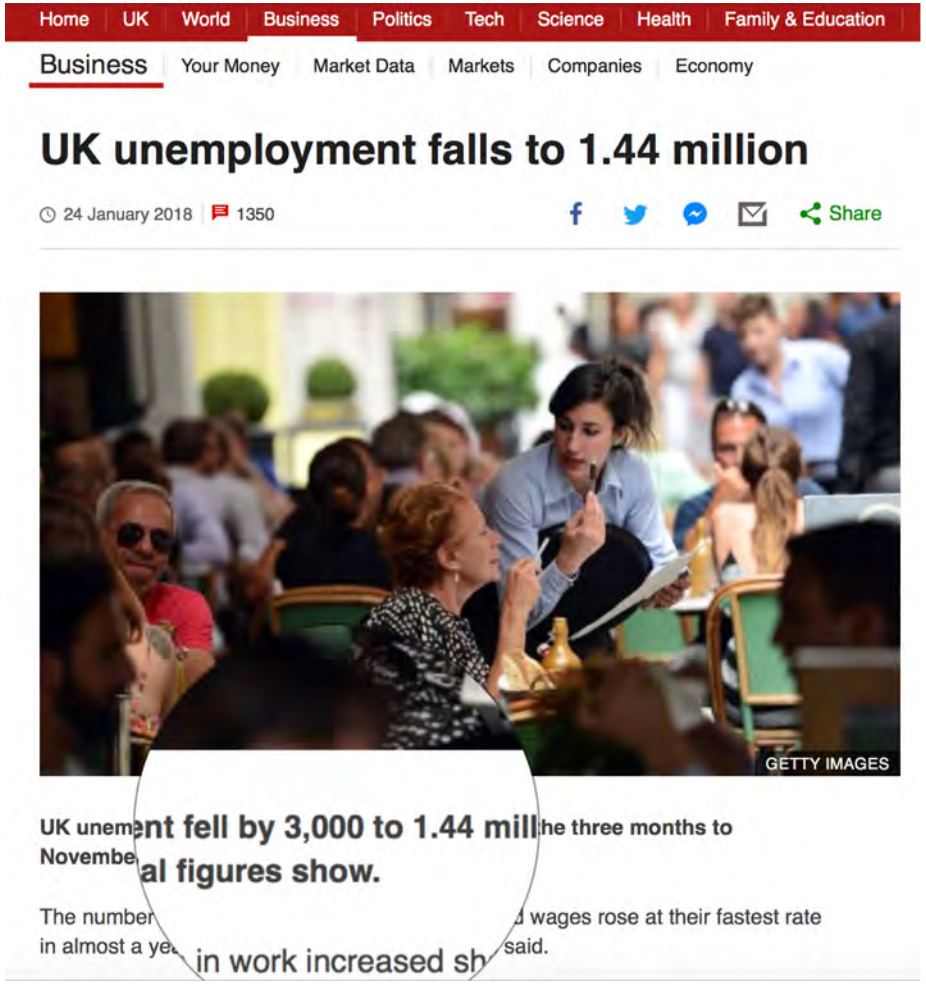
UK unemployment falls to 1.44 million

24 January 2018 | 1350 | Share

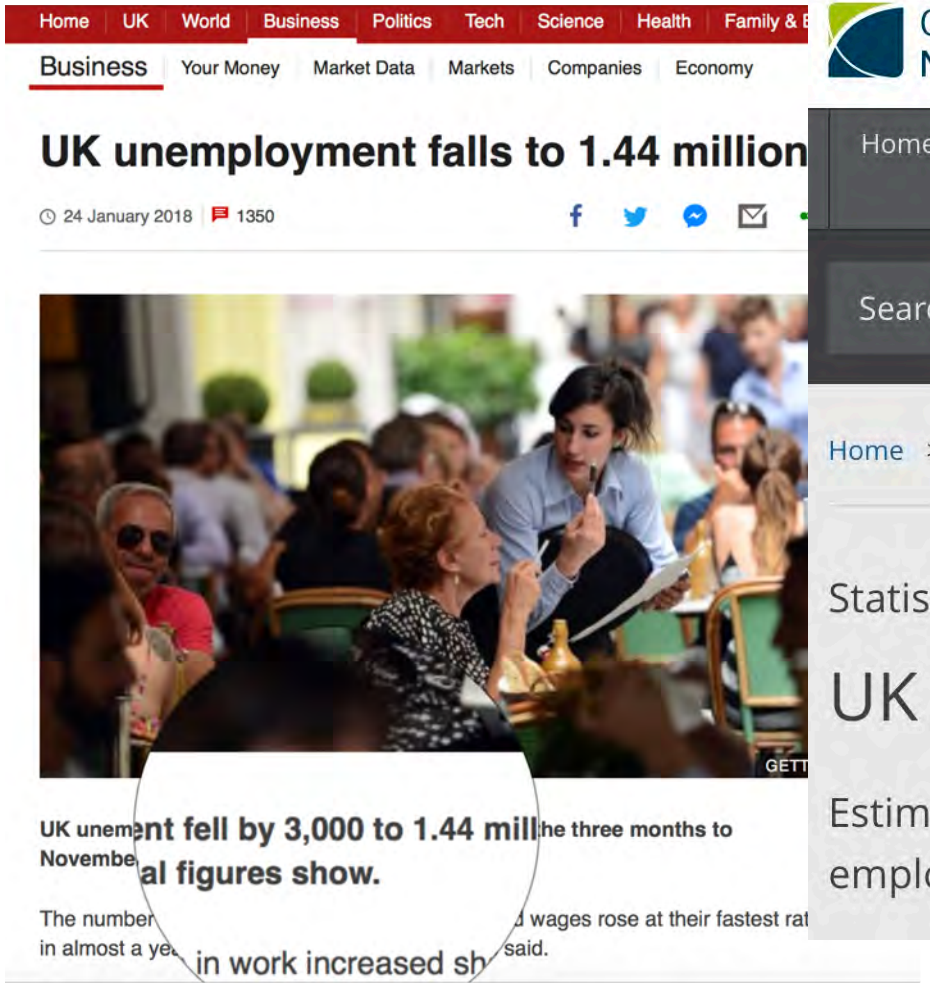
UK unemployment fell by 3,000 to 1.44 million in the three months to November, official figures show.

The number of those in work increased sharply and wages rose at their fastest rate in almost a year, the Office for National Statistics said.

Uncertainty about statistics



Uncertainty about statistics



Uncertainty about statistics

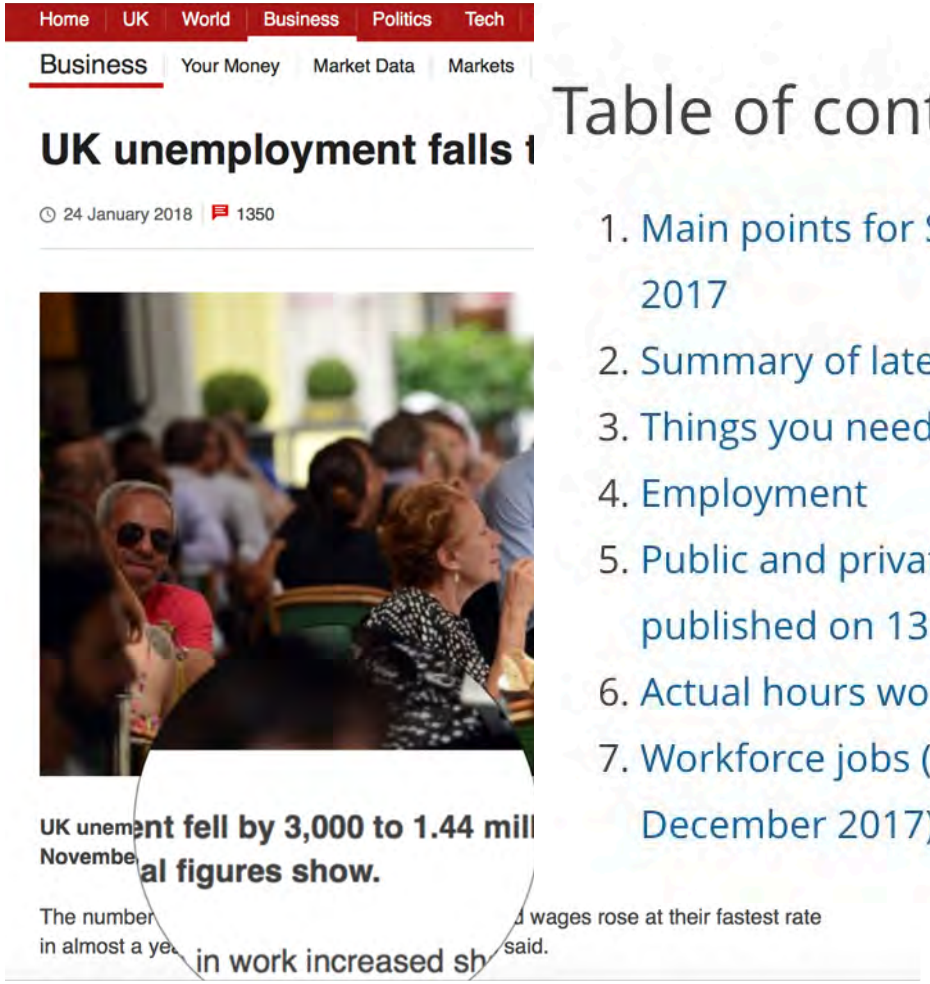
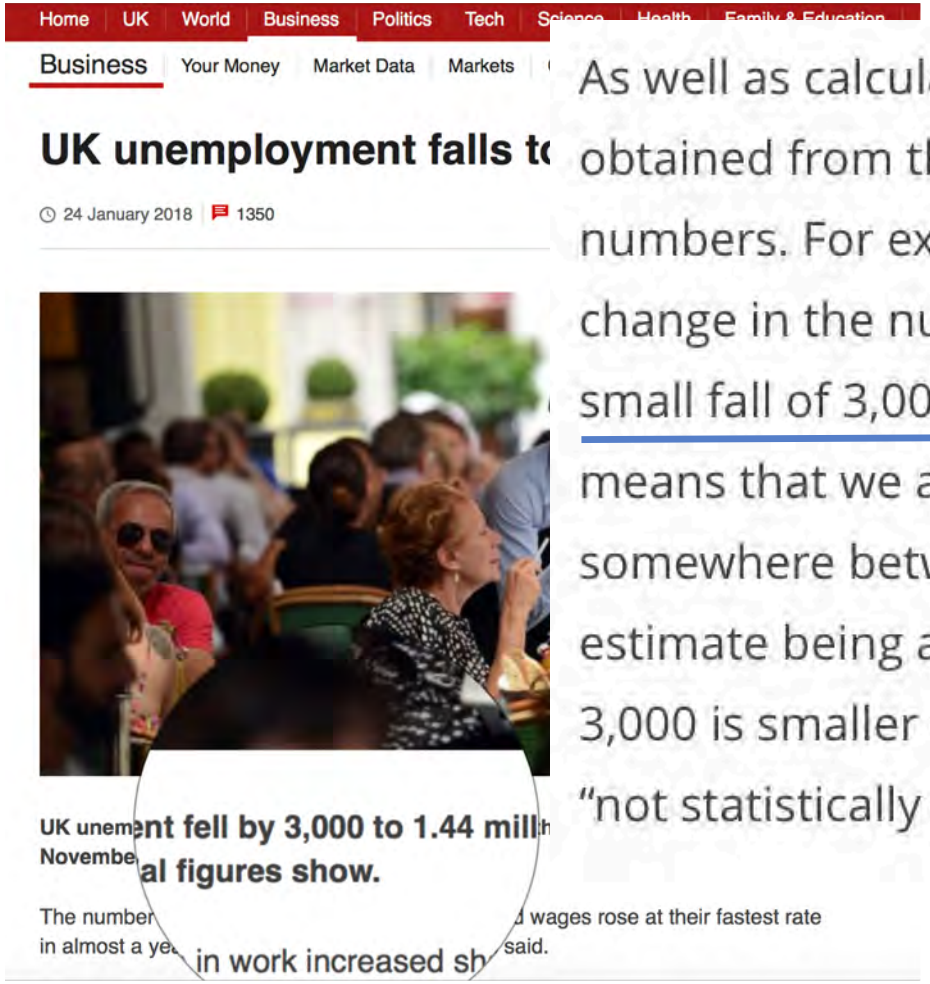


Table of contents

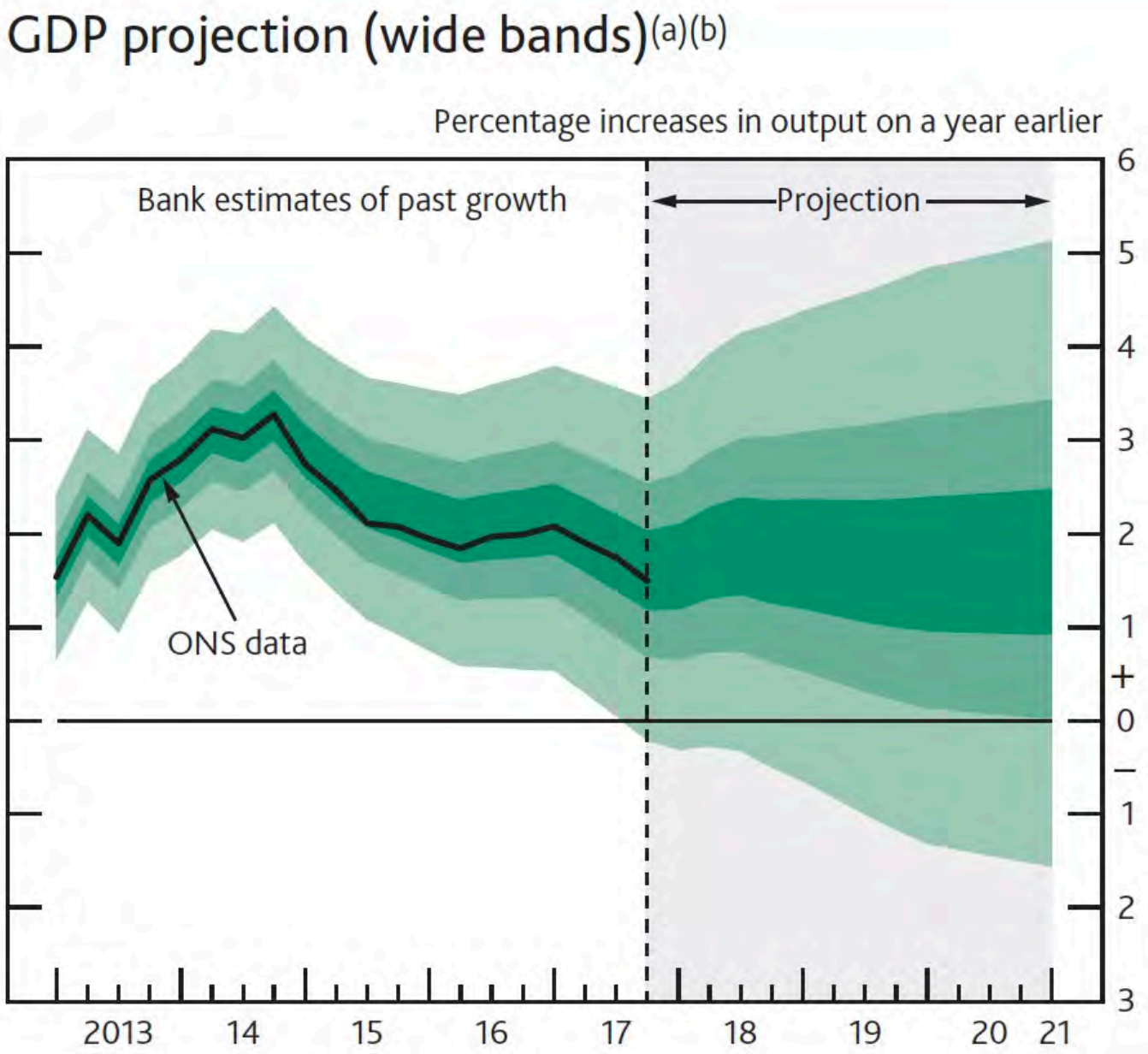
Uncertainty about statistics



As well as calculating precision measures around the numbers and rates obtained from the survey, we can also calculate them for changes in the numbers. For example, for September to November 2017, the estimated change in the number of unemployed people since June to August 2017 was a small fall of 3,000, with a 95% confidence interval of plus or minus 77,000. This means that we are 95% confident the actual change in unemployment was somewhere between an increase of 74,000 and a fall of 80,000, with the best estimate being a small fall of 3,000. As the estimated fall in unemployment of 3,000 is smaller than 77,000, the estimated fall in unemployment is said to be "not statistically significant".

February 2018 Inflation report

- ONS do not provide ‘error’ on GDP

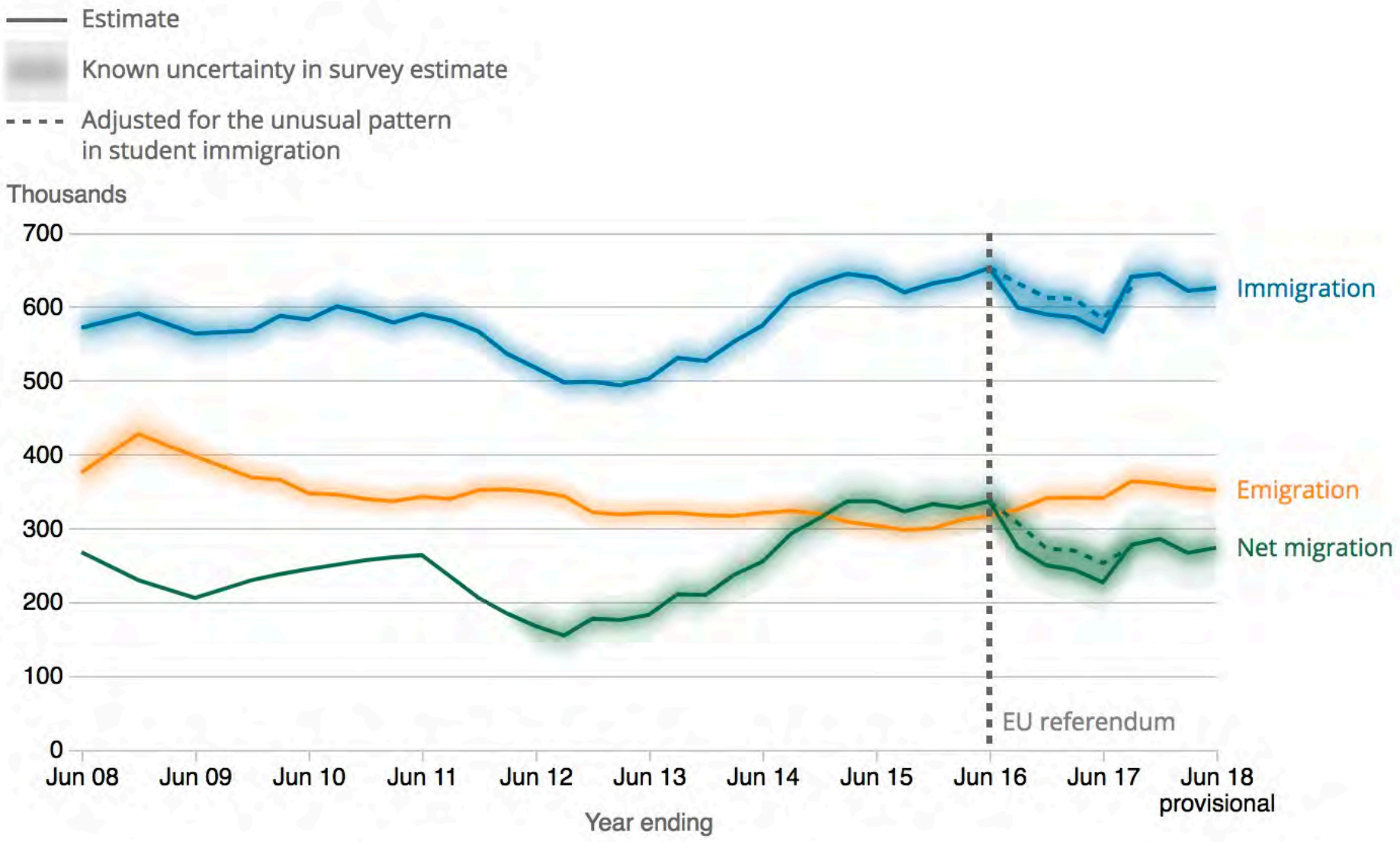


UK migration report November 2018

Only visualises sampling error

Quality issues as verbal caveats

Figure 1: Long-Term International Migration, UK, year ending June 2008 to year ending June 2018



Source: Long-Term International Migration, Office for National Statistics

Communicating uncertainty

- Our empirical research suggests that ‘confident uncertainty’ does not reduce trust in the source – audiences expect it.

- Relevance: future official statistics will be increasingly based on complex analysis of routine data

Fairness

There are many reasons for feeling an algorithm is ‘unfair’…..

What is YOUR heart age? Take this quick quiz to find out your stroke risk

A HEART Age Test is being promoted by The NHS and Public Health England. Here's how to check how healthy your vital organ is.

What's your heart age?

HOW HEALTHY IS YOUR HEART?

The Heart Age Test:

- Tells you your heart age compared to your real age
- Explains why it's important to know your blood pressure and cholesterol numbers
- Gives advice on how to reduce your heart age

START

Full terms and conditions can be read here

This tool is a collaboration between the NHS website, Public Health England, UCL and the British Heart Foundation. More information about partners

Full credits can be read here

HOW HEALTHY IS YOUR HEART?

PLEASE GIVE US SOME DETAILS ABOUT YOU

Date of birth 16 (Day) 08 (Month) 1953 (Year)

Gender Male Female

Why is this asked?

Ethnic group White

Why is this important?

Postcode CB5 8HL

Why is this being asked?

Do you have cardiovascular disease? Yes No

What is cardiovascular disease?

Do you smoke? No

YOUR HEART AGE IS ABOUT 69

Compared to a person of the same age, gender and ethnicity without raised risk factors.

On average, someone like you can expect to live to the age of **81** without having a heart attack or stroke.

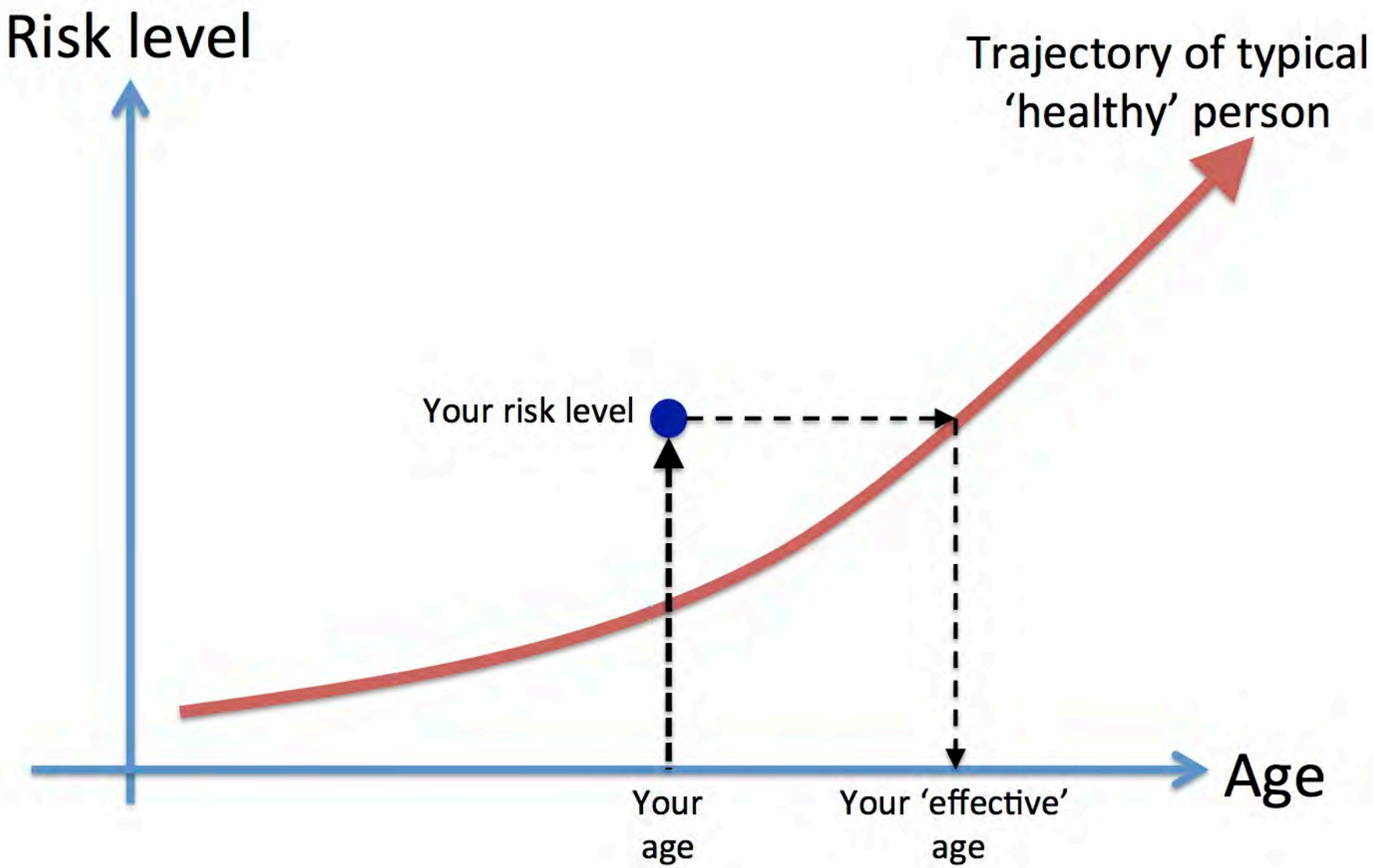
About your calculation

See how your heart age changes if you:

- Lose weight
- Lower cholesterol
- Reduce blood pressure

What is the ‘effective age’ of your organs?

- “*Lung age*”, “*brain age*”, etc etc
- Generic idea: what is the age of a ‘healthy’ person who has the same risk/function as you?


Risk level
Trajectory of typical 'healthy' person
Your risk level
Age
Your age
Your 'effective' age

Phase 3: RCT of 'heart age'

Effectiveness of the Heart Age tool for improving modifiable cardiovascular risk factors in a Southern European population: a randomized trial

Angel A Lopez-Gonzalez[1], Antoni Aguilo[2], Margalida Frontera[2], Miquel Bennasar-Veny[2], Irene Campos[1], Teofila Vicente-Herrero[3], Matias Tomas-Salva[4], Joan De Pedro-Gomez[2] and Pedro Tauler[2]

- > 3000 subjects individually randomised to
  - Heart Age calculator
  - Framingham risk score
  - Control
- At 12 months, reduction in risk score
  - Heart Age > Risk Score > Control

Comments from esteemed colleagues

- *‘What a load of c**p’ (Maths professor)*
- *‘It just annoys me that it says I have raised risk factors when I have none.’ (BBC producer)*
- *‘But what utter b******s this whole thing is.’ (General Practitioner)*
- *‘I could have programmed that in my sleep – just a load of random numbers designed to p**s people off.’ (Maths professor)*

What irritated people so much?

- Nearly everyone has increased heart age
- Exercise not in equation – seen as ‘not fair’

So who was responsible for all this?

COPYRIGHT and LICENSING

The JBS3 Cardiovascular Risk Assessment was created by the Understanding Uncertainty team of the University of Cambridge (UoC), working with the British Cardiovascular Society (BCS). The current version of the risk assessment was released in 2012 and is copyright the University of Cambridge. It is released under version 3 of the GNU General Public Licence. The source code, containing a copy of this license is published on GitHub. If the Tool is, subject to

- Reveals that we were responsible for adapting an existing model to provide Heart Age
- …. but used by 2.9 million people in 3 days

BMJ

RESEARCH

Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database

Julia Hippisley-Cox, professor of clinical epidemiology and general practice,[1] Carol Coupland, associate professor in medical statistics,[1] John Robson, senior lecturer general practice,[2] Peter Brindle, research and evaluation programme director[3]

- based on regression analysis (2.3 million people)
- but no question about physical fitness, as not in GP database

- now going to incorporate exercise……..

Table 2 | Adjusted hazard ratios* for cardiovascular disease for individual predictor variables in the derivation cohort of 2 343 759 patients

Variables	Adjusted hazard ratio (95% CI)	
	Women	Men
Body mass index†	1.32 (1.22 to 1.44)	1.54 (1.45 to 1.63)
Systolic blood pressure (per 20 mm Hg increase)	1.13 (1.12 to 1.14)	1.11 (1.10 to 1.12)
Total cholesterol:HDL cholesterol ratio (per unit increase)	1.17 (1.16 to 1.18)	1.18 (1.17 to 1.18)
Townsend score (per 5 unit increase)‡	1.13 (1.11 to 1.14)	1.06 (1.05 to 1.07)
Smoking status:		
Non-smoker	1.00	1.00
Former smoker	1.17 (1.14 to 1.21)	1.18 (1.16 to 1.21)
Light smoker (<10 cigarettes/day)	1.39 (1.33 to 1.45)	1.38 (1.34 to 1.43)
Moderate smoker (10-19/day)	1.57 (1.52 to 1.63)	1.55 (1.51 to 1.60)
Heavy smoker (≥20/day)	1.84 (1.77 to 1.91)	1.79 (1.74 to 1.84)
Ethnic group:		
White or not recorded	1.00	1.00
Indian	1.42 (1.28 to 1.58)	1.50 (1.38 to 1.63)
Pakistani	2.04 (1.78 to 2.34)	2.05 (1.84 to 2.28)
Bangladeshi	1.61 (1.30 to 1.98)	2.14 (1.85 to 2.46)
Other Asian	1.14 (0.92 to 1.4 0)	1.32 (1.12 to 1.56)
Caribbean	1.03 (0.91 to 1.16)	0.71 (0.63 to 0.81)
Black African	0.69 (0.54 to 0.89)	0.70 (0.56 to 0.86)
Chinese	0.77 (0.55 to 1.08)	0.79 (0.58 to 1.06)
Other	0.99 (0.85 to 1.16)	0.90 (0.78 to 1.04)
Clinical conditions:		
Family history of early coronary heart disease§	1.67 (1.63 to 1.71)	1.84 (1.80 to 1.88)
Type 2 diabetes	1.67 (1.60 to 1.73)	1.60 (1.55 to 1.66)
Treated hypertension	1.33 (1.30 to 1.36)	1.37 (1.34 to 1.40)
Rheumatoid arthritis	1.43 (1.35 to 1.53)	1.37 (1.26 to 1.50)
Atrial fibrillation	1.89 (1.78 to 2.01)	1.63 (1.54 to 1.72)
Chronic renal disease	1.67 (1.44 to 1.95)	1.59 (1.39 to 1.83)

Conclusions

- Need to demonstrate the trustworthiness of claims both
  - **by** an algorithm
  - **about** an algorithm
- Phased evaluation of quality and impact
- Can formally rank algorithms
- Explanation in multiple forms and levels
- Confident communication of uncertainty
- Many reasons why people might feel an algorithm was unfair
- Basic statistical science might help!

Thanks to …

Titanic

- Maria Skoularidou

Predict

- George Farmer, Alex Freeman, Gabriel Recchia, Paul Pharoah, Jem Rashbass,

Migration

- Sarah Dryhurst

Heart Age

- Mike Pearson

Google Scholar

label:statistics

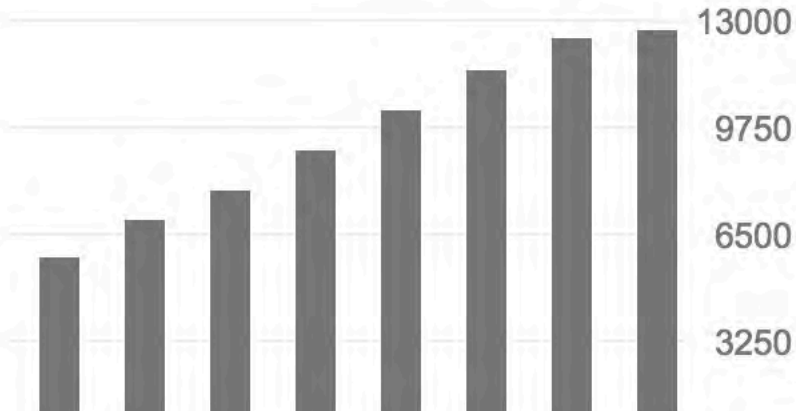
14

Leo Breiman 1928-2005

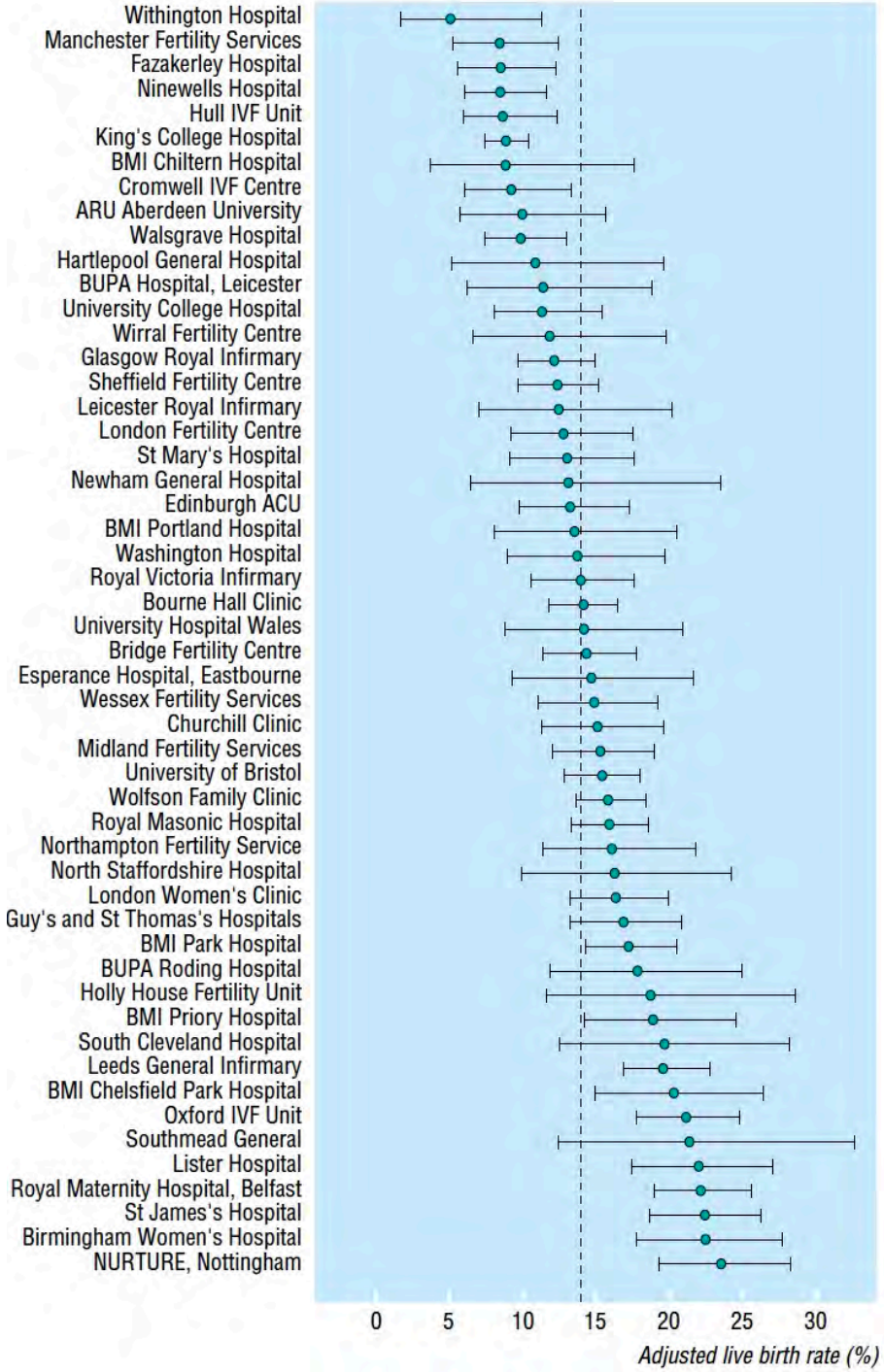
Professor of Statistics, UC Berkeley

Verified email at stat.berkeley.edu

Data Analysis Statistics Machine Learning

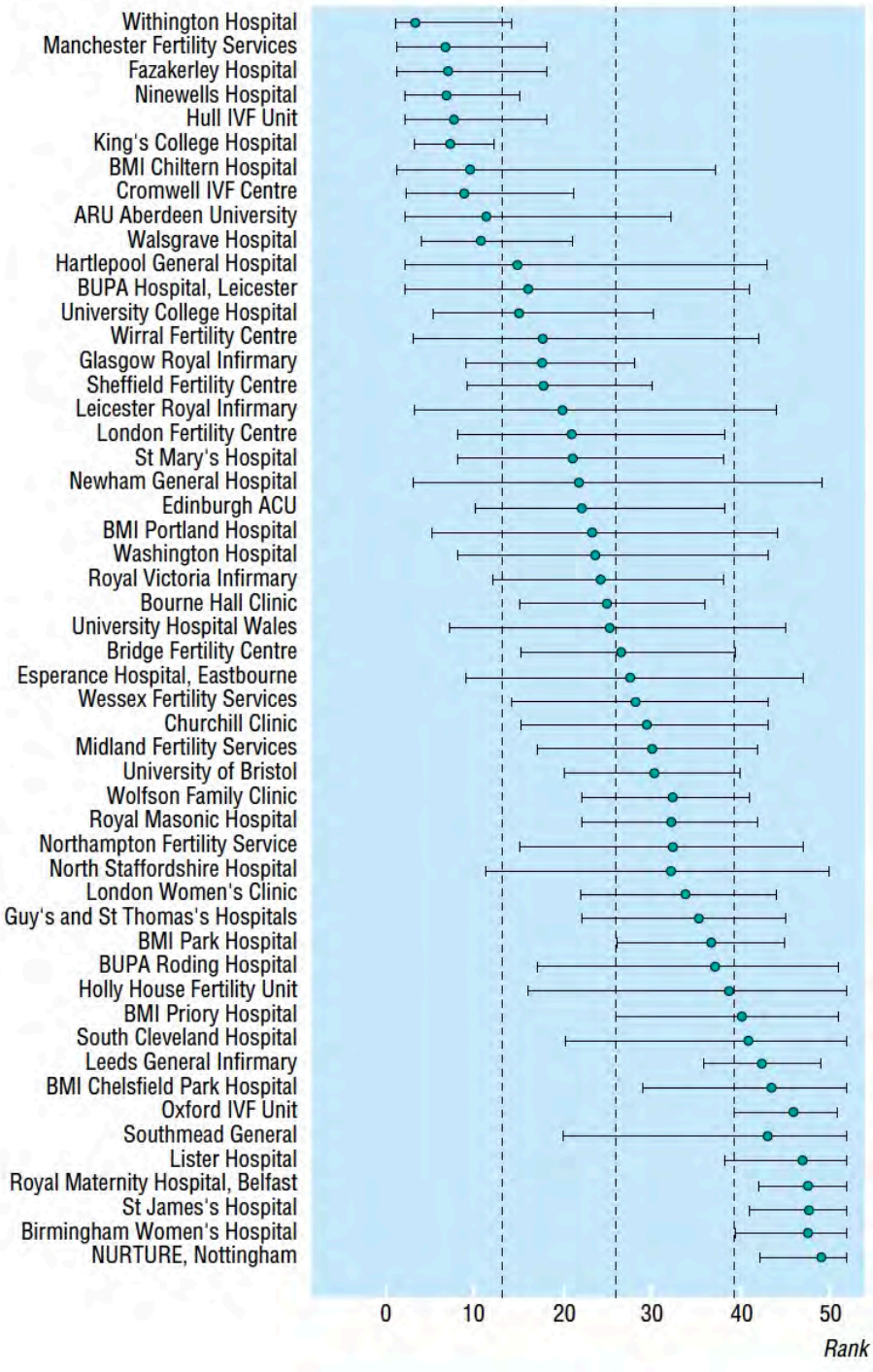


Cited by 120297



- Comparing success rates of IVF clinics
- League table is misleading
- Simulate set of ‘success rates’ from their distributions
- Rank each set
- Repeat say 1,000 times
- Get distribution over ranks of institutions

Marshall et al, BMJ, 1998



Tipping points – what is the crucial item of evidence?

The mystery of the lost star

A statistical detective story

In July 2005 the Healthcare Commission released its annual "star ratings" for English National Health Service (NHS) trusts[1], in which acute or specialist hospitals, mental health services, ambulance services and primary care trusts were each given 0, 1, 2 or 3 stars. There was some surprise that the Cambridge University Hospitals NHS Foundation Trust (better known as Addenbrooke's Hospital) dropped from the 3 stars obtained in 2004 to 2 stars. **David Spiegelhalter** investigated.

"Out of all the trusts in England, it is likely there will be at least

Unfortunately we only just missed out on three stars because we did not perform so well in the areas of delayed discharges and cancelled operations despite making progress over the past year

Malcolm Stamp

Chief Executive of Cambridge Addenbrooke's Hospital

‘Star rating’ based on (very) complex hierarchical algorithm mixing scores and rules

	Balanced scorecard						
Key targets	BS = 0	BS = 1	BS = 2	BS = 3	BS = 4	BS = 5	BS = 6
Fail: > 12 penalty points	0 stars	0 stars	0 stars	0 stars	0 stars	0 stars	0 stars
Moderate Fail: 7–12	0 stars	1 star	1 star	1 star	1 star	1 star	1 star
Borderline: 3–6	1 star	1 star	1 star	1 star	2 stars	2 stars	2 stars
Pass: ≤ 2	1 star	2 stars	2 stars	2 stars	2 stars	3 stars	3 stars

Table 5. Rules for obtaining a final star rating based on key targets and balanced scorecard, for acute and specialist trusts

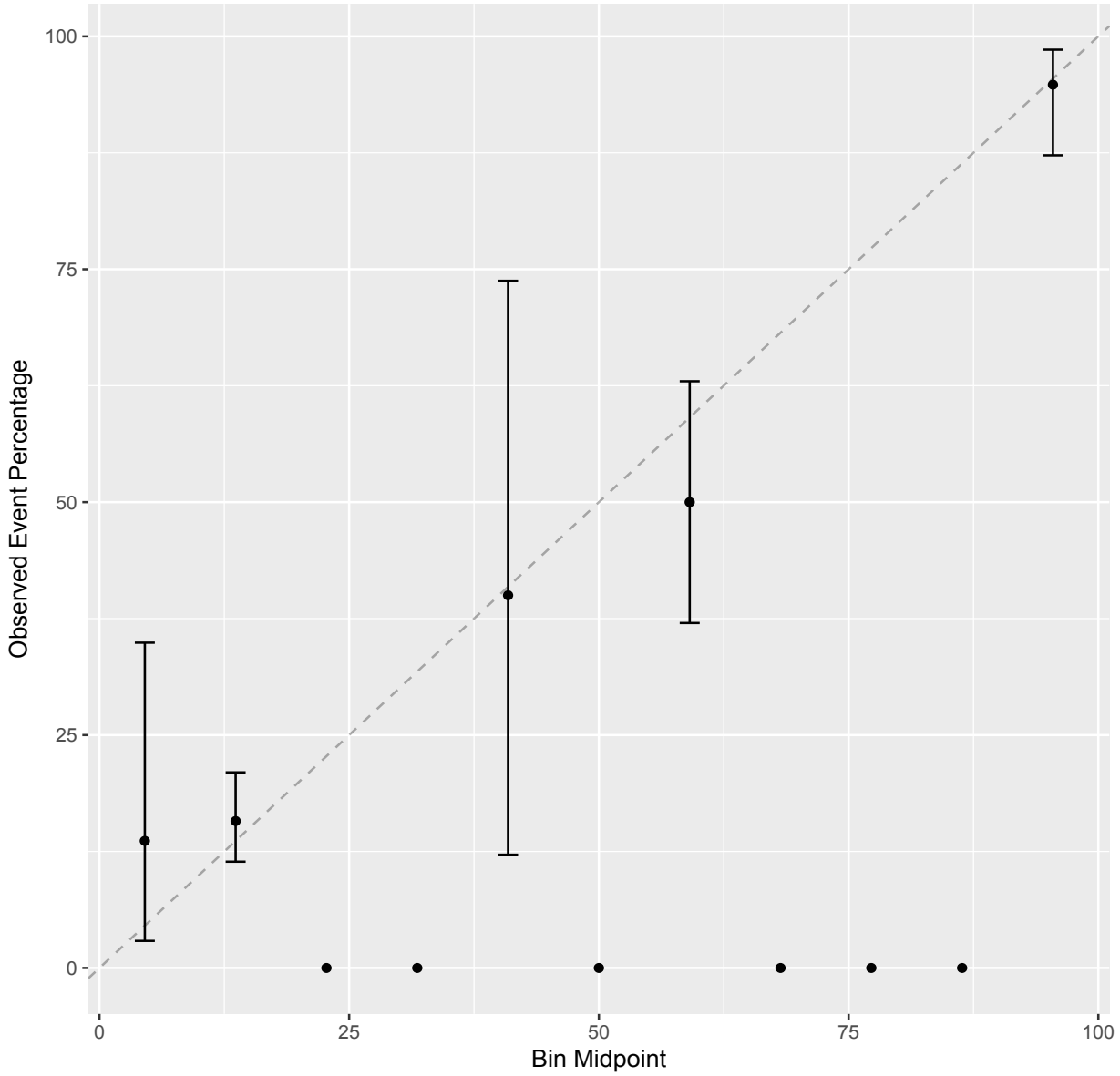
After a lot of manual work, found the crucial piece of evidence that tipped Addenbrooke’s …

If just four more junior doctors out of 417 had complied with the 'New Deal on working hours', then…

- Addenbrooke's rate on this indicator would have been 395/417 = 94.7% compliance.
- Rounded to 95%, giving 1 point for *Junior Doctors' Hours*
- Gives a band score of 4 for the *Workforce Indicator*
- Brings total band score to 21 in the *Capability and Capacity* focus area
- Gives a focus score of 2.
- The Balanced Scorecard would be 5
- Combined with the key targets, would have given Addenbrooke's 3 stars!

Probabilities should be well- calibrated

- Simple classification tree for Titanic problem is well-calibrated
- The probabilities mean what they say - they are trustworthy.



A simple test for calibration

- $X_i = 1$ if event occurs, $X_i = 0$ otherwise, p_i is probability given to event occurring.
- Mean squared error = mean Brier score = $\overline{B} = \frac{1}{n}\sum_i (X_i - p_i)^2$.
- Since $X_i^2 = X_i$, we can rewrite mean Brier score as

$$\begin{aligned} \overline{B} &= \frac{1}{n}\sum_i (X_i - p_i)(1 - 2p_i) + \frac{1}{n}\sum_i p_i(1 - p_i). \\ &= \text{'lack of calibration'} + \text{'separation'} \end{aligned}$$

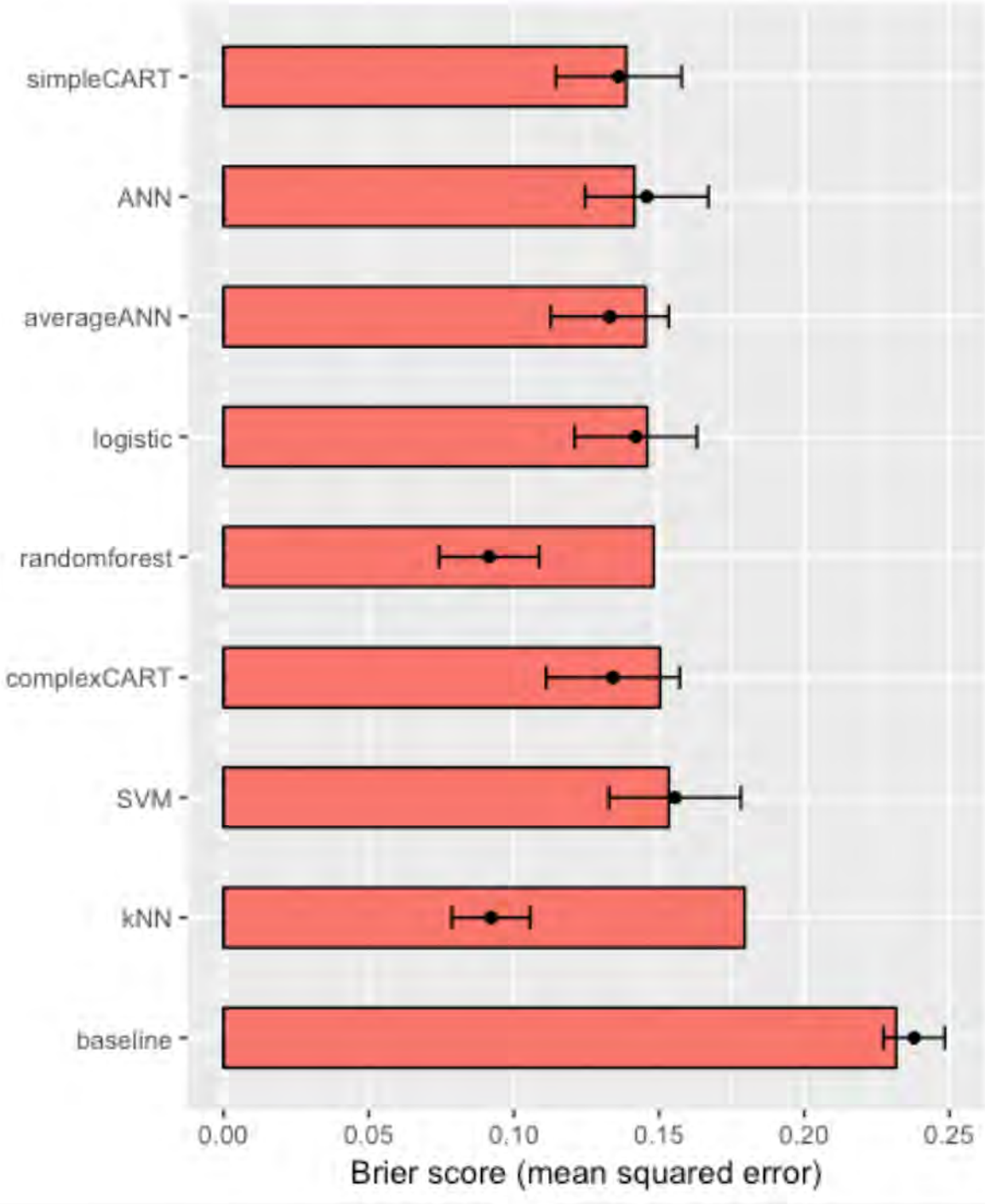
- Under null hypothesis of perfect calibration, $E_0(X_i) = p_i$, $V_0(X_i) = p_i(1 - p_i)$
- So

$$\begin{aligned} E_0(\overline{B}) &= \frac{1}{n}\sum_i p_i(1 - p_i) \\ V_0(\overline{B}) &= \frac{1}{n^2}\sum_i \left[p_i(1 - p_i)(1 - 2p_i)^2\right] \end{aligned}$$

- $Z = (\overline{B} - E_0(\overline{B}))/\sqrt{V_0(\overline{B})}$ is a popular standardised global test of calibration

DJS, SIM, 1986

- Expected mean Brier score, if perfectly calibrated
- randomforest and kNN are very overconfident
- ’baseline’ is a bit cautious



Age at diagnosis – 65 +

Age must be between 25 and 85

Post Menopausal? Yes | No | Unknown

ER status Positive | Negative

HER2 status Positive | Negative | Unknown

Ki-67 status Positive | Negative | Unknown

Positive means more than 10%

Tumour size (mm) – 20 +

Tumour grade 1 | 2 | 3

Detected by Screening | Symptoms | Unknown

Detected as part of a preventive screening programme

Positive nodes – 2 +

Micrometastases Yes | No | Unknown

Enabled when positive nodes is zero

Treatment Options

Hormone Therapy No | Yes

Available when ER-status is positive

Chemotherapy None | 2nd gen | 3rd gen

Trastuzumab No | Yes

Available with chemotherapy when HER2 status is positive

Bisphosphonates No | Yes

Available for post-menopausal women

Print

Results

Table | Curves | Chart | Texts | Icons

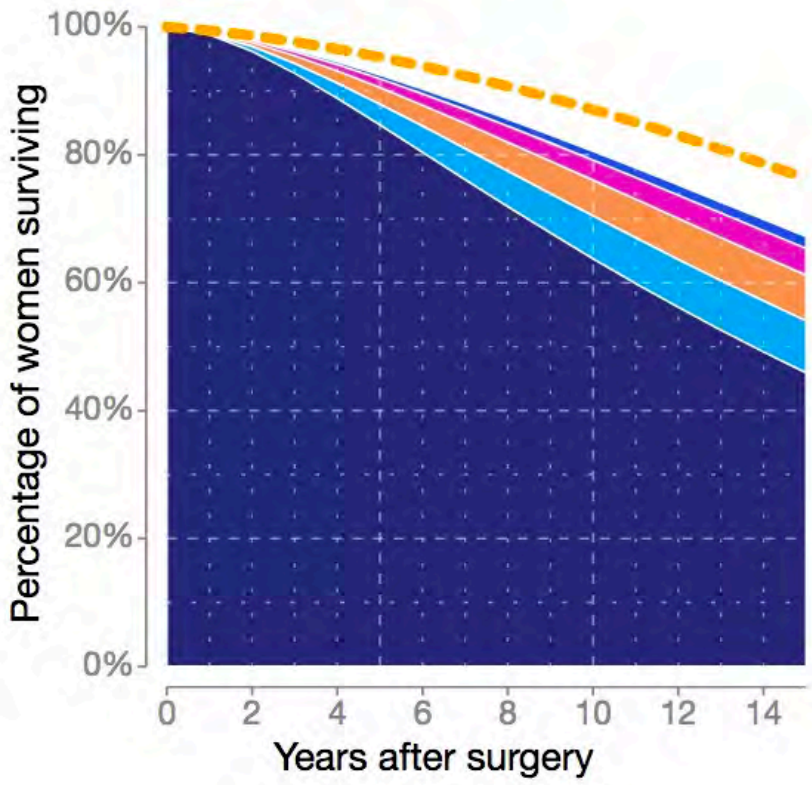
These results are for women who have already had surgery. This table shows the percentage of women who survive at least 5 | 10 | 15 years after surgery, based on the information you have provided.

Treatment	Additional Benefit	Overall Survival %
Surgery only	-	46%
+ Hormone therapy	8%	54%
+ Chemotherapy	7%	61%
+ Trastuzumab	4%	65%
+ Bisphosphonates	2%	67%

If death from breast cancer were excluded, 76% would survive at least 15 years.

Table | **Curves** | Chart | Texts | Icons

These results are for women who have already had surgery. This graph shows the percentage of women surviving up to 15 years. These results are based on the inputs and treatments you selected.



Survival rate excluding deaths from breast cancer.

Additional benefit of bisphosphonates

Additional benefit of trastuzumab

Additional benefit of chemotherapy

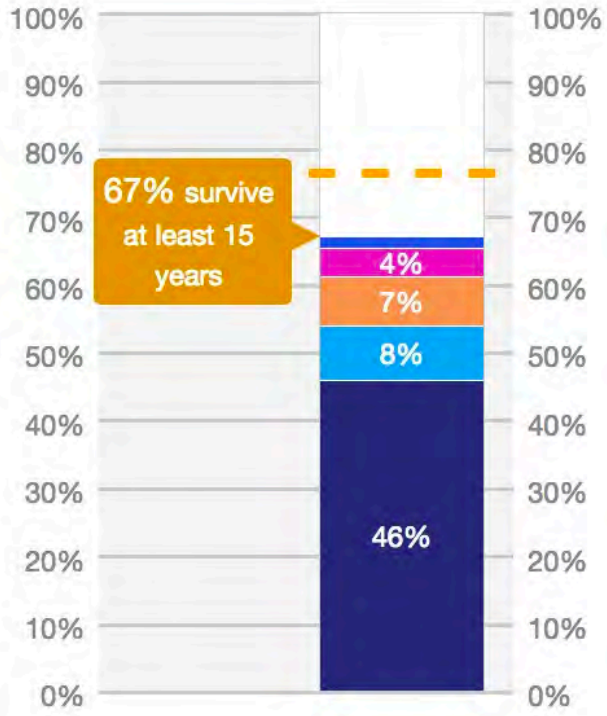
Additional benefit of hormone therapy

Surgery only

Table | Curves | **Chart** | Texts | Icons

These results are for women who have already had surgery. Based on the inputs and treatments you selected, this graph shows the percentage of women surviving at least 5 | 10 | **15** years after surgery.

Overall Survival



Survival rate excluding deaths from breast cancer.

Additional benefit of bisphosphonates is 2% at 15 years.

Additional benefit of trastuzumab is 4% at 15 years.

Additional benefit of chemotherapy is 7% at 15 years.

Additional benefit of hormone therapy is 8% at 15 years.

Surgery only survival is 46% at 15 years.

Table | Curves | Chart | **Texts** | Icons

These results are for women who have already had surgery. This display shows the outcomes for 100 women based on the inputs and treatments you have selected 5 | 10 | **15** years after surgery.

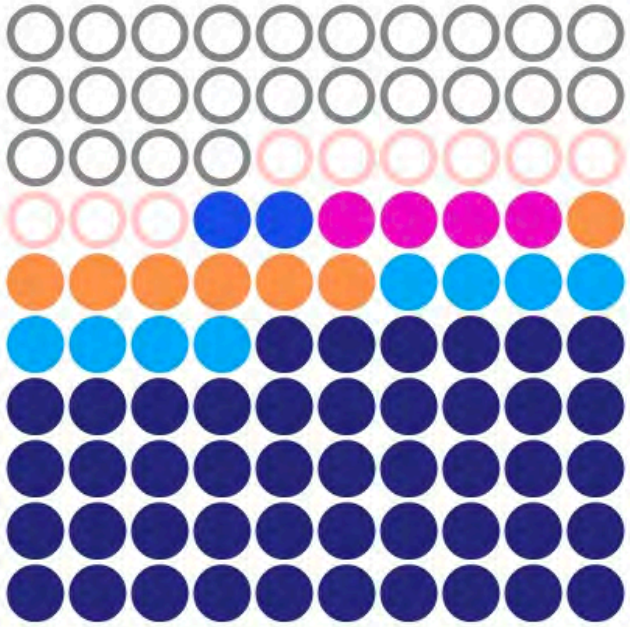
46 out of 100 women treated with surgery only are alive at 15 years.

- 54 out of 100 women treated with hormone therapy are alive (an extra 8).
- 61 out of 100 women treated with hormone therapy, and chemotherapy are alive (an extra 15).
- 65 out of 100 women treated with hormone therapy, chemotherapy, and trastuzumab are alive (an extra 19).
- 67 out of 100 women treated with hormone therapy, chemotherapy, trastuzumab, and bisphosphonates are alive (an extra 21).

Of the women who would not survive, 24 would die due to causes not related to breast cancer.

Table | Curves | Chart | Texts | **Icons**

These results are for women who have already had surgery. This display shows the outcomes for 100 women based on the inputs and treatments you have selected 5 | 10 | **15** years after surgery.



- 24 deaths due to other causes
- 9 breast cancer related deaths
- 2 extra survivors due to bisphosphonates
- 4 extra survivors due to trastuzumab
- 7 extra survivors due to chemotherapy
- 8 extra survivors due to hormone therapy
- 46 survivors with surgery alone

Uncertainty?

Table | Curves | Chart | Texts | Icons

These results are for women who have already had surgery. This table shows the percentage of women who survive at least 5 | 10 | 15 years after surgery, based on the information you have provided.

Treatment	Additional Benefit	Overall Survival %
Surgery only	-	52%
+ Hormone therapy	6.9% (4.0% – 8.6%)	59%
+ Chemotherapy	5.9% (4.3% – 7.2%)	64%
+ Trastuzumab	3.4% (2.4% – 4.7%)	68%
+ Bisphosphonates	1.5% (0.5% – 2.2%)	69%

If death from breast cancer were excluded, 76% would survive at least 15 years.

Show ranges? Yes | No

Assumed treatment effects

Table 3: Treatment Risk-factor coefficients

Treatment	log(*RR*)	approx se of log(*RR*)	Hazard ratio Relative risk	Source
hormone therapy up to 10 years (if ER+)	-0.386	0.08	0.68	Early Breast Cancer Trialists' Collaborative Group (2011) p777
trastuzumab (if HER2+)	-0.357	0.08	0.70	unpublished meta-analysis of 4 large randomised trials
Bisphosphonates (if post-menopausal)	-0.198	0.06	0.82	Early Breast Cancer Trialists' Collaborative Group (2015)
2nd gen chemotherapy	-0.248	0.12	0.78	Early Breast Cancer Trialists' Collaborative Group (2012)
3rd gen chemotherapy	-0.446	0.13	0.64	Early Breast Cancer Trialists' Collaborative Group (2012)

ben goldacre
@bengoldacre

Following

This test is ridiculous. Try it. PHE's tool tells a woman in her 30s that her heart age is older than her real age because she's not had her cholesterol done. And tells her to get her cholesterol done by GP. There is no evidence for this, pointless excess GP workload...

Public Health England @PHE_uk
Did you know, having a heart age older than your actual age means you are at a greater risk of having a heart attack or stroke? Check your heart age using our #HeartAgeTest: bit.ly/2v1vL2d

11:39 PM - 4 Sep 2018

NHS heart check tool attacked by doctors for 'sending healthy 30-year-olds to GP needlessly'

'This test is ridiculous'

Alex Matthews-King Health Correspondent | Friday 7 September 2018 15:45 | 2 comments