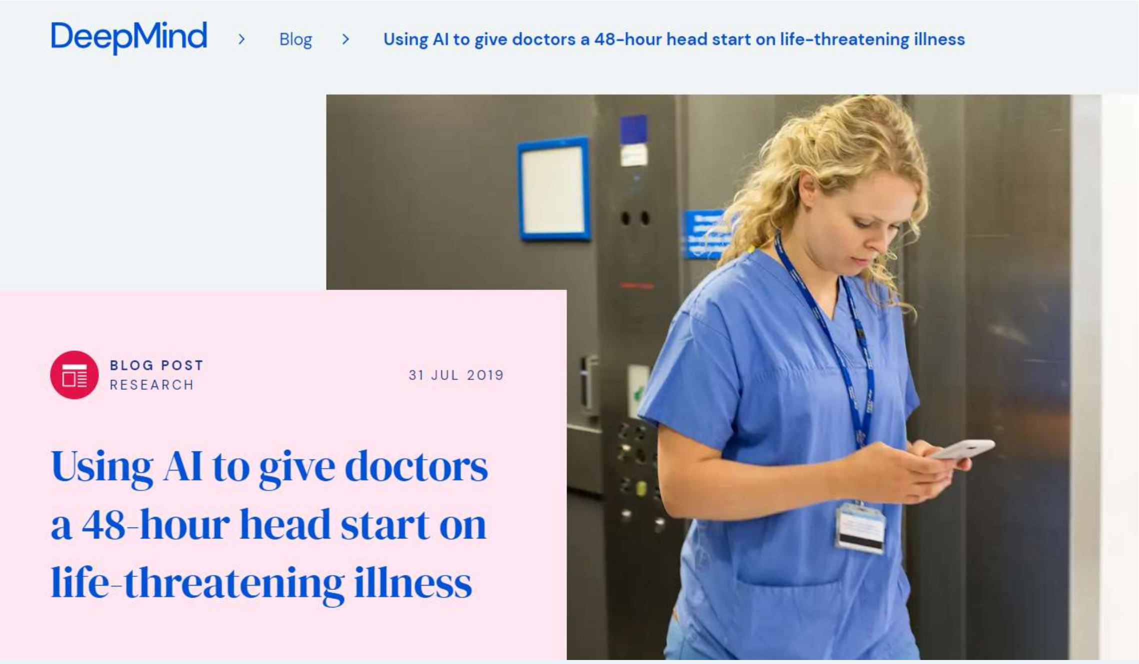
AMDS
Clinical Development and Analytics

Novartis benchmarking initiative: making sense of AI

Mark Baillie (with Conor Moloney & Janice Branson)
BBS, Basel
November 01, 2019

NOVARTIS | Reimagining Medicine



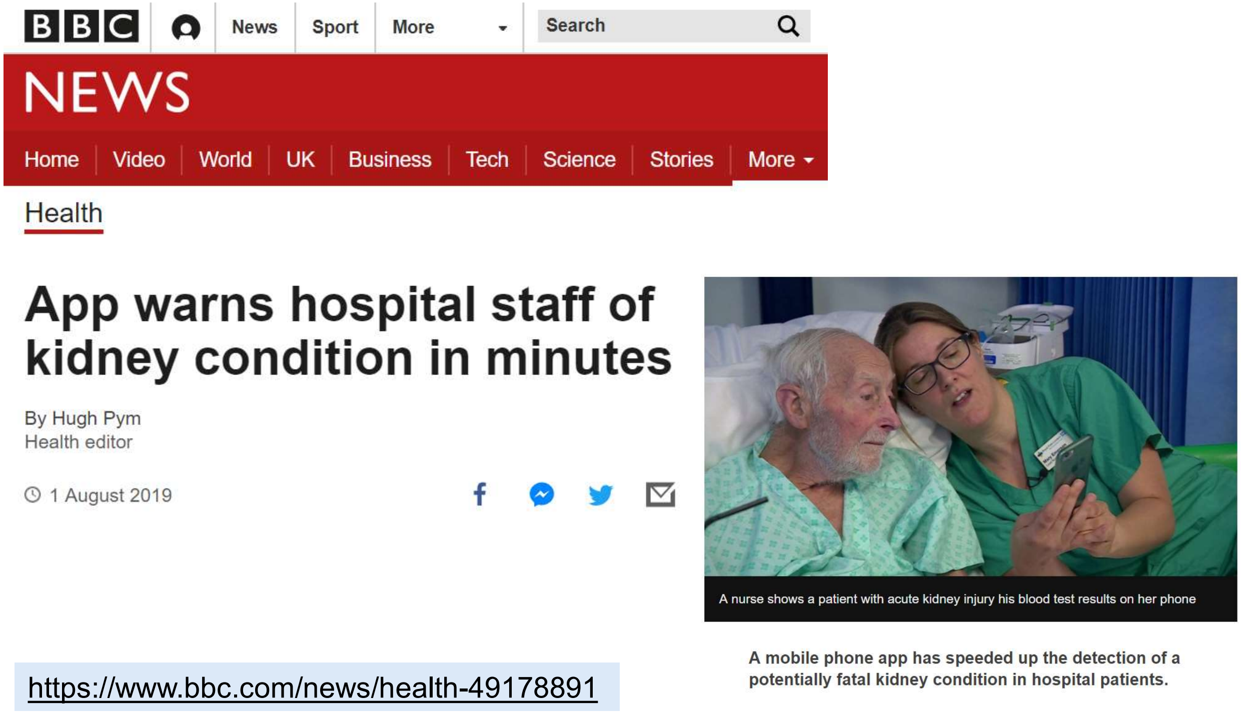
DeepMind › Blog › Using AI to give doctors a 48-hour head start on life-threatening illness

BLOG POST
RESEARCH

31 JUL 2019

Using AI to give doctors a 48-hour head start on life-threatening illness

https://deepmind.com/blog/article/predicting-patient-deterioration



App warns hospital staff of kidney condition in minutes

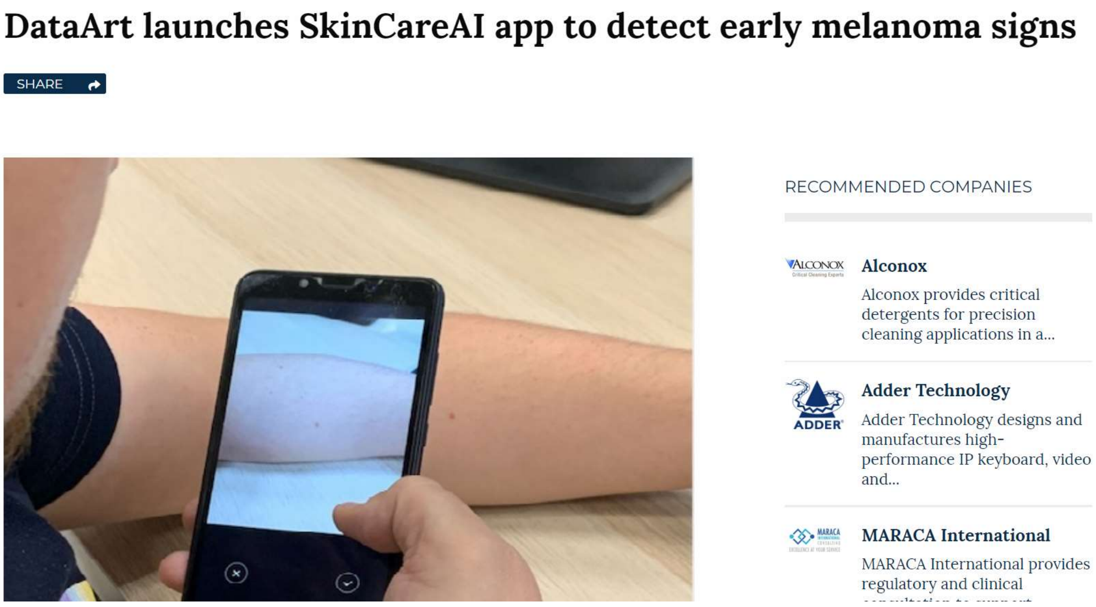
By Hugh Pym
Health editor

1 August 2019

A nurse shows a patient with acute kidney injury his blood test results on her phone

A mobile phone app has speeded up the detection of a potentially fatal kidney condition in hospital patients.

https://www.bbc.com/news/health-49178891



DataArt launches SkinCareAI app to detect early melanoma signs

RECOMMENDED COMPANIES

Alconox
Alconox provides critical detergents for precision cleaning applications in a...

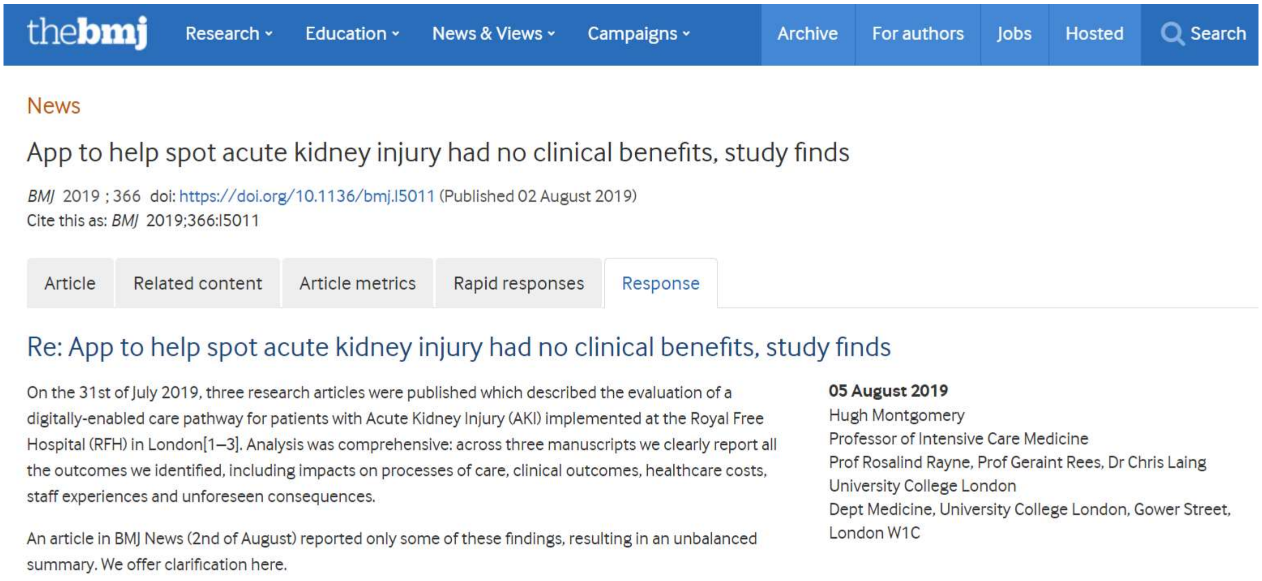
Adder Technology
Adder Technology designs and manufactures high-performance IP keyboard, video and...

MARACA International
MARACA International provides regulatory and clinical consultation to support

https://www.medicaldevice-network.com/news/dataart-launches-skincareai-app/

How do we know it works?

NOVARTIS | Reimagining Medicine



News

App to help spot acute kidney injury had no clinical benefits, study finds

BMJ 2019 ; 366 doi: https://doi.org/10.1136/bmj.l5011 (Published 02 August 2019)
Cite this as: *BMJ* 2019;366:l5011

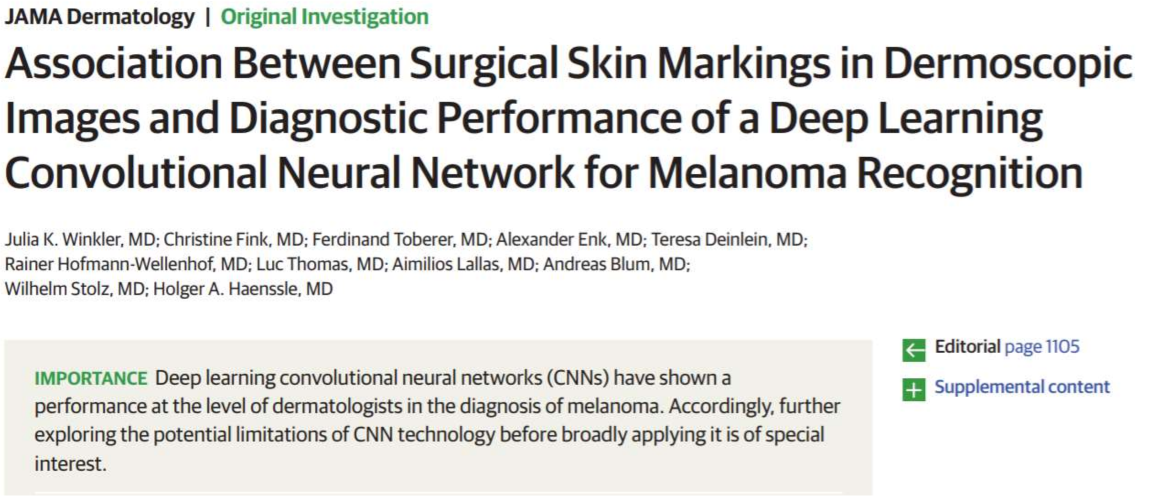
Re: App to help spot acute kidney injury had no clinical benefits, study finds

On the 31st of July 2019, three research articles were published which described the evaluation of a digitally-enabled care pathway for patients with Acute Kidney Injury (AKI) implemented at the Royal Free Hospital (RFH) in London[1–3]. Analysis was comprehensive: across three manuscripts we clearly report all the outcomes we identified, including impacts on processes of care, clinical outcomes, healthcare costs, staff experiences and unforeseen consequences.

An article in BMJ News (2nd of August) reported only some of these findings, resulting in an unbalanced summary. We offer clarification here.

05 August 2019
Hugh Montgomery
Professor of Intensive Care Medicine
Prof Rosalind Rayne, Prof Geraint Rees, Dr Chris Laing
University College London
Dept Medicine, University College London, Gower Street, London W1C

https://www.bmj.com/content/366/bmj.l5011/rr



JAMA Dermatology | Original Investigation

Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition

Julia K. Winkler, MD; Christine Fink, MD; Ferdinand Toberer, MD; Alexander Enk, MD; Teresa Deinlein, MD; Rainer Hofmann-Wellenhof, MD; Luc Thomas, MD; Aimilios Lallas, MD; Andreas Blum, MD; Wilhelm Stolz, MD; Holger A. Haenssle, MD

IMPORTANCE Deep learning convolutional neural networks (CNNs) have shown a performance at the level of dermatologists in the diagnosis of melanoma. Accordingly, further exploring the potential limitations of CNN technology before broadly applying it is of special interest.

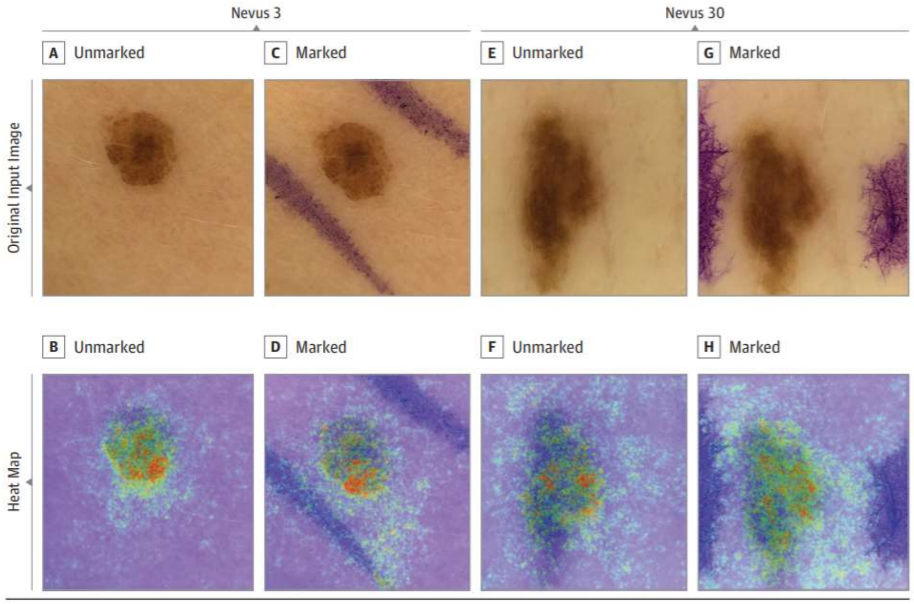
Editorial page 1105

Supplemental content

https://jamanetwork.com/journals/jamadermatology/fullarticle/2740808

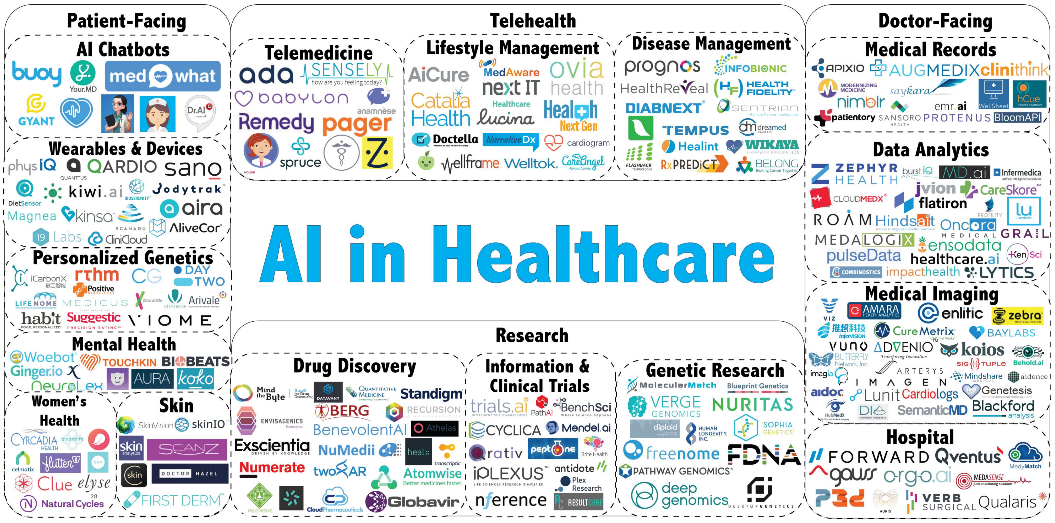
NOVARTIS | Reimagining Medicine

Figure 3. Heat Maps of 2 Benign Nevi With Unchanged Melanoma Probability Scores After Addition of In Vivo Skin Markings



ining Medicine

How do we know it works?



https://techburst.io/ai-in-healthcare-industry-landscape-c433829b320c

How do we systematically evaluate?

- A standard process for benchmarking:
  - Common task framework
  - Reporting guidelines
- This process aims to:
  - **evaluate** and **compare** «innovtation» on relevant tasks
  - **de-risk** engagement
  - **reduce** internal resources for evaluation

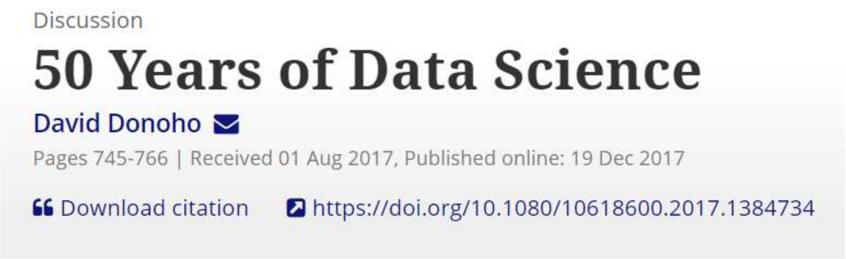
NOVARTIS | Reimagining Medicine

Why benchmarking?

- Machine learning, statistical learning, AI, etc. are experimental fields
- Most new methodological improvements are assessed using standard benchmark datasets – "the common task framework"
- Using tasks and benchmarks developed at Novartis will enable us to better understand claims on effectiveness
- There is also a real need to develop new benchmarks which reflect real world problems in the biomedical space to advance understanding.

NOVARTIS | Reimagining Medicine

Common task framework



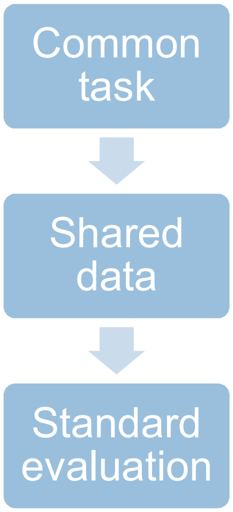
Discussion

50 Years of Data Science

David Donoho

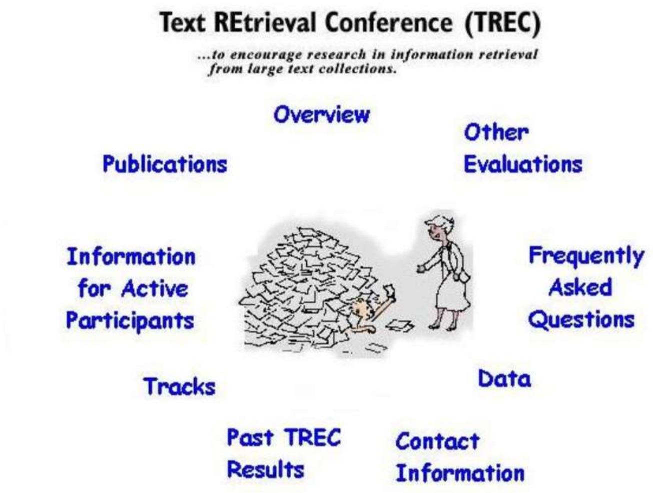
Pages 745-766 | Received 01 Aug 2017, Published online: 19 Dec 2017

Download citation https://doi.org/10.1080/10618600.2017.1384734



https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1384734

Common task framework

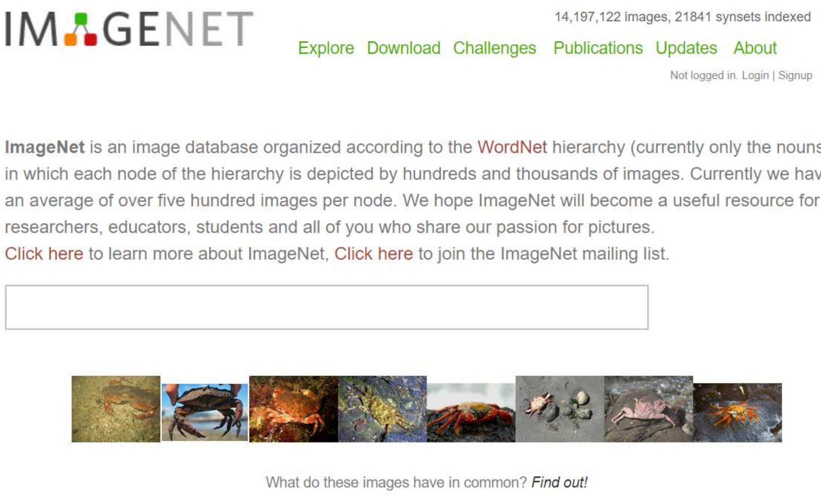


Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

https://trec.nist.gov/

Common task framework



IMAGENET
14,197,122 images, 21841 synsets indexed
Explore Download Challenges Publications Updates About
Not logged in. Login | Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.

Click here to learn more about ImageNet, Click here to join the ImageNet mailing list.

What do these images have in common? *Find out!*

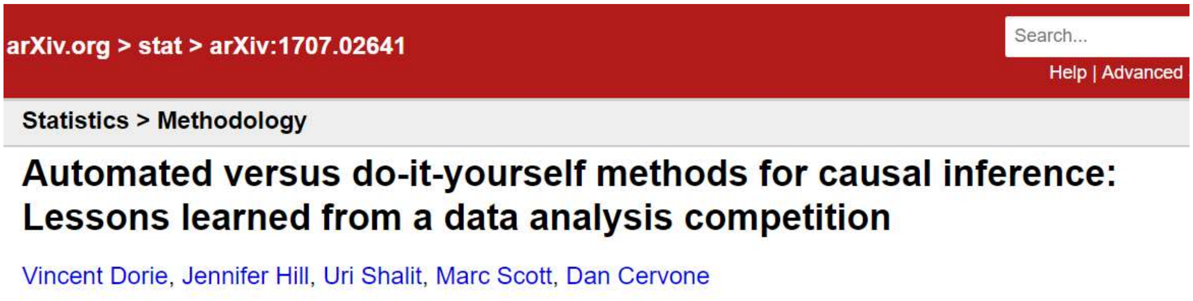
http://www.image-net.org/

Common task framework

A community platform for NGS assay evaluation and regulatory science exploration.

https://precision.fda.gov

Common task framework



arXiv.org > stat > arXiv:1707.02641

Statistics > Methodology

Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, Dan Cervone

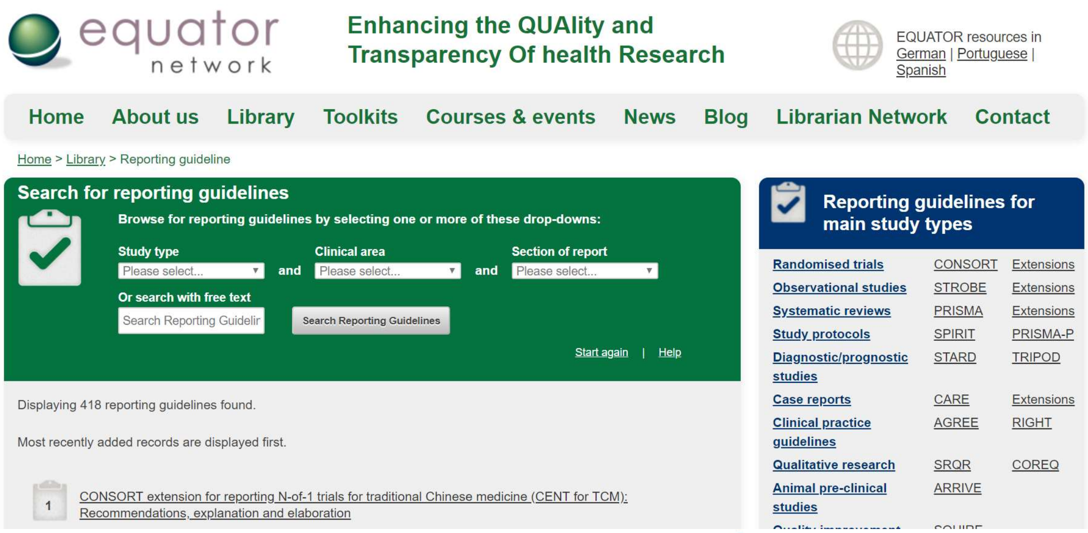
https://arxiv.org/abs/1707.02641

An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question.

- John Tukey

https://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711

Reporting guidelines



Enhancing the QUAlity and Transparency Of health Research

EQUATOR resources in German | Portuguese | Spanish

Home > Library > Reporting guideline

Search for reporting guidelines

Browse for reporting guidelines by selecting one or more of these drop-downs:

Study type: Please select... and Clinical area: Please select... and Section of report: Please select...

Or search with free text: Search Reporting Guidelir [Search Reporting Guidelines]

Start again | Help

Displaying 418 reporting guidelines found.

Most recently added records are displayed first.

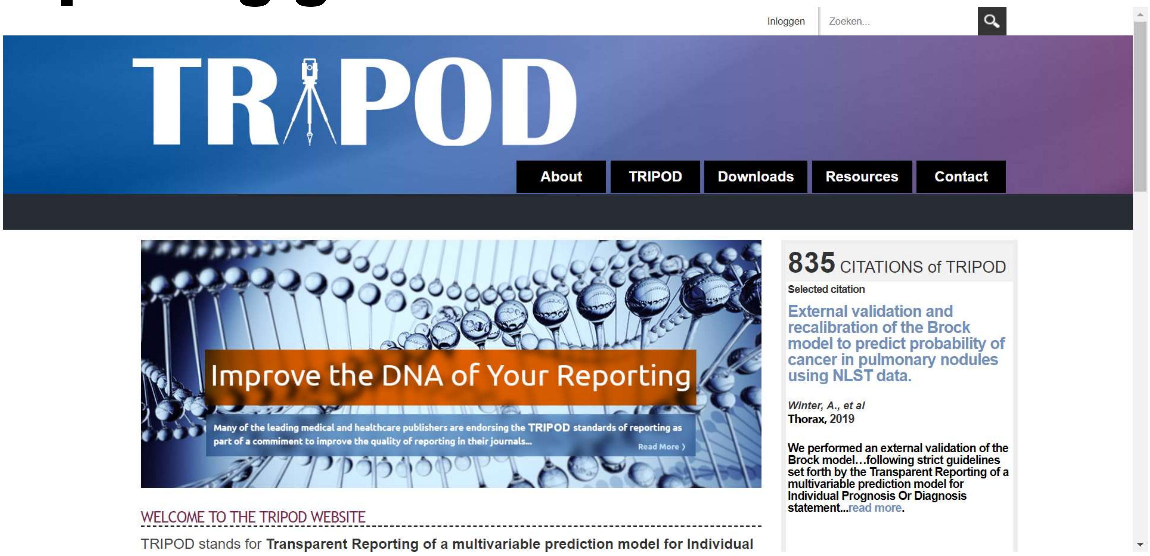
1 CONSORT extension for reporting N-of-1 trials for traditional Chinese medicine (CENT for TCM): Recommendations, explanation and elaboration

Reporting guidelines for main study types

Randomised trials	CONSORT	Extensions
Observational studies	STROBE	Extensions
Systematic reviews	PRISMA	Extensions
Study protocols	SPIRIT	PRISMA-P
Diagnostic/prognostic studies	STARD	TRIPOD
Case reports	CARE	Extensions
Clinical practice guidelines	AGREE	RIGHT
Qualitative research	SRQR	COREQ
Animal pre-clinical studies	ARRIVE	

https://www.equator-network.org/reporting-guidelines/

Reporting guidelines



Improve the DNA of Your Reporting

Many of the leading medical and healthcare publishers are endorsing the **TRIPOD** standards of reporting as part of a commiment to improve the quality of reporting in their journals...

Read More ›

WELCOME TO THE TRIPOD WEBSITE

TRIPOD stands for **Transparent Reporting of a multivariable prediction model for Individual**

835 CITATIONS of TRIPOD

Selected citation

External validation and recalibration of the Brock model to predict probability of cancer in pulmonary nodules using NLST data.

Winter, A., et al
Thorax, 2019

We performed an external validation of the Brock model…following strict guidelines set forth by the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis statement...read more.

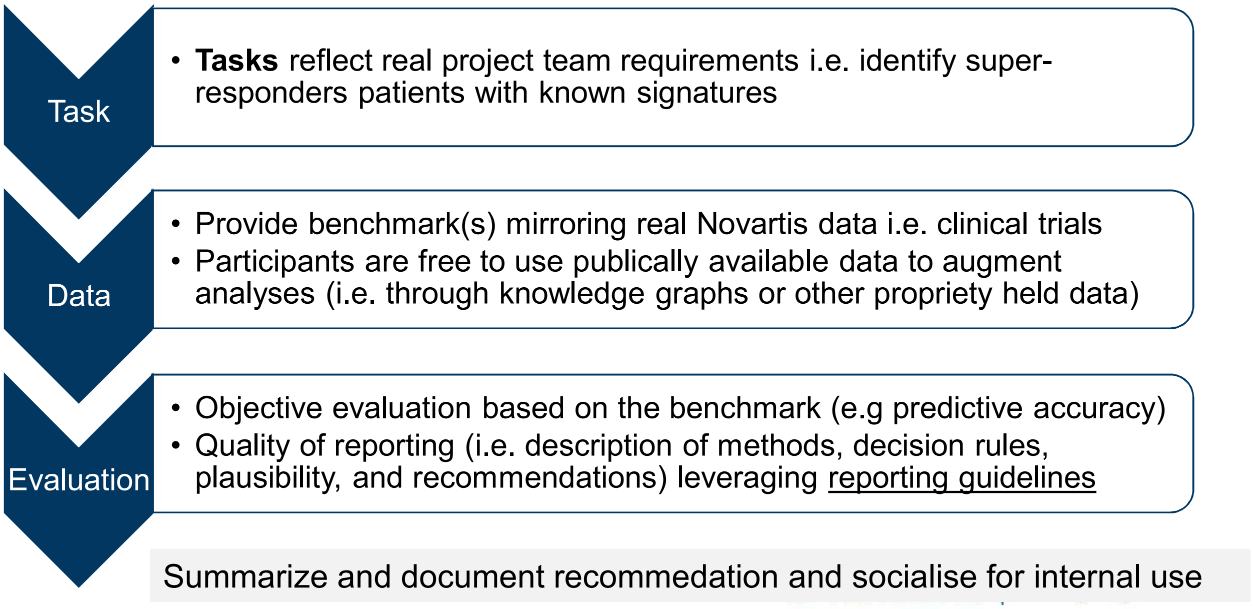
https://www.tripod-statement.org/

Why reporting guidelines such as TRIPOD?

- TRIPOD is an evidence-based, minimum set of recommendations for reporting prediction modeling studies in biomedical sciences.
- TRIPOD is part of a wider set of guidelines under the https://www.equator-network.org/ including CONSORT for clinical trials
- TRIPOD includes both prognostic and diagnostic prediction models as well as prediction model development, validation, updating or extending studies (i.e. the core of AI/ML).
- TRIPOD offers a standard way for reporting the results of prediction modeling studies and thus aiding their critical appraisal, interpretation and uptake by potential users.
- TRIPOD and other related reporting guidelines have been adopted by many top tier scientific journals

NOVARTIS | Reimagining Medicine

Task-based benchmarking



Task

- **Tasks** reflect real project team requirements i.e. identify super-responders patients with known signatures

Data

- Provide benchmark(s) mirroring real Novartis data i.e. clinical trials
- Participants are free to use publically available data to augment analyses (i.e. through knowledge graphs or other propriety held data)

Evaluation

- Objective evaluation based on the benchmark (e.g predictive accuracy)
- Quality of reporting (i.e. description of methods, decision rules, plausibility, and recommendations) leveraging reporting guidelines

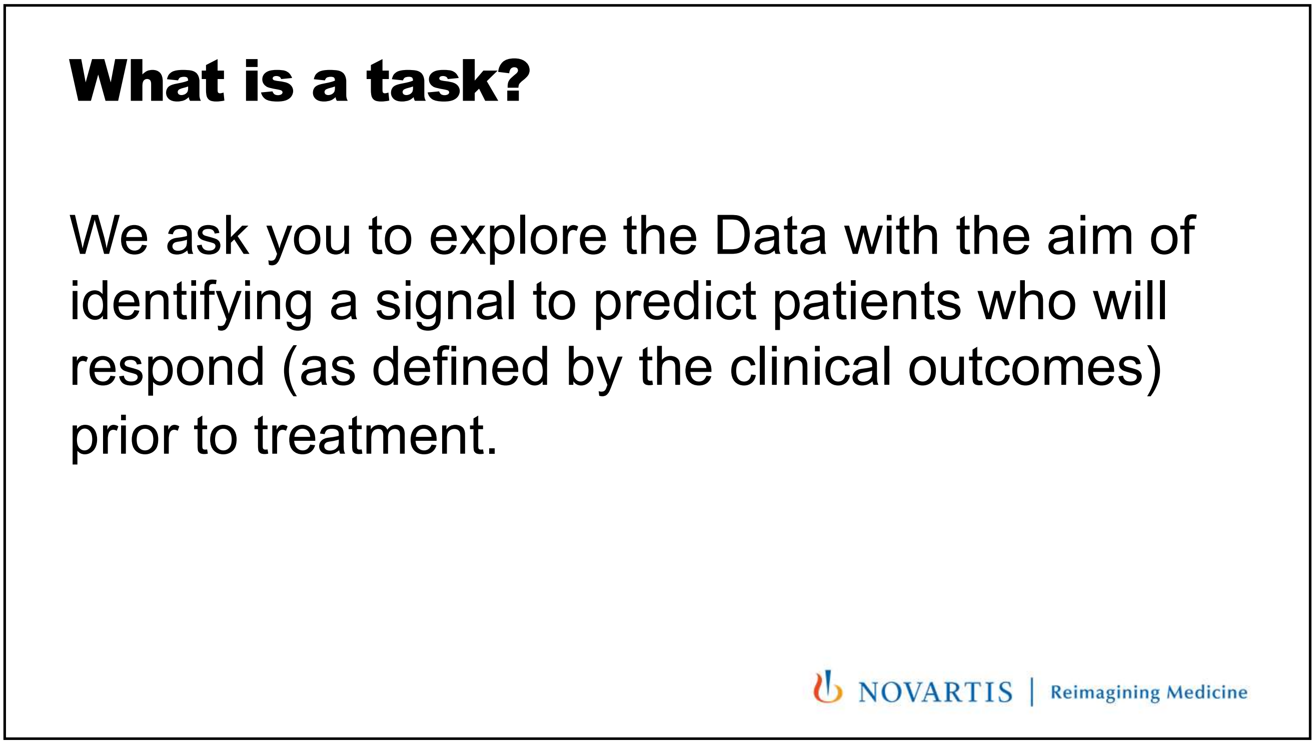
Summarize and document recommedation and socialise for internal use

What is a task?

task
noun
\ ˈtask \

- **:** a usually assigned piece of work often to be finished within a certain time
- **:** something hard or unpleasant that has to be done

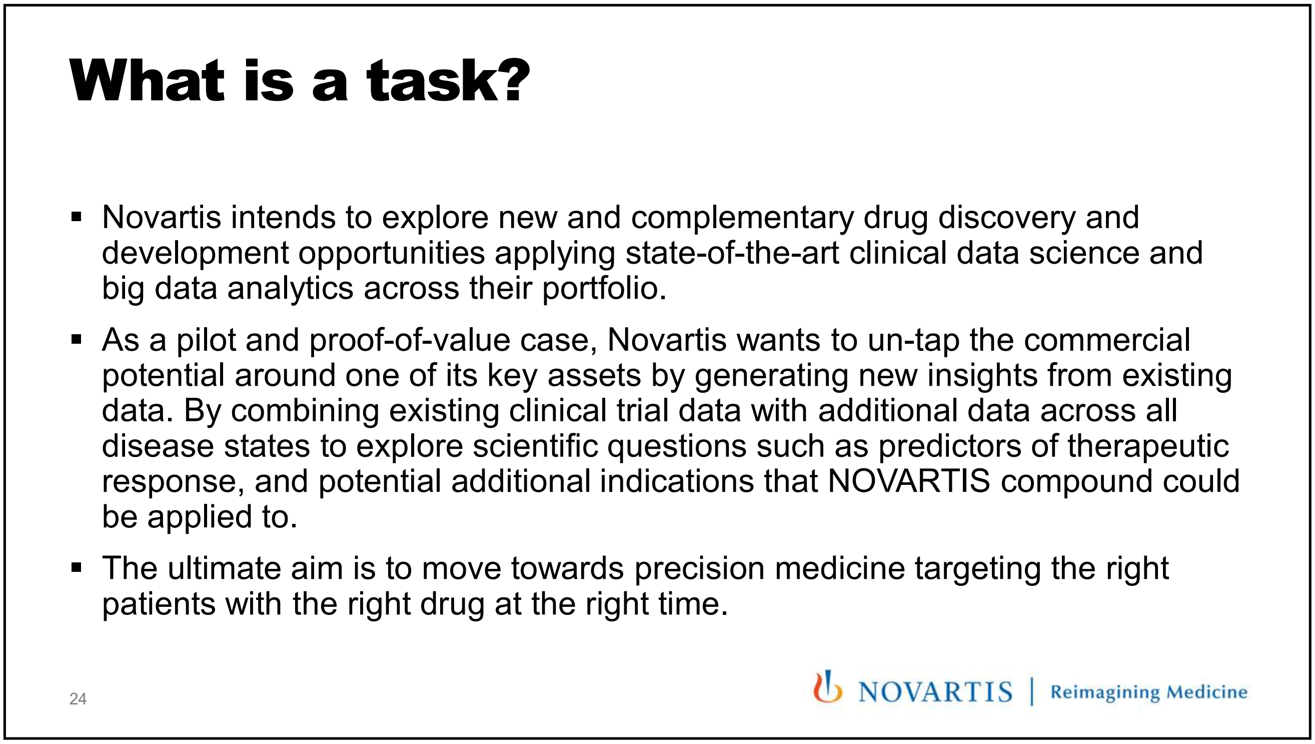
https://www.merriam-webster.com/dictionary/task



What is a task?

We ask you to explore the Data with the aim of identifying a signal to predict patients who will respond (as defined by the clinical outcomes) prior to treatment.

NOVARTIS | Reimagining Medicine



What is a task?

- Novartis intends to explore new and complementary drug discovery and development opportunities applying state-of-the-art clinical data science and big data analytics across their portfolio.
- As a pilot and proof-of-value case, Novartis wants to un-tap the commercial potential around one of its key assets by generating new insights from existing data. By combining existing clinical trial data with additional data across all disease states to explore scientific questions such as predictors of therapeutic response, and potential additional indications that NOVARTIS compound could be applied to.
- The ultimate aim is to move towards precision medicine targeting the right patients with the right drug at the right time.

24

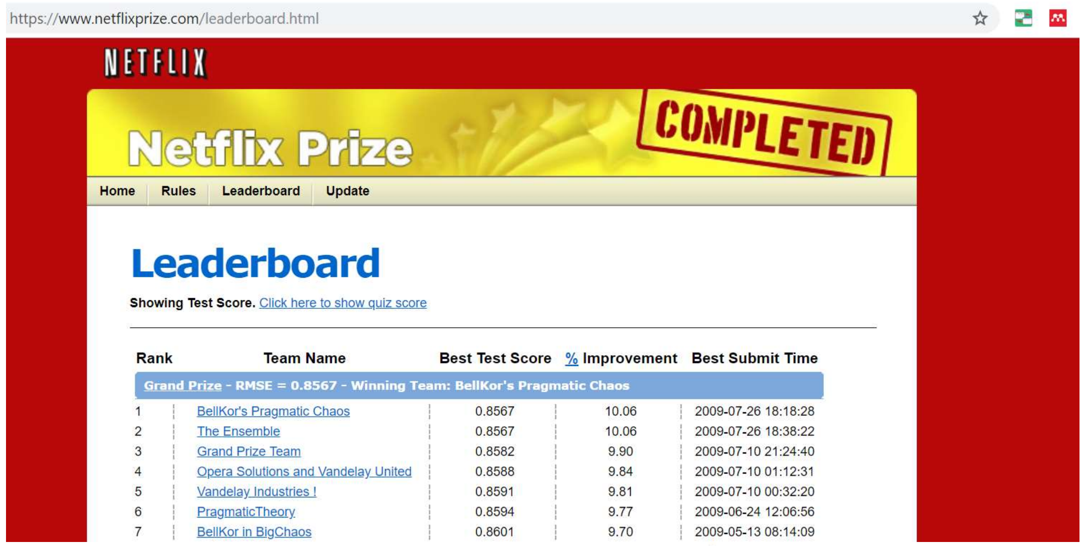
NOVARTIS | Reimagining Medicine

Example Benchmark Data

An example (secure) transfer to participants:

- Two phase 3 studies
  - 2,000 randomized patients
  - 180 clinical and genetic predictors (anonymized)
  - 5 clinical outcomes (endpoints)
- Additional supporting materials to provide context
  - Data dictionary
  - Data specifications
  - Trial manuscripts

Evaluation is task dependent

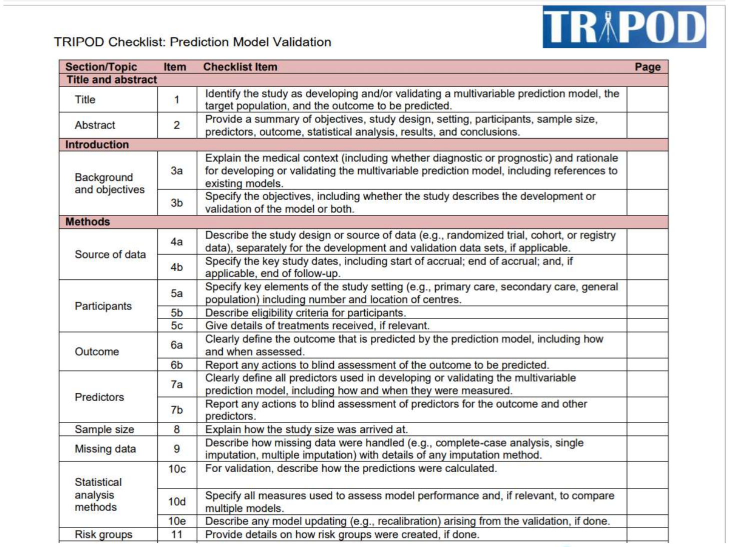


Leaderboard

Showing Test Score. Click here to show quiz score

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09

Evaluation is task dependent

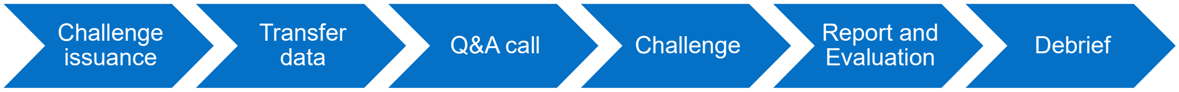


TRIPOD

TRIPOD Checklist: Prediction Model Validation

Section/Topic	Item	Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
	5b	Describe eligibility criteria for participants.	
	5c	Give details of treatments received, if relevant.	
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
	6b	Report any actions to blind assessment of the outcome to be predicted.	
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	
Sample size	8	Explain how the study size was arrived at.	
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
Statistical analysis methods	10c	For validation, describe how the predictions were calculated.	
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
	10e	Describe any model updating (e.g., recalibration) arising from the validation, if done.	
Risk groups	11	Provide details on how risk groups were created, if done.	

Putting it all together



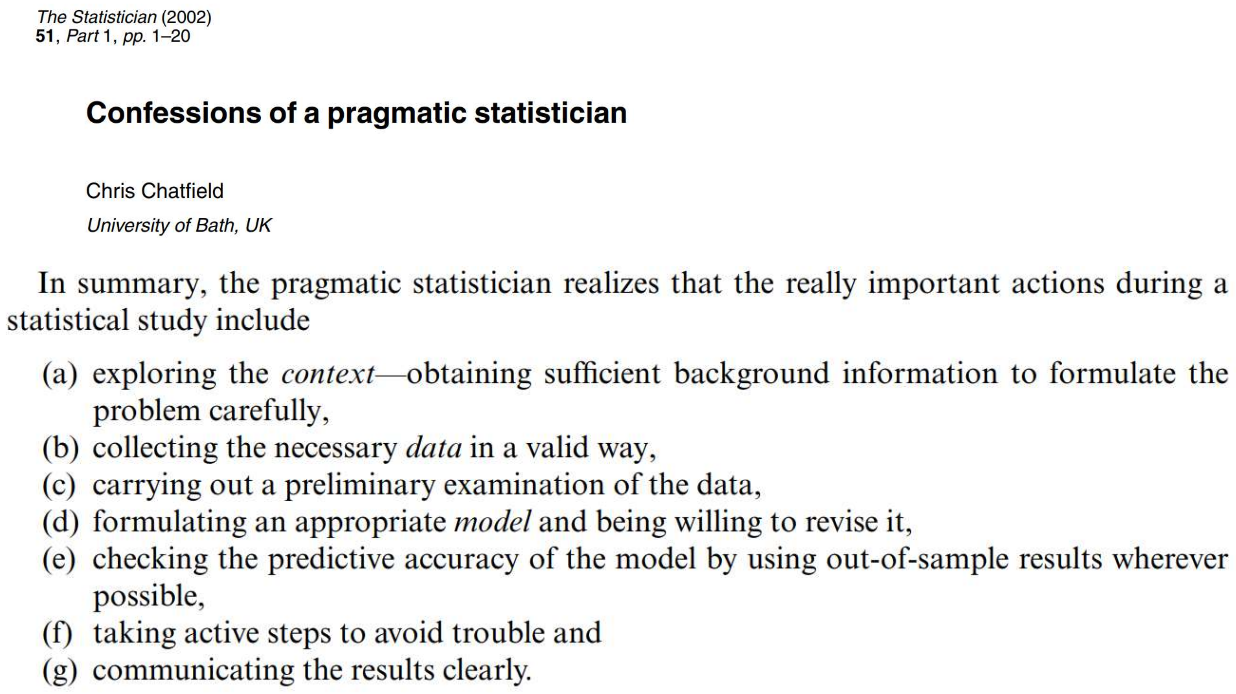
Challenge issuance → Transfer data → Q&A call → Challenge → Report and Evaluation → Debrief

- We have been evaluating the approach as a proof of concept
  - Issue issuance document with detailed information on challenge
  - Transfer data through secured service on receipt of signed document
  - Set up introductory call
  - Participant submits a short report documenting solution
  - Evaluation primarily based on the TRIPOD guidelines
  - Debrief call

Progress and learnings so far

- Learnings
- Black boxes
- Synthetic data

NOVARTIS | Reimagining Medicine



The Statistician (2002)
51, *Part* 1, *pp.* 1–20

Confessions of a pragmatic statistician

Chris Chatfield
University of Bath, UK

In summary, the pragmatic statistician realizes that the really important actions during a statistical study include

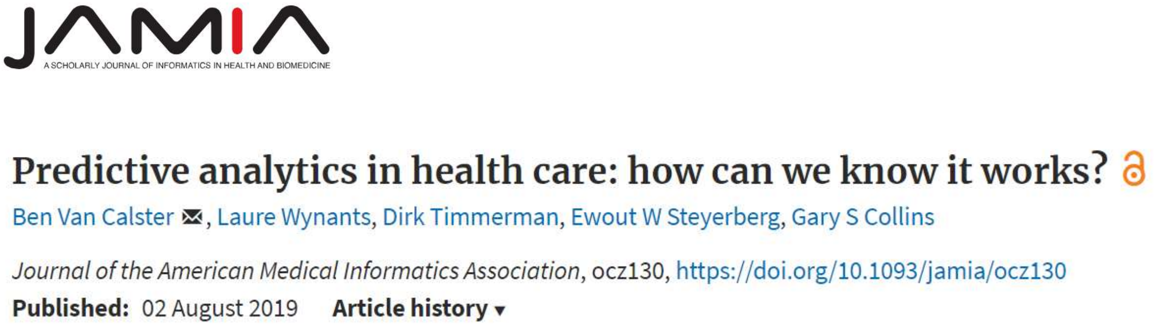
(a) exploring the *context*—obtaining sufficient background information to formulate the problem carefully,
(b) collecting the necessary *data* in a valid way,
(c) carrying out a preliminary examination of the data,
(d) formulating an appropriate *model* and being willing to revise it,
(e) checking the predictive accuracy of the model by using out-of-sample results wherever possible,
(f) taking active steps to avoid trouble and
(g) communicating the results clearly.

Black boxes?

- The advantage of benchmarking is that we define the task and the evaluation approach, therefore allowing us to assess the output of any black box
- Using synthetic data, we can set up tests to assess when a black box approach works or potentially fails
- Part of the assessment is to identify if the vendor is open to sharing methodological and implementation details about their approach
- Hiding algorithmic details for specific tasks such as disease progression is also considered **unethical** by many in the scientific community https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocz130/5542900
- Identifying early on a vendor approach to sharing information will help guide teams on future engagement and to ameliorate potential risks

NOVARTIS | Reimagining Medicine

Black boxes?



JAMIA

A SCHOLARLY JOURNAL OF INFORMATICS IN HEALTH AND BIOMEDICINE

Predictive analytics in health care: how can we know it works?

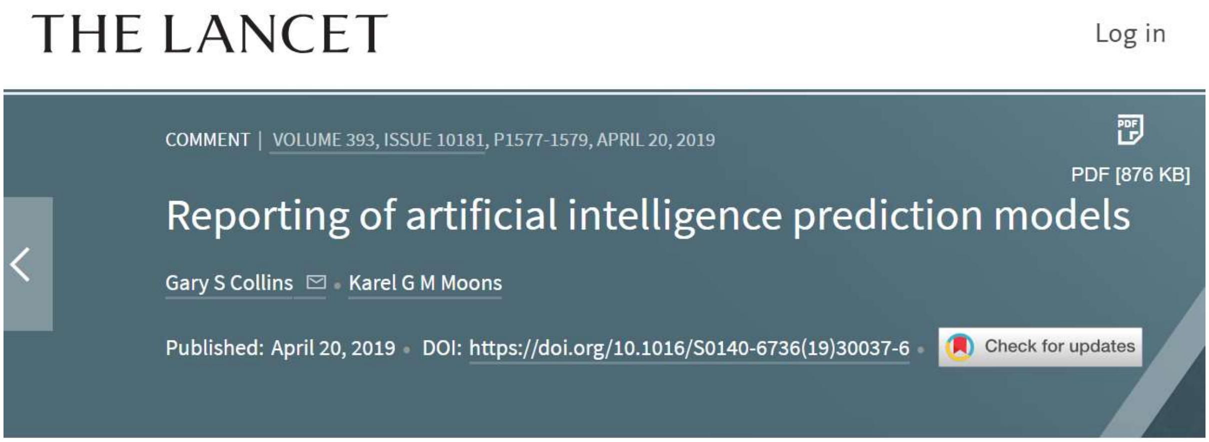
Ben Van Calster, Laure Wynants, Dirk Timmerman, Ewout W Steyerberg, Gary S Collins

Journal of the American Medical Informatics Association, ocz130, https://doi.org/10.1093/jamia/ocz130

Published: 02 August 2019 **Article history** ▾

https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocz130/5542900

Black boxes?



THE LANCET

Log in

COMMENT | VOLUME 393, ISSUE 10181, P1577-1579, APRIL 20, 2019

PDF [876 KB]

Reporting of artificial intelligence prediction models

Gary S Collins • Karel G M Moons

Published: April 20, 2019 • DOI: https://doi.org/10.1016/S0140-6736(19)30037-6 •

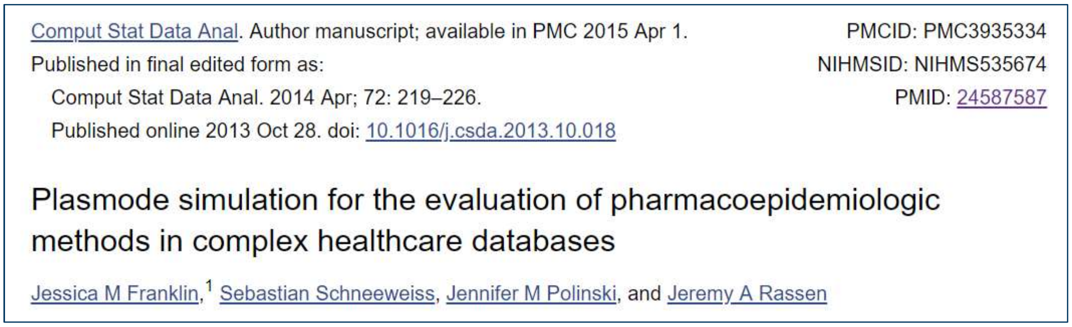
Check for updates

https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(19)30037-6/fulltext

Synthetic data

- Synthetic data is generated from real data, is not real data but has the same statistical properties.
- Synthetic data is generated using (statistical machine learning and deep learning) models from real data sampling pseudo patients from these models.
- Because it is not real data, it will not have the same privacy risks as real data. We can explicitly test that assumption.
- We can also introduce artificial signals (plasmode simulation) for the purpose of evaluation e.g. we introduce which patients will respond to a drug and why.
- We have developed this internally for the initial project.

NOVARTIS | Reimagining Medicine

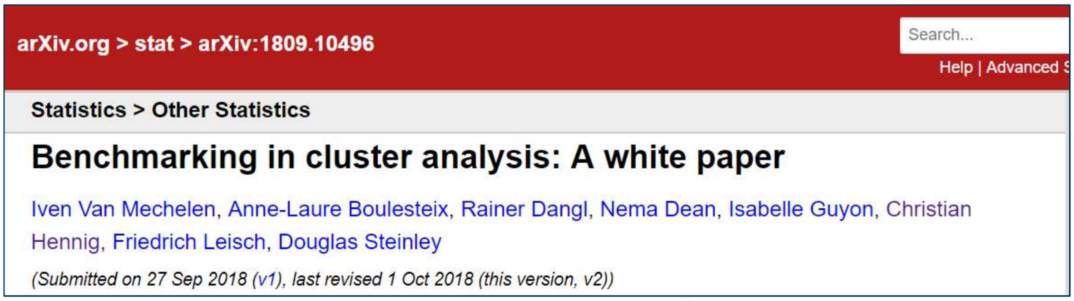


Comput Stat Data Anal. Author manuscript; available in PMC 2015 Apr 1.
Published in final edited form as:
Comput Stat Data Anal. 2014 Apr; 72: 219–226.
Published online 2013 Oct 28. doi: 10.1016/j.csda.2013.10.018

PMCID: PMC3935334
NIHMSID: NIHMS535674
PMID: 24587587

Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases

Jessica M Franklin,[1] Sebastian Schneeweiss, Jennifer M Polinski, and Jeremy A Rassen



Statistics > Other Statistics

Benchmarking in cluster analysis: A white paper

Iven Van Mechelen, Anne-Laure Boulesteix, Rainer Dangl, Nema Dean, Isabelle Guyon, Christian Hennig, Friedrich Leisch, Douglas Steinley

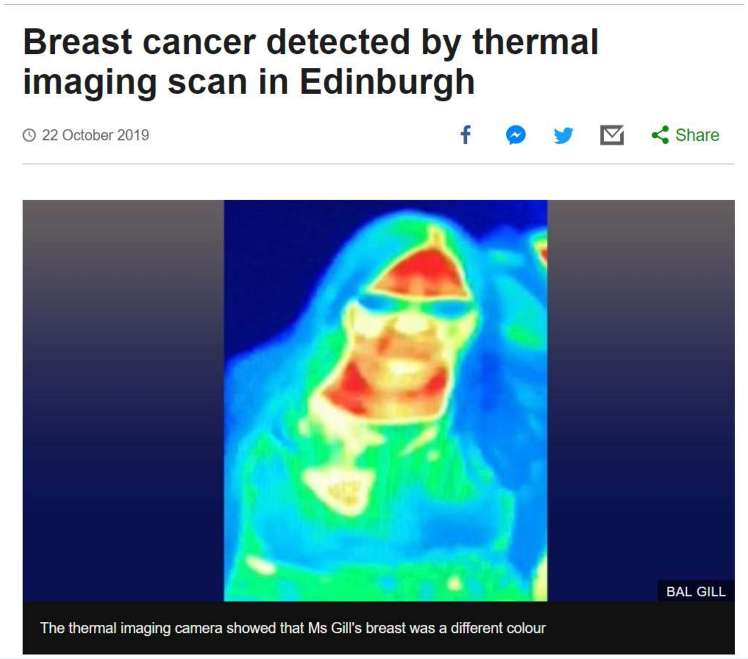
(Submitted on 27 Sep 2018 (v1), last revised 1 Oct 2018 (this version, v2))

NOVARTIS | Reimagining Medicine

Next steps: scaling up

- We have tested this approach, the next step is to scale up:
  - across the wider organization (i.e. all development units, countries, etc.)
  - develop a centralized knowledge base accessible across Novartis of all ongoing and completed engagements
  - company-wide disseminate of findings
  - company-wide coordination to avoid rework or duplication of effort
- Develop new challenges that will enable us to better understand claims on effectiveness
- Develop a plan to proactively engage scientifically community on methodology research
  - There is also a real need to develop new benchmarks which reflect real world

NOVARTIS | Reimagining Medicine



Breast cancer detected by thermal imaging scan in Edinburgh

22 October 2019

Share

The thermal imaging camera showed that Ms Gill's breast was a different colour

https://www.bbc.com/news/uk-scotland-edinburgh-east-fife-50139540

It’s not innovative if it doesn't work

NOVARTIS | Reimagining Medicine

AMDS
Clinical Development and Analytics

Thank you

Mark Baillie (with Conor Moloney & Janice Branson)
BBS, Basel
November 01, 2019

NOVARTIS | Reimagining Medicine