1.1 Clinical Problems and Demonstrators

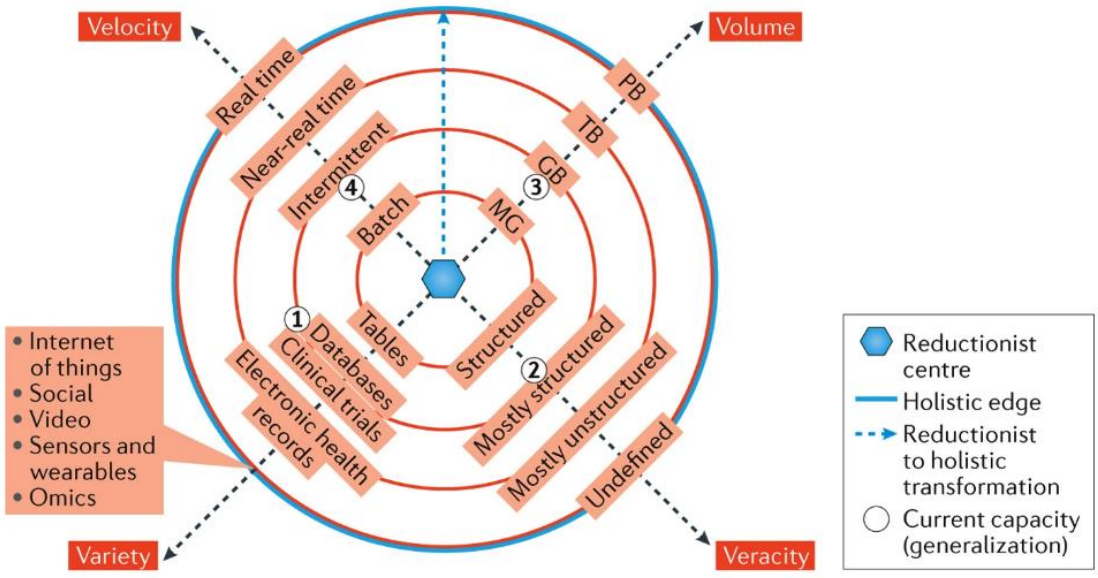
Dr Eoin Mckinney

Outline

- Definitions
- The problem with translation
- Example of success?
- Progress, challenges, and opportunities
    - Deployment
    - Data access
    - Beyond supervised assistance
    - To interpret or not to care?
- Summary
- Q&A

Definition of AI and ML

- AI: capability of a machine to imitate intelligent human behaviour
- ML: the field of study that gives computers the ability to learn without explicitly being programmed



Nature Reviews | Drug Discovery

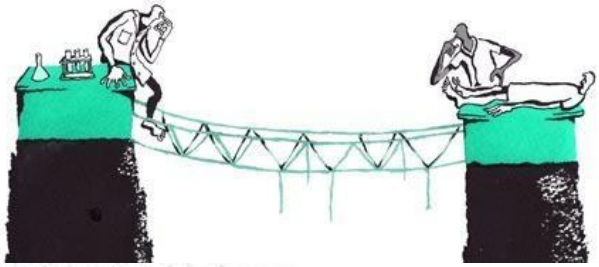
Big data can be defined as having four dimensions: volume (data size), variety (data type), veracity (data noise and uncertainty) and velocity (data flow and processing). Currently, FDA approval decisions are generally based on data of limited variety, mainly from clinical trials and preclinical studies (**1**) that are mostly structured (**2**), in data sets usually no more than a few gigabytes in size (**3**), that are processed intermittently as part of regulatory submissions (**4**). The expansion of big data in the four dimensions (grey lines) calls for increasing organizational and technical capacity. This could transform big data into smart data by enabling a holistic approach to personalization of therapies that takes patient, disease and environmental characteristics into account.

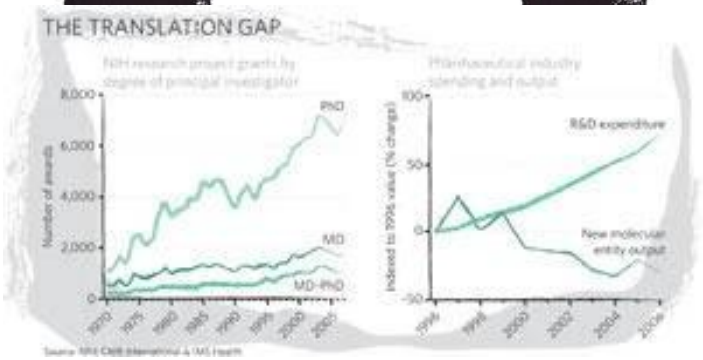
- https://www.nature.com/articles/nrd.2017.26
- contexts in terms of application has changed drastically and will change more
- Moving outwards from the circle
    - data moving towards real-time information
    - New methods required

"In the years ahead, AI is poised to broadly reshape medicine" – Eric Topol, Nature Medicine 2022

"AI and machine learning have the potential to truly revolutionise the delivery of healthcare, to the great benefit of patients, clinicians and the wider medical ecosystem" – Mihaela van der Schaar, CCAIM 2022

- Why are we still poised?
- How can we change to realisation?
    - The AI chasm (Nature 453;12 2008) — Crossing the valley of death
        - https://www.nature.com/articles/453840a

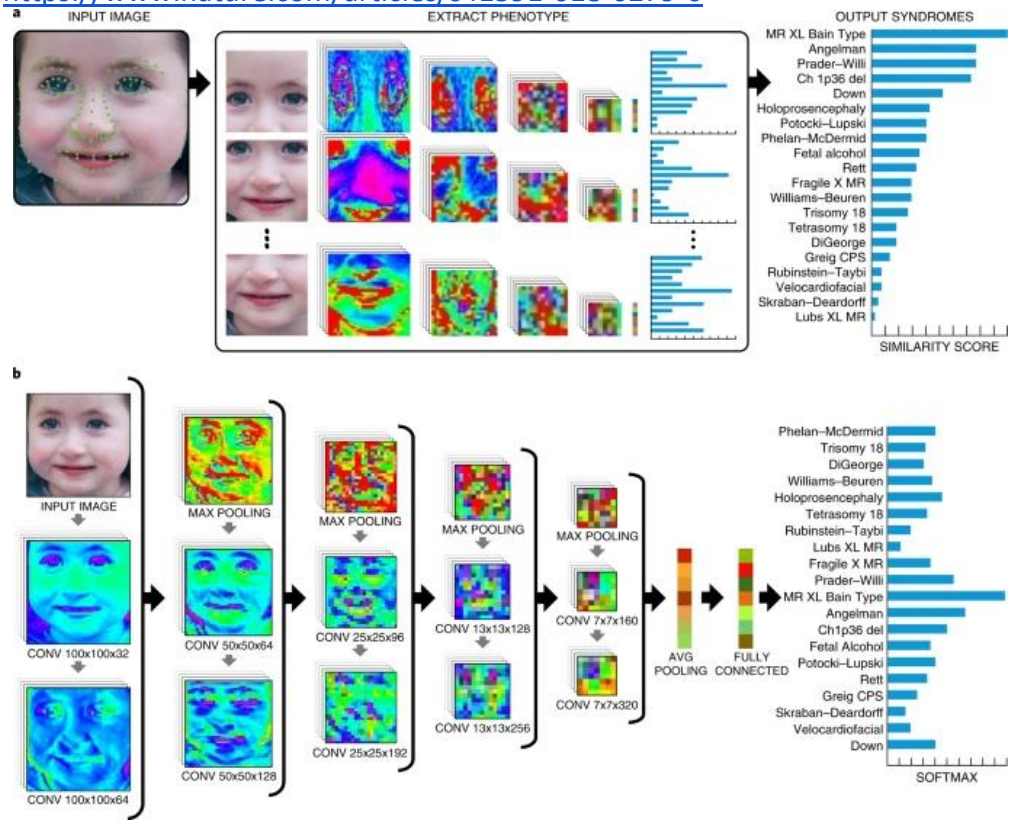




    - Getting towards impact is difficult
        - Presentation will cover key aspects
        - e.g. not translatable in 5 to 10 years. no impact made following publications

An example given: Identifying facial phenotypes of genetic disorders using deep learning

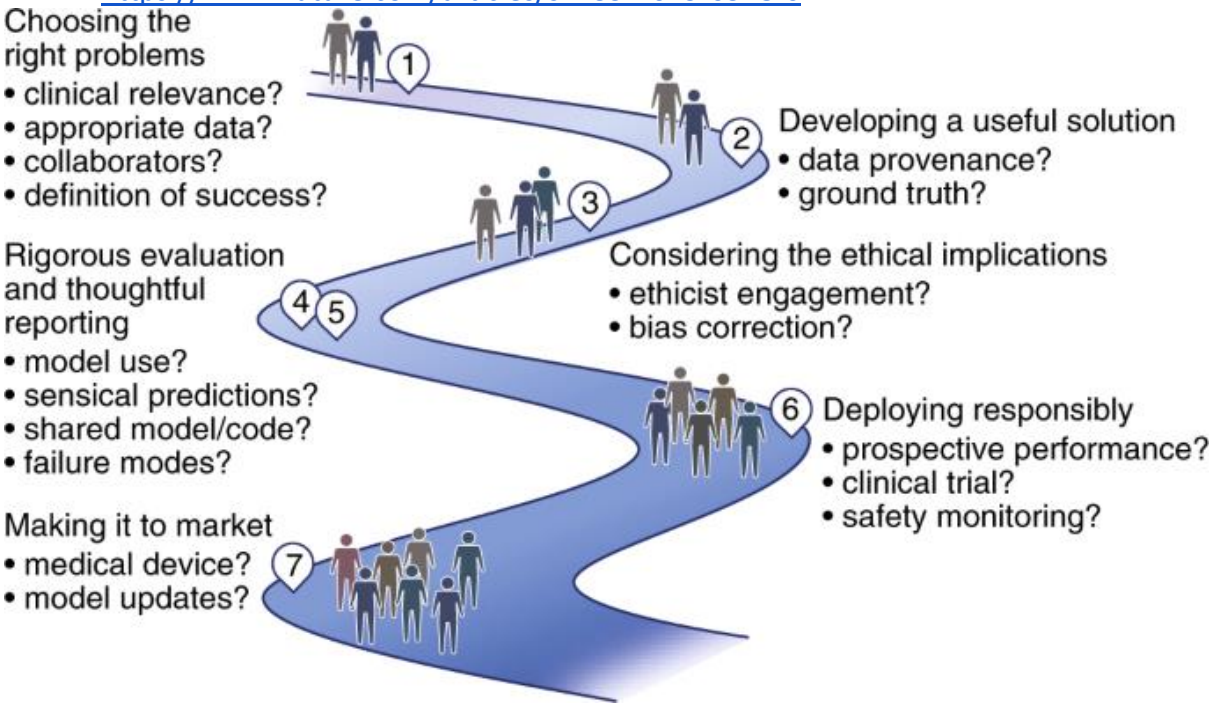
- https://www.nature.com/articles/s41591-018-0279-0



- DeepGestalt
    - Potentially adds considerable value to phenotypic evaluation in clinical genetics, genetic testing, research and precision medicine

Road to impact

- https://www.nature.com/articles/s41591-019-0548-6

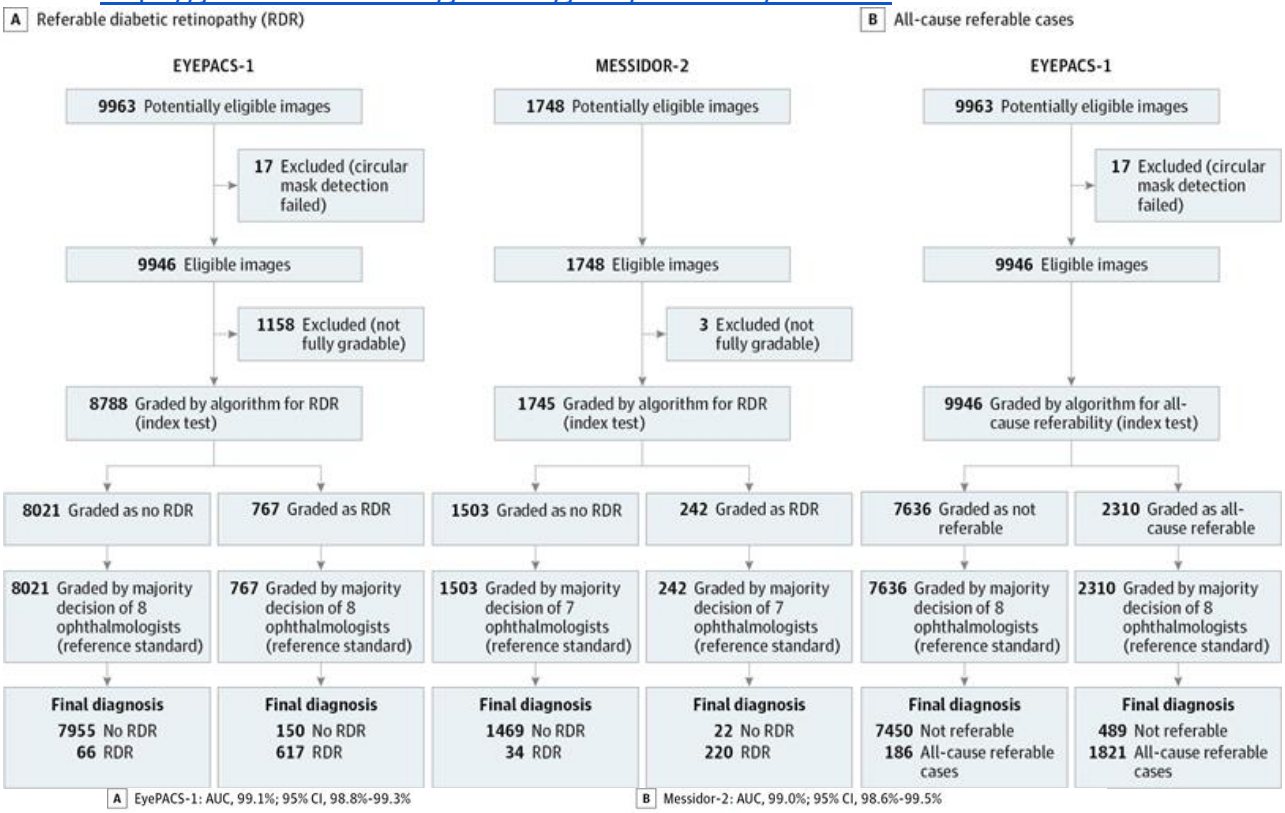


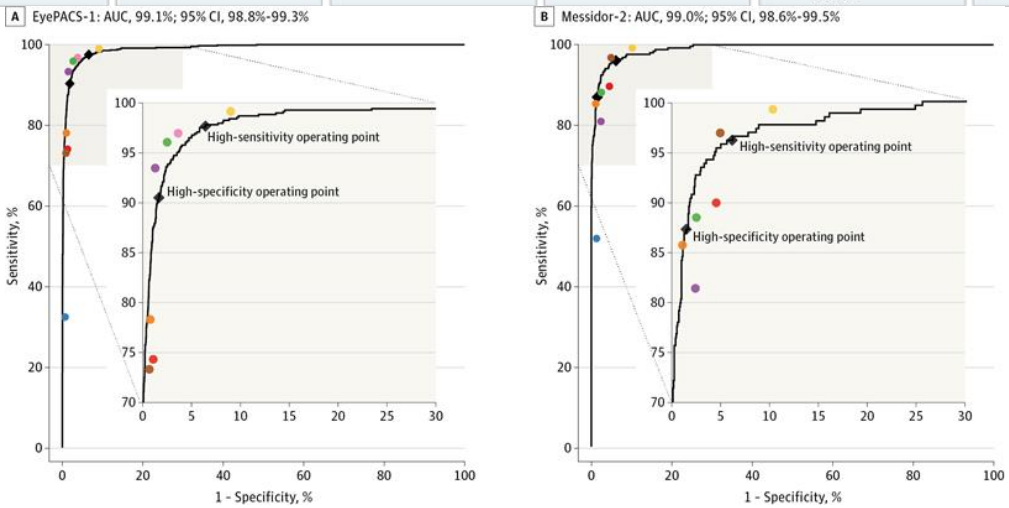
- 2. Useful model solution should result in
    - important impacts

- 4. Model use:
    - Critical consideration
- 6. Deploying responsibly:
    - Ensuring when something is deployed, it remains useful
- 7. Making it to market:
    - Up to date model

Making it to market and staying up to date – Retinal photographs example:

- https://jamanetwork.com/journals/jama/fullarticle/2588763





- CNN developed to train and diagnose diabetic retinopathy
- high sensitivity and specificity when compared to **expert opinions**
    - replacing medical individual with algorithms
    - still require camera and photo-taking
    - well defined goals and well annotated dataset (ground truth)

- Years after: pivotal trial on the model
  - **Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices**
  - https://www.nature.com/articles/s41746-018-0040-6

	Point estimate	95% CI	Superiority endpoint
Sensitivity	87.2%	81.8%–91.2%	85.0%
Specificity	90.7%	88.3%–92.7%	82.5%

Point estimates for sensitivity and specificity were calculated on the 819 participants that were analyzable, using the prespecified logistic regression. The superiority endpoints were previously discussed with FDA.

  - Requirements of FDA to define the superiority end-point
    - not algorithm better than human: does not have to be perfect
    - drop-outs of samples from data (10%): unreadable
      - not included in clinical trials, point estimates cleared superiority when added later on
    - pivotal trial: testing whether algorithm better than
      - does not consider economical impact
      - uncertain if the algorithm can make healthcare system effective
    - Further studies conducted years after: https://onlinelibrary.wiley.com/doi/10.1111/aos.13613

Table 4. Classification of diagnosis according to the IDx-DR device compared to the gold standard and accuracy measures [95% confidence intervals (CIs)] of the IDx-DR device using the EURODIAB and ICDR classification score

IDx-DR	EURODIAB			ICDR			Total
	Human grading			Human grading			
	No RDR	MDR	VTDR	No RDR	MDR	VTDR	
No RDR	732	1	1	711	22	1	734
MDR	101	3	4	76	28	4	108
VTDR	43	4	9	38	10	8	56
Total	876	8	14	825	60	13	898

	RDR	VTDR	RDR	VTDR
Se	0.91 (0.69–0.98)	0.64 (0.36–0.86)	0.68 (0.56–0.79)	0.62 (0.32–0.85)
Sp	0.84 (0.81–0.86)	0.95 (0.93–0.96)	0.86 (0.84–0.88)	0.95 (0.93–0.96)
PPV	0.12 (0.08–0.18)	0.16 (0.08–0.29)	0.30 (0.24–0.38)	0.14 (0.07–0.27)
NPV	1.00 (0.99–1.00)	0.99 (0.99–1.00)	0.97 (0.95–0.98)	0.99 (0.99–1.00)

Accuracy measures are presented with 95% CI.

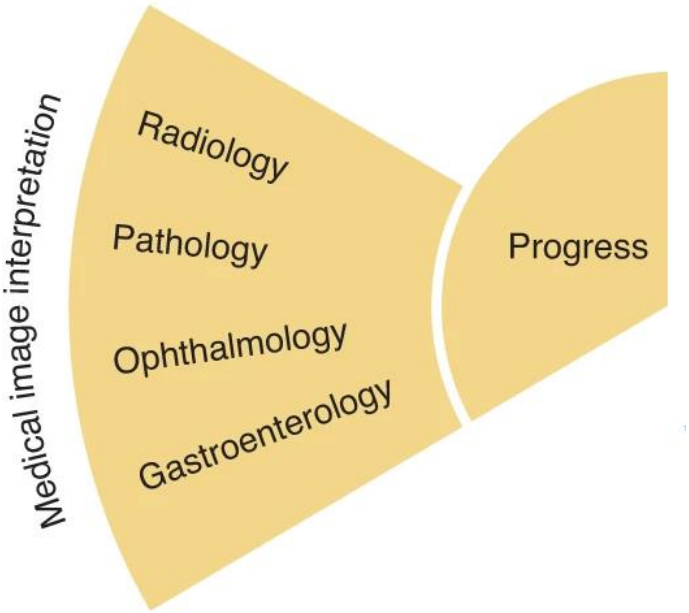
ICDR = International Clinical Diabetic Retinopathy severity scale, MDR = moderate diabetic retinopathy, NPV = negative predictive value, PPV = positive predictive value, RDR = referable diabetic retinopathy (moderate and vision-threatening diabetic retinopathy), Se = sensitivity, Sp = specificity, VTDR = vision-threatening diabetic retinopathy.

        - Risk missing cases due to severe end-points,
            - sensitivity can be more important than specificity
            - drop of up to 40%, images not up to quality -- cost efficiency
    - In the end, Algorithm passed
        - unanswered questions:
            - Will it be used in hospitals?
            - Is it cost saving?
    - Some of these are due to reproducibility:
        - Independent replication shows low reproducibility
            - **Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs**
                - https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0217541
            - **Evaluation of Artificial Intelligence–Based Grading of Diabetic Retinopathy in Primary Care**

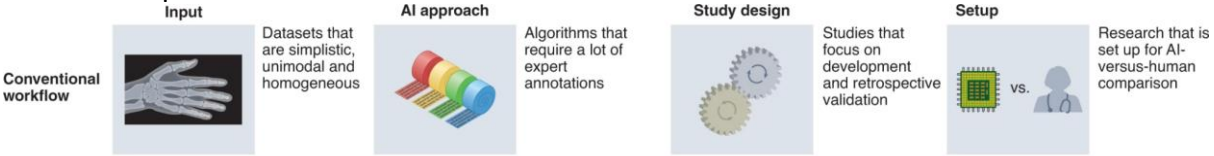
        - https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2703944
    - Reproducibility: reproducing the results shared in the initial publication
        - Challenging
        - commercial sensitivity: no code sharing or dataset sharing
        - random seeding: may not be reproducibility
        - Data release changes
        - code version updates
        - inherent methods (stochastic descent)
    - Independent replication:
        - Substantially impacts from new data quality
        - Much more lower than published results
- Publication showing 35 highly-cited biomarker publications:
    - Comparison of Effect Sizes Associated With Biomarkers Reported in Highly Cited Individual Articles and in Subsequent Meta-analyses
    - https://jamanetwork.com/journals/jama/article-abstract/900417
    - Once reached to clinical deployment
        - performance is expected to reduce significantly

Problems tackled

- AI in health and medicine (https://www.nature.com/articles/s41591-021-01614-0)



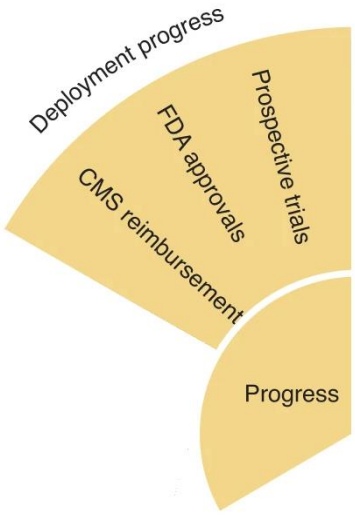
- Increasing success and usage in DL and ML methods, especially in medical imaging interpretation



- **Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review**
    - https://www.nature.com/articles/s41746-021-00524-2

Deployment

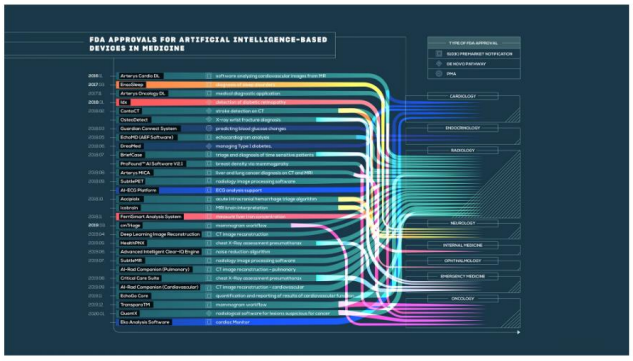
- https://www.linkedin.com/pulse/fda-proposed-regulatory-framework-modifications-samd-adafinoaiei



-
- Orphan device assignment for FDA approval
- assisted device = low risk
- Comparison with human performance
- Stringent data quality selection

Fig. 1: An infographic about the 29 FDA-approved, AI/ML-based medical technologies.

From: The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database



The devices have features such as date and type of FDA approval; name of the device, its short description and which primary and secondary medical specialty it is related to.

https://www.nature.com/articles/s41746-020-00324-0

- Much progress based on images, followed by cardiology (ECG)
    - approval is not enough, further demonstration of impact beyond approval

Data access

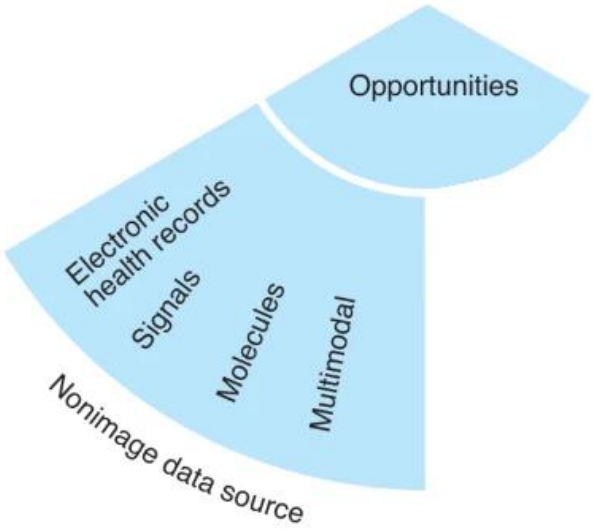
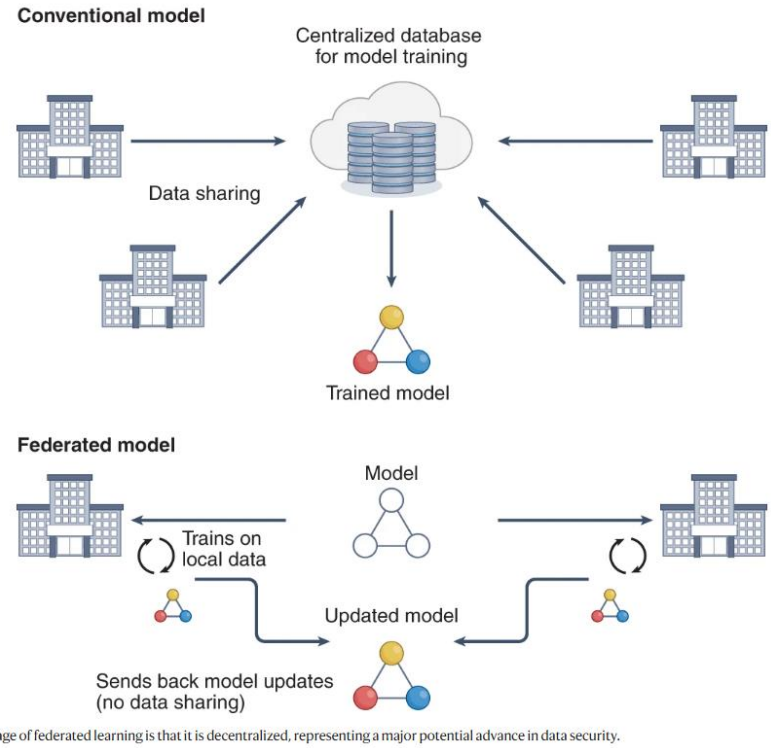


Fig. 4: Evolving procedures for data sharing.

From: AI in health and medicine



An advantage of federated learning is that it is decentralized, representing a major potential advance in data security.

- Historically, local learning and independently validated
  - replaced by central learning mechanism (data and parameters housed centrally)
  - challenges in:
    - collaboration
    - security
    - data ownership
  - Replaced by federated model

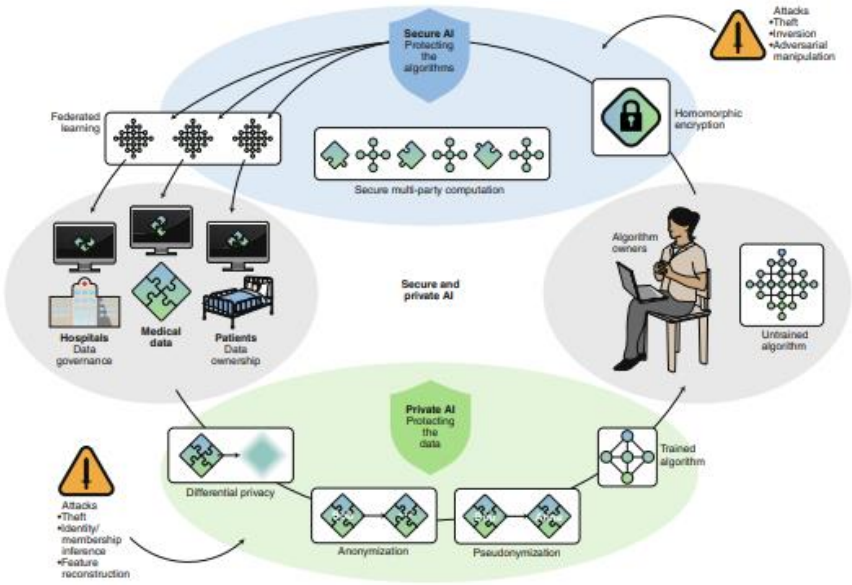
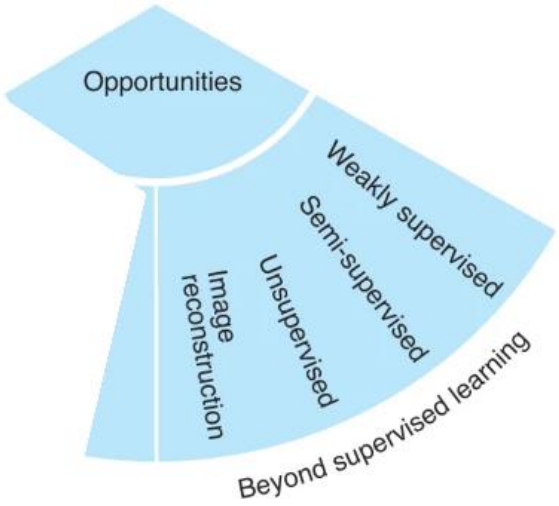


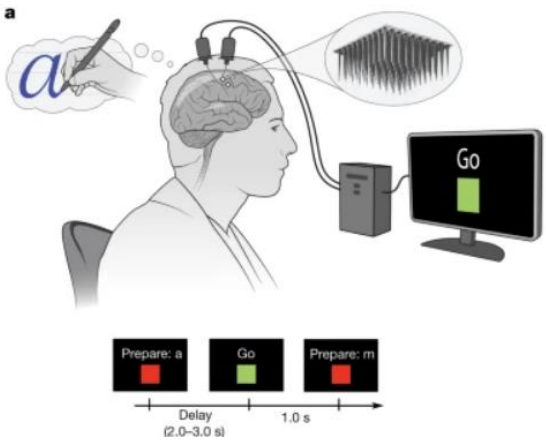
Fig. 1 | Secure and private AI. Schematic overview of the relationships and interactions between data, algorithms, actors and techniques in the field of secure and private AI.

- Further advances: swarm learning (block chain access)
  - https://mediatum.ub.tum.de/doc/1602022/1602022.pdf
  - https://www.nature.com/articles/s41586-021-03583-3

Beyond supervised assistance

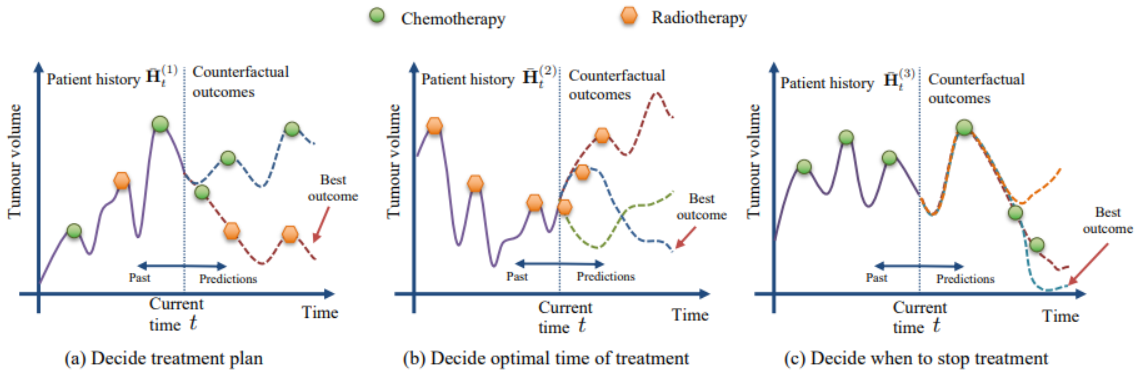


- **Example: High-performance brain-to-text communication via handwriting**
  - https://www.nature.com/articles/s41586-021-03506-2



a, To assess the neural representation of attempted handwriting, participant T5 attempted to handwrite each character one at a time, following the instructions given on a computer screen (bottom panels depict what is shown on the screen, following the timeline). Credit: drawing of the human silhouette created by E. Woodrum.

  - rich longitudinal NN
    - RNN to extract and learn what letter has been written
    - time-series information with clear ground truth
- Learning from an entire health record
  - Complexity of different information increases greatly
    - Application of NN made better
      - e.g. death, length of stay
      - better supporting the healthcare system
  - Individual treatment effect inference
    - **ESTIMATING COUNTERFACTUAL TREATMENT OUTCOMES OVER TIME THROUGH ADVERSARIALLY BALANCED REPRESENTATIONS**
      - https://arxiv.org/pdf/2002.04083.pdf

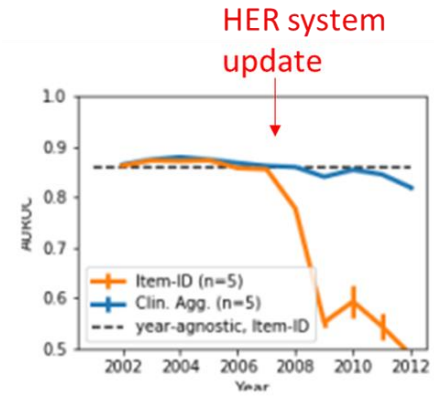


(a) Decide treatment plan (b) Decide optimal time of treatment (c) Decide when to stop treatment

Figure 1: Applicability of CRN in cancer treatment planning. We illustrate 3 patients with different covariate and treatment histories $\bar{\mathbf{H}}_t$. For a current time t, CRN can predict counterfactual trajectories (the coloured dashed branches) for planned treatments in the future. Through the counterfactual predictions, we can decide which treatment plan results in the best patient outcome (in this case, the lowest tumour volume). This way, CRN can be used to perform all of the following: choose optimal treatments (a), find timing when treatment is most effective (b) decide when to stop treatment (c).

- Decide treatment plan, optimal time of treatment, when to stop treatment
- Another example:
  - When treatment should be stopped,

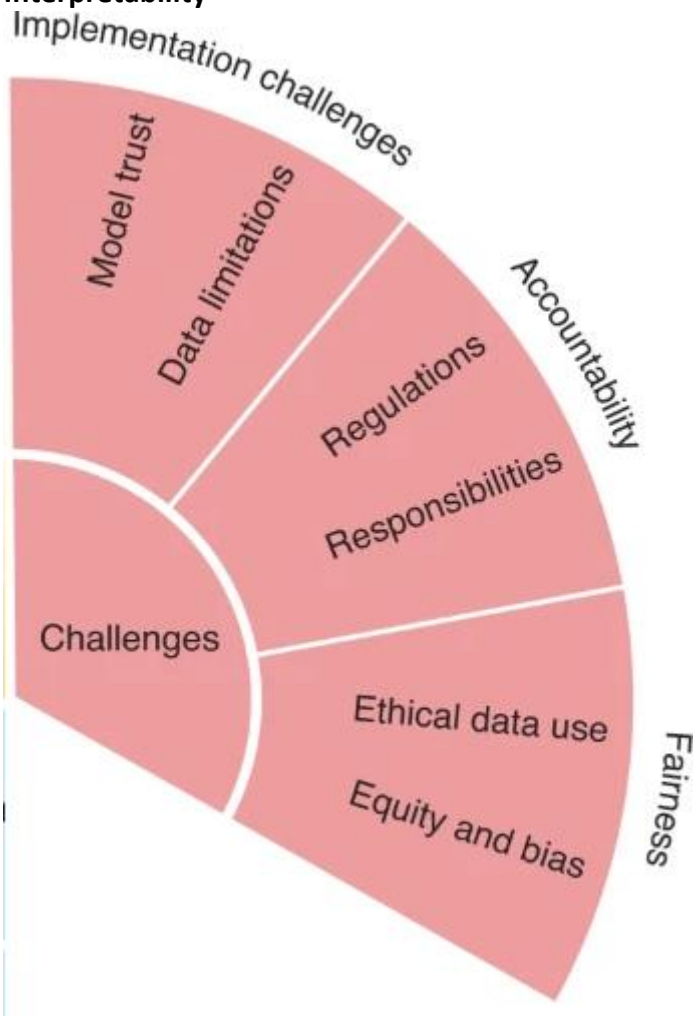
- Rethinking clinical prediction: considering year of care and feature aggregation in ML
  - **Why machine learning must consider year of care and feature aggregation**
    - https://arxiv.org/pdf/1811.12583.pdf



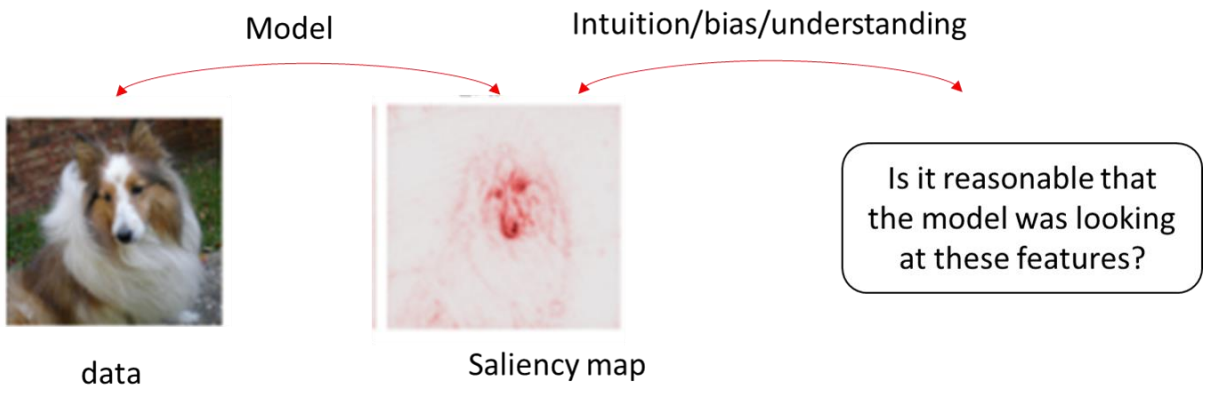
(c) Mortality AUC, models trained yearly on all prior data.

  - Dynamic system: changes with time
    - data changes dependent on the year and where collected
    - if learning is off-line at initial stage, the changes of healthcare system will lead to cliff-edge
      - constant learning
      - particular challenge at deployment

Interpretability



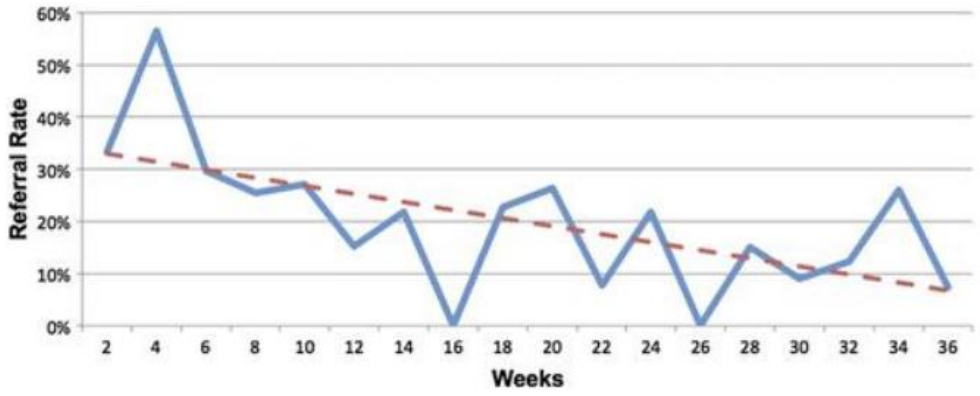
- Concepts of interpretability and why is it important
- Interpretability:
  - step 1: insights into decision-making process
  - step2 : mechanistic understanding of whether the process was sensible
    - e.g. saliency map
    - much harder to define and look at
    - potentially biased: based on your own intuition
    - many are not interpretable
  - e.g. feature of a dog (dog or cat?)



- What do clinicians want?

- Blackbox models usually outperform transparent models
  - arguably should develop more blackbox?
- Feature importance
  - Clinicians wants to understand the individual variances
- Instance level explanations
  - i.e. what other instances should be considered from the model if two predictions are similar?
- Temporal explanations?
  - how changes in longitudinal is factored in?
- Transparent design
  - Not always available
- Avoidance of alarm fatigue
  - as demonstrated previously
  - why the prediction is made thus understanding what risks the alarm fatigue

Figure 3



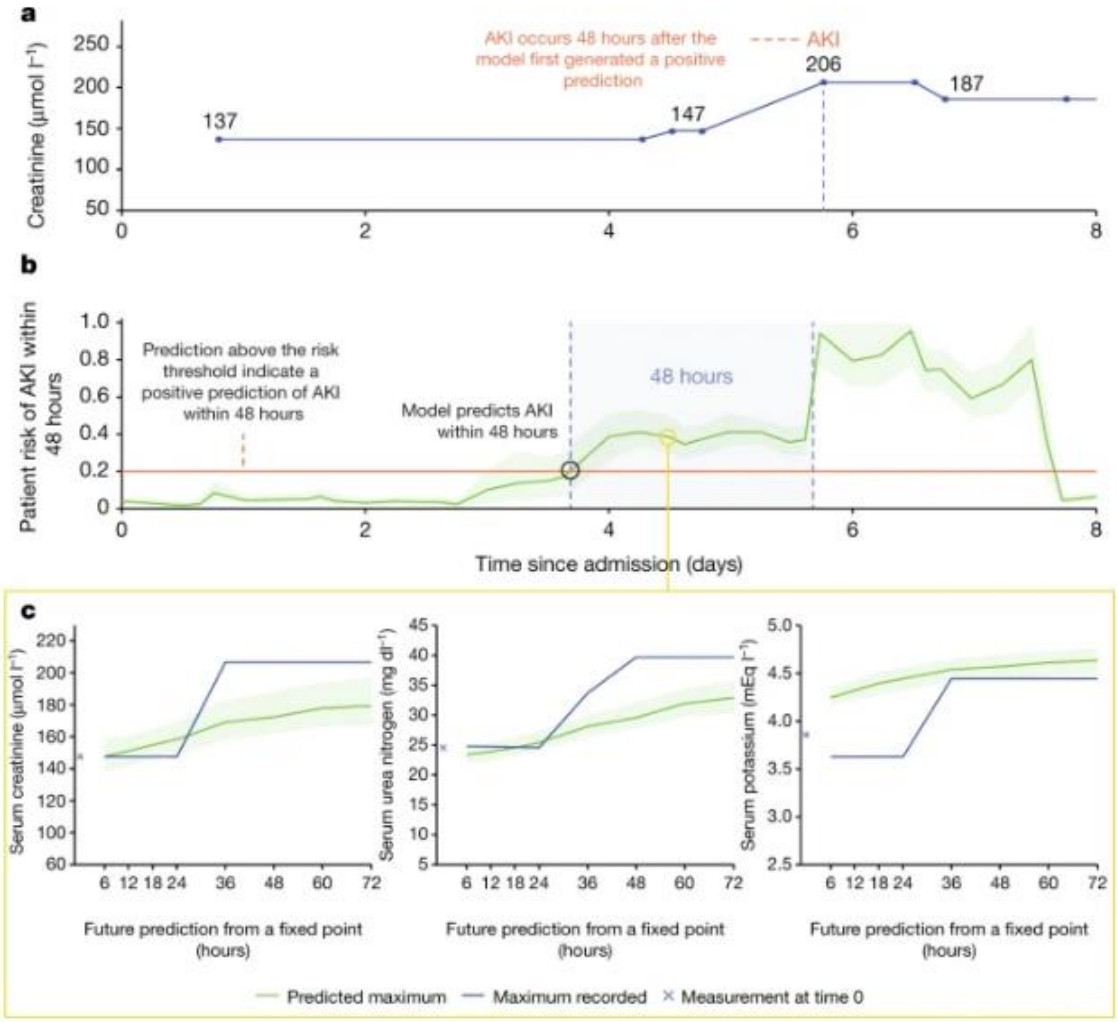
Open in new tab Download slide

Physician-generated referral rates using clinical trial alerts (CTAs) are plotted at 2-week intervals over the 36-week study. The solid line tracks referrals rates at each time point. The dashed line represents the linear regression line through each time point. Referral rates declined at a rate of 4.9% per 2-week time period (p=0.0294).

- **Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study**
- https://doi.org/10.1136/amiajnl-2011-000743

- Example of alarm fatigue:

Fig. 1: Illustrative example of risk prediction, uncertainty and predicted future laboratory values.



The first 8 days of admission for a male patient aged 65 with a history of chronic obstructive pulmonary disease. **a**, Patient creatinine measurements during admission. Creatinine measurements, showing AKI occurring on day 5. **b**, Model predictions for any AKI within 48 h. Continuous risk predictions: the model predicted increased AKI risk 48 h before it was observed. A risk above 0.2 (corresponding to 33% precision) was the threshold above which AKI was predicted. Lighter green borders on the risk curve indicate uncertainty, taken as the range of 100 ensemble predictions (after these were trimmed for the highest and lowest 5 values). **c**, Laboratory value predictions 4.5 days into admission. Predictions of the maximum future observed values of creatinine, urea and potassium.

- **A clinically applicable approach to continuous prediction of future acute kidney injury**
  - https://www.nature.com/articles/s41586-019-1390-1
- model can predict kidney injury with 2 day lead time
- ratio to true to false positive was 2:1
  - early stage of clinical performance, showing alarm fatigue

- A need to understand?
  - Can highly complex models be explained? do they need to be?
  - Legal perspective:
    - EU GDPR Article 13-15
    - must be interpretable: "meaningful information"
  - Risk of apparently reassuring saliency maps
    - saliency maps:
      - or feature weights
      - human needs to interpret the actual feature and quantifying it
  - Explaining explanations: quantifying interpretation?
  - risk of bias/ model debugging
- **Intuitive interpretation**

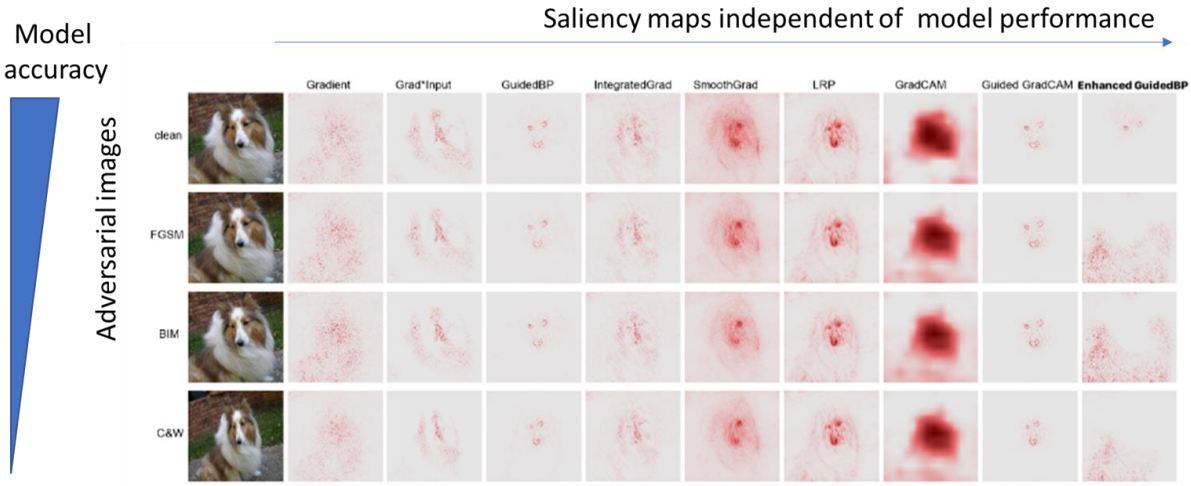
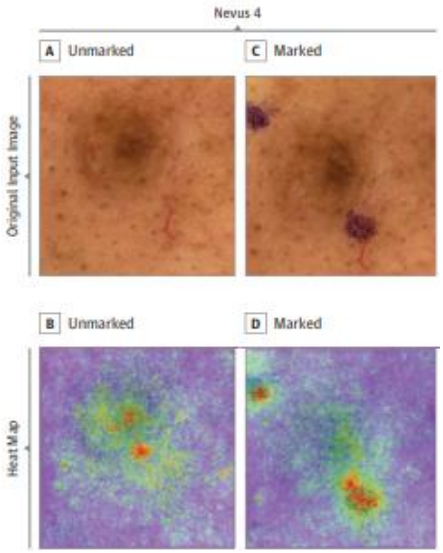


Figure 3: This figure shows SMs of clean image and adversary ones. The first column lists the original image and its adversary ones. Our Enhanced GuidedBP reacts the adversary attacks strongly, while all other the SMs produce similar SMs.

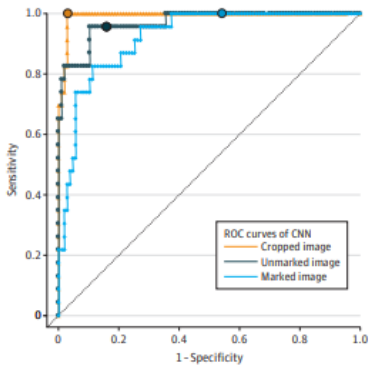
  - **Saliency Methods for Explaining Adversarial Attacks**
    - https://www.semanticscholar.org/reader/d18a5a20e76d60206caee92b23db529329e34061
  - should we be reassured if what the model is looking at “make sense”?
  - Saliency maps = make sense?
    - same features being looked at even reduced model accuracy
    - falsely reassured that the model is working as expected
  - “You could have many explanations for what a computer model is doing. Do you just pick the one you ‘want’ to be correct?” – Cynthia Rudin

Bias and debugging



The heat maps were created by vanilla (meaning basic) gradient descent backpropagation. A and E, Unmarked input images. B and F, Heat maps reveal relevant pixels for the convolutional neural network's (CNN's) prediction of benign nevi. C and G, Marked input images. D and H, Heat maps reveal that skin markings are of high relevance for CNN's changed prediction of malignant melanomas, while the nevus itself is mostly ignored.

Figure 5. Receiver Operating Characteristic (ROC) Curves of the Performance of the Convolutional Neural Network (CNN) Diagnostic Classification



Shown are ROC curves of the CNN in the original unmarked image set vs marked and cropped images (dichotomous classification). The sensitivity and specificity of the CNN for each image set at the a priori operation point are indicated by a circle on the curve and were as follows: (1) unmarked images: sensitivity, 95.7%; specificity, 84.1%; area under the curve (AUC), 0.969; (2) marked images: sensitivity, 100%; specificity, 45.8%; AUC, 0.922; and (3) cropped images: sensitivity, 100%; specificity, 97.2%; AUC, 0.993.

- **Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition**
  - https://jamanetwork.com/journals/jamadermatology/fullarticle/2740808
- Model were markedly influenced by the skin marking on the marked lesions
  - falsely supporting diagnosis
- Other example risk due to model bias

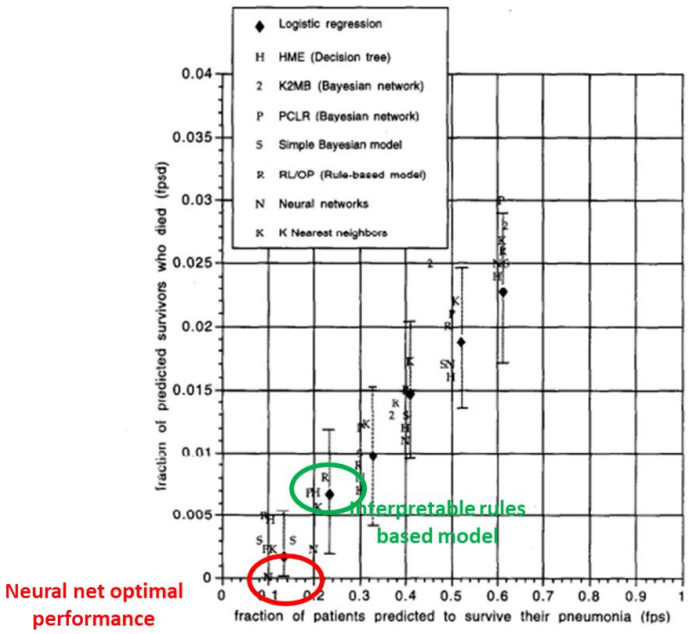
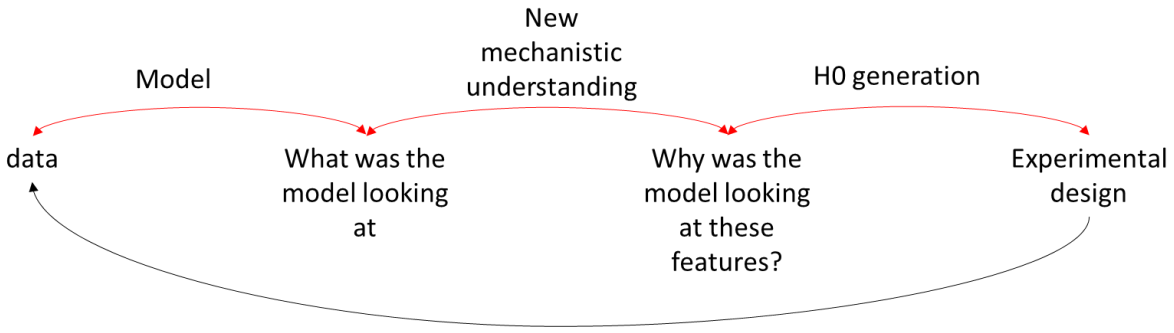


Fig. 1. A plot of the fpsd error rate as a function of fps for the eight models named in the legend. Each vertical bar through a logistic regression data point indicates a 95% confidence interval around that point. Some of the letters in the plot have been moved slightly to the right or left to prevent their overlapping each other.

  - **An evaluation of machine-learning methods for predicting pneumonia mortality**
    - https://www.sciencedirect.com/science/article/pii/S0933365796003673
    - the hospital data implied the rule “HasAsthma (x)⇒LowerRisk(x).”

        - data leak from asthmatic patients in the training dataset, not factored in
            - asthma was shown as low risk due to their chronic asthma treatment
    - **Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission**
        - https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/KDD2015FinalDraftIntelligibleModels4HealthCare_igt143e-caruanaA.pdf
        - Similarly
- **Clinical demonstrators to facilitate interpretation and debugging**
    - More information: https://www.vanderschaar-lab.com/engagement-sessions/revolutionizing-healthcare/
    - software will be showcased later in the week
        - clinical demonstrator allowing visualization of potential data leak for those unfamiliar to saliency maps
        - interpretable allowing debugging before clinical translation
- Interpretability in genomics
    - skipped
    - fundamental for interpretability, hypothesis generation rely on the results



- Summary
    - Definitions
    - Problem with translation
    - Example of successful models
    - Progress, remaining challenges, and opportunities
    - changes in infrastructure for deployment
    - Data accessibility
    - Beyond supervised assistance
    - interpretability very important for translation