Artificial Intelligence in Radiology: Panacea or Snake Oil?

Safwan S. Halabi MD - @SafwanHalabi

Associate Professor of Radiology,
Northwestern University Feinberg School of Medicine

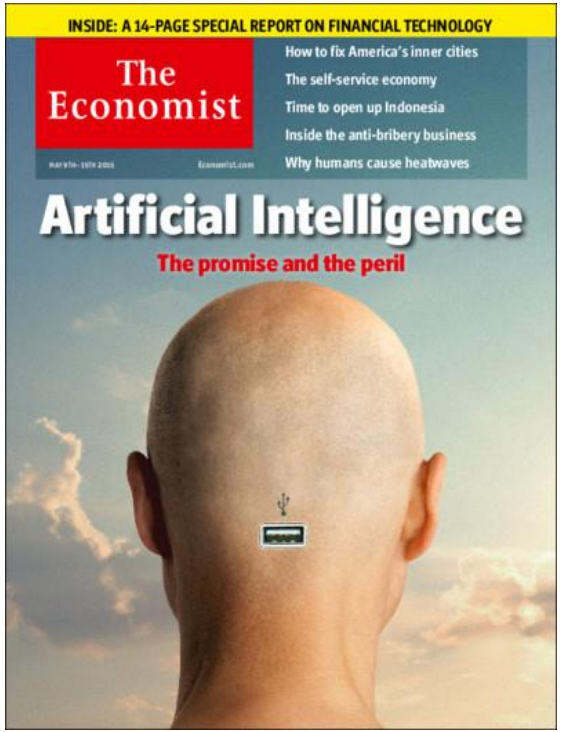
Vice-Chair of Imaging Informatics,
Ann & Robert H. Lurie Children's Hospital of Chicago

Disclosures

Advisor	Bunkerhill Health
Consultant	Change Healthcare
Board Member	Society of Imaging Informatics in Medicine (SIIM)
Member	RSNA Informatics Committee
Chair	Data Science Standards Subcommittee

Objectives

1. Define artificial intelligence and augmented intelligence in medicine
2. Identify the current and future uses of AI tools in medicine and radiology
3. Discuss how AI could lead to healthcare disparities
4. Discuss the legal, ethical, and liability concerns of AI applications in medicine


INSIDE: A 14-PAGE SPECIAL REPORT ON FINANCIAL TECHNOLOGY
The Economist
How to fix America's inner cities
The self-service economy
Time to open up Indonesia
Inside the anti-bribery business
Why humans cause heatwaves
Artificial Intelligence
The promise and the peril


The Economist
Theresa May v Brussels
Ten years on: banking after the crisis
South Korea's unfinished revolution
Biology, but without the cells
The world's most valuable resource
amazon
Google


The Economist
Inside China's Ministry of Truth
Trump in trouble
Who are the Niger Delta Avengers?
The flaws in executive pay
Motorcycles that fly
March of the machines
A SPECIAL REPORT ON ARTIFICIAL INTELLIGENCE

AI
Armageddon?

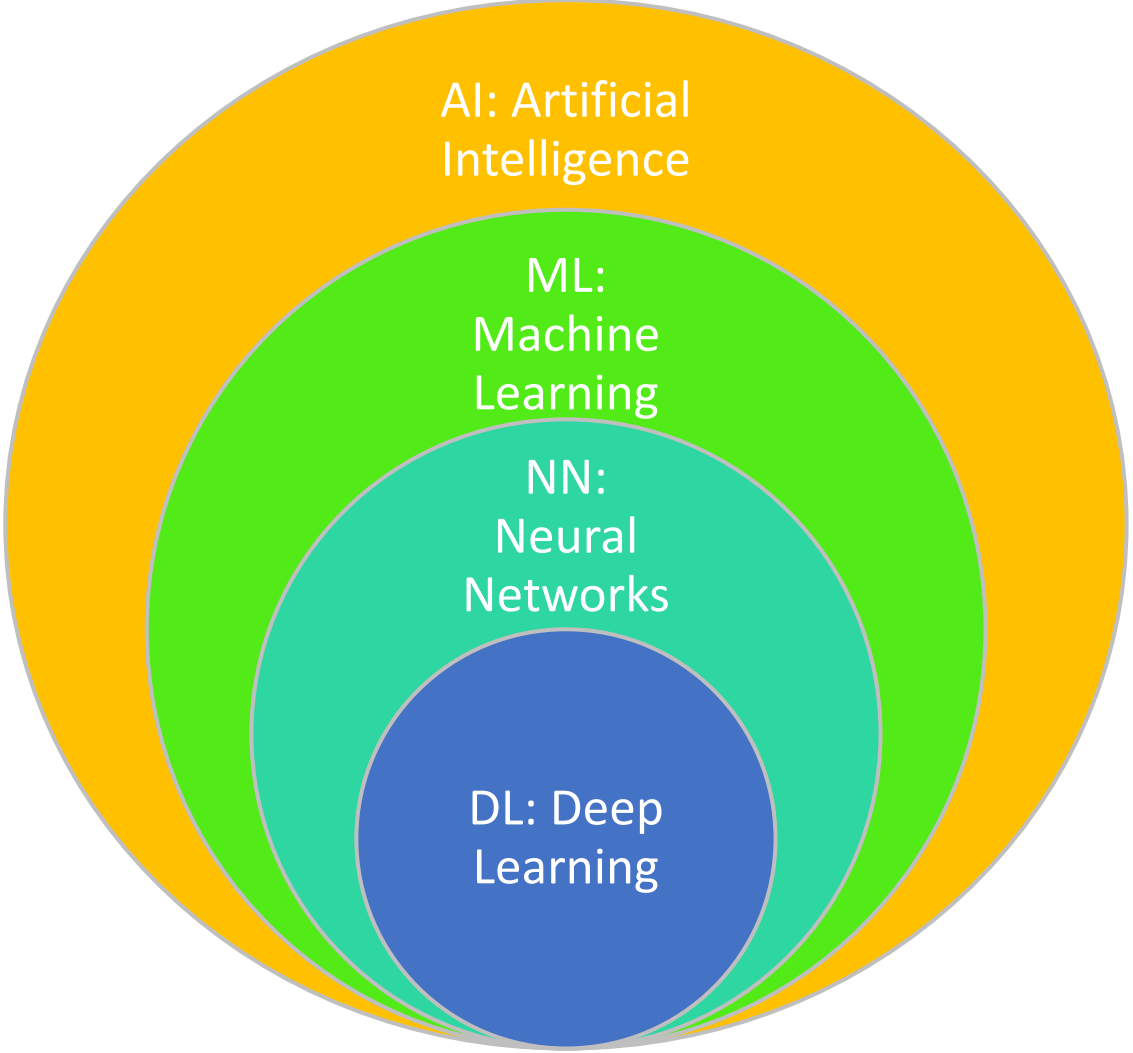
Elon Musk: ‘Mark my words — A.I. is far more dangerous than nukes’

Catherine Clifford | 1:22 PM ET Tue, 13 March 2018

Or Fake News?

Facebook CEO Mark Zuckerberg: Elon Musk's doomsday AI predictions are 'pretty irresponsible'

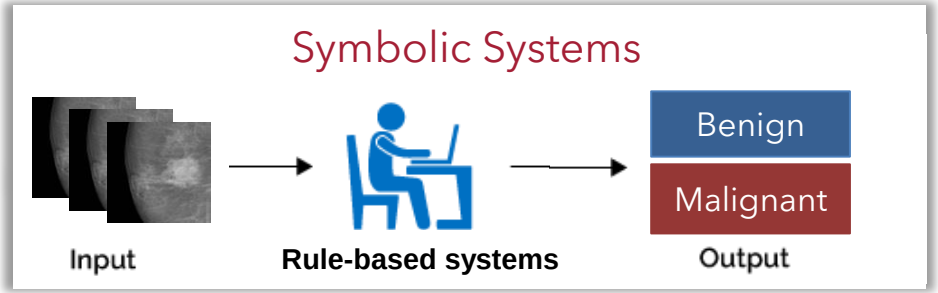
Catherine Clifford | 1:05 PM ET Mon, 24 July 2017



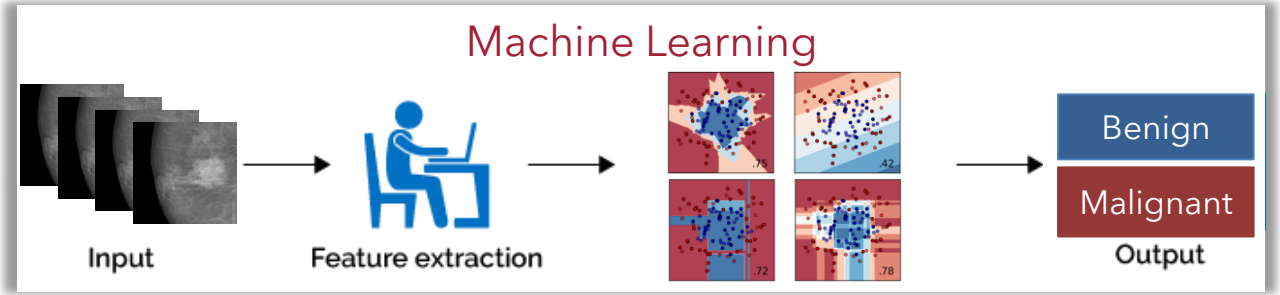
Definitions

- AI: When computers do things that normally require human intelligence
- ML: Rapid automatic construction of algorithms from data
- NN: Powerful form of machine learning
- DL: Neural networks with many layers

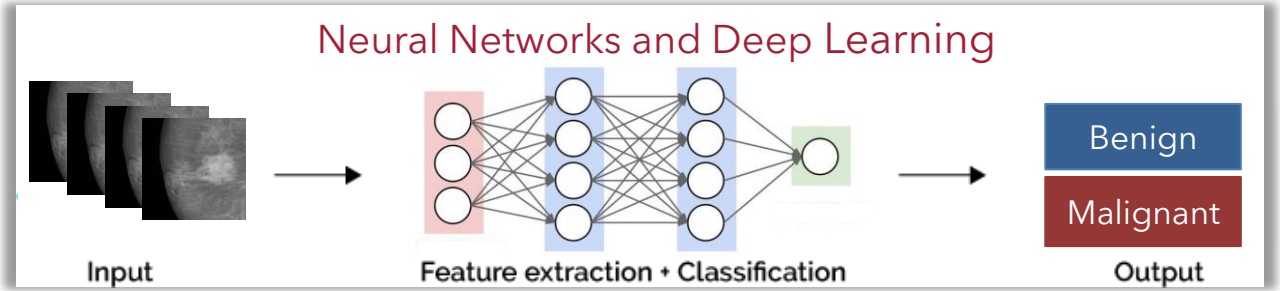
AI v1.0:
1950s-1980s



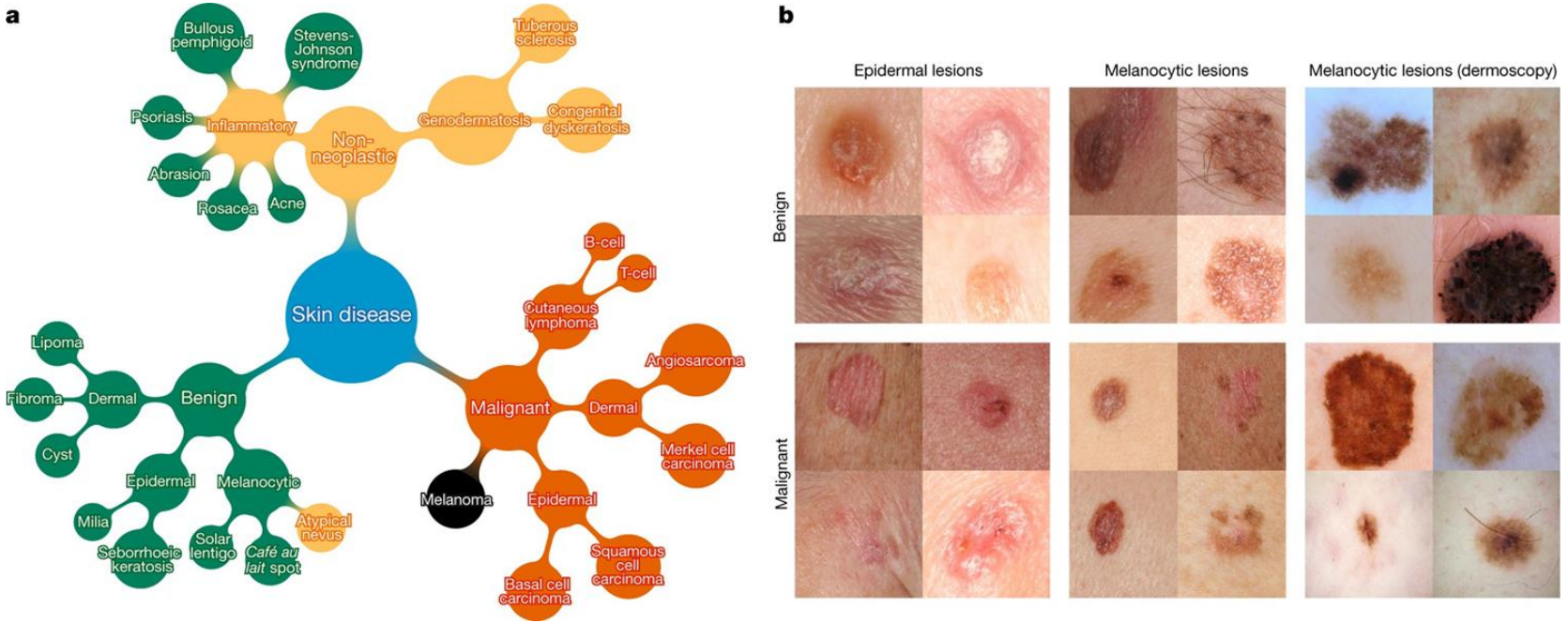
AI v2.0:
1980s-2010s



AI v3.0:
2010-
present

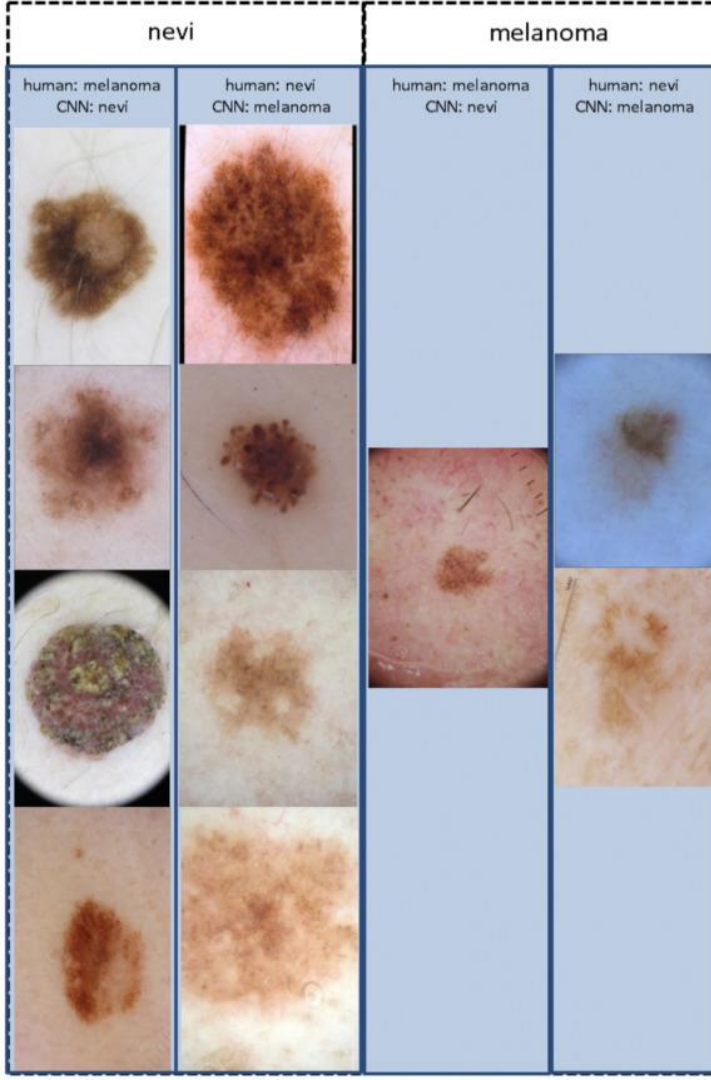


Applications of AI in Medicine



Dermatology

- In a head-to-head comparison, the AI outperformed 136 of 157 participating dermatologists.
- AI outperformed dermatologists of all hierarchical subgroups (from junior to chief physicians) in dermoscopic melanoma image classification.



European Journal of Cancer. April 2019.
DOI: https://doi.org/10.1016/j.ejca.2019.04.001

Ophthalmology

Original Investigation | Innovations in Health Care Delivery

FREE

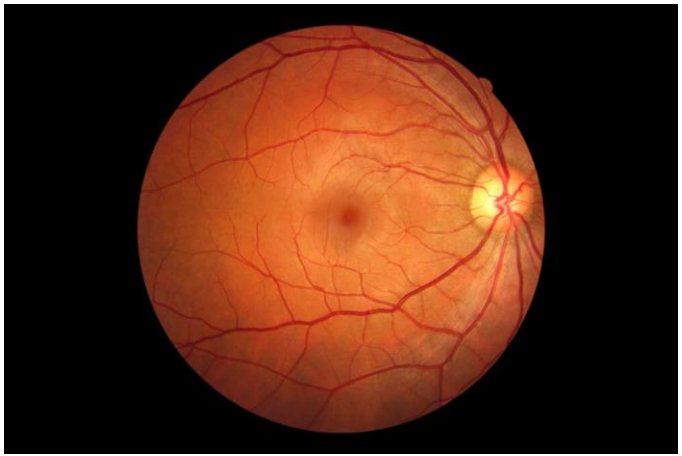
December 13, 2016

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD[1]; Lily Peng, MD, PhD[1]; Marc Coram, PhD[1]; et al

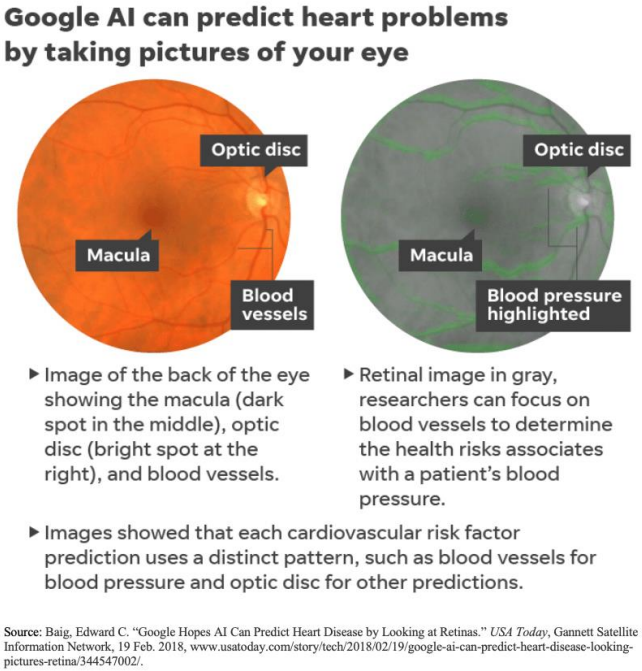
» Author Affiliations | Article Information

JAMA. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216



- Prevalence of diabetic retinopathy is approximately 28.5% in the United States and 18% in India.
- Annual screening for those with no retinopathy or mild diabetic retinopathy
- Repeat examination in 6 months for moderate diabetic retinopathy, and an ophthalmologist referral for treatment evaluation within a few weeks to months for severe or worse diabetic retinopathy or the presence of referable diabetic macular edema.

Eye Predicts the Heart



- AI can use retinal images to predict…
  - Patient's age
  - Gender
  - Smoking status
  - Systolic blood pressure
  - Cardiovascular risk factors
  - Risk of major adverse cardiac events occurring over the next five years

Cardiology

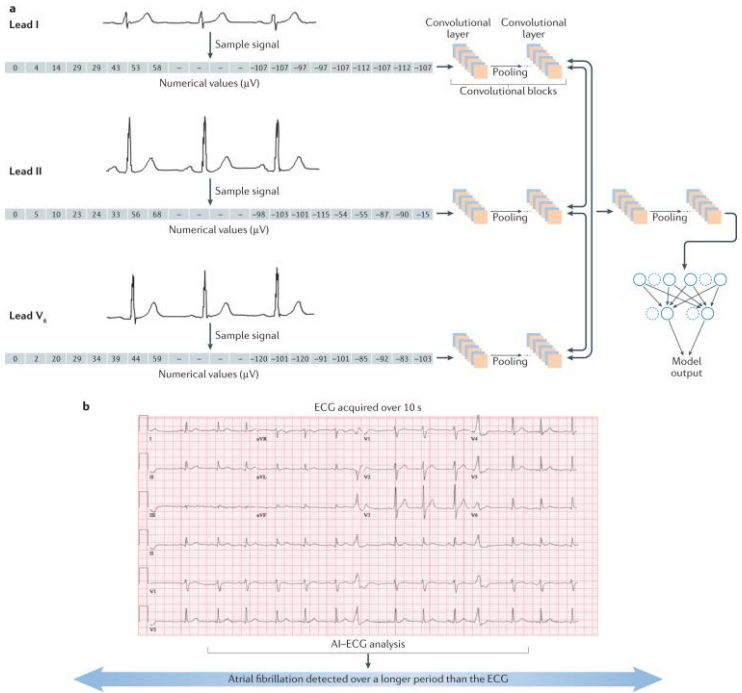
Review Article | Published: 01 February 2021

Artificial intelligence-enhanced electrocardiography in cardiovascular disease management

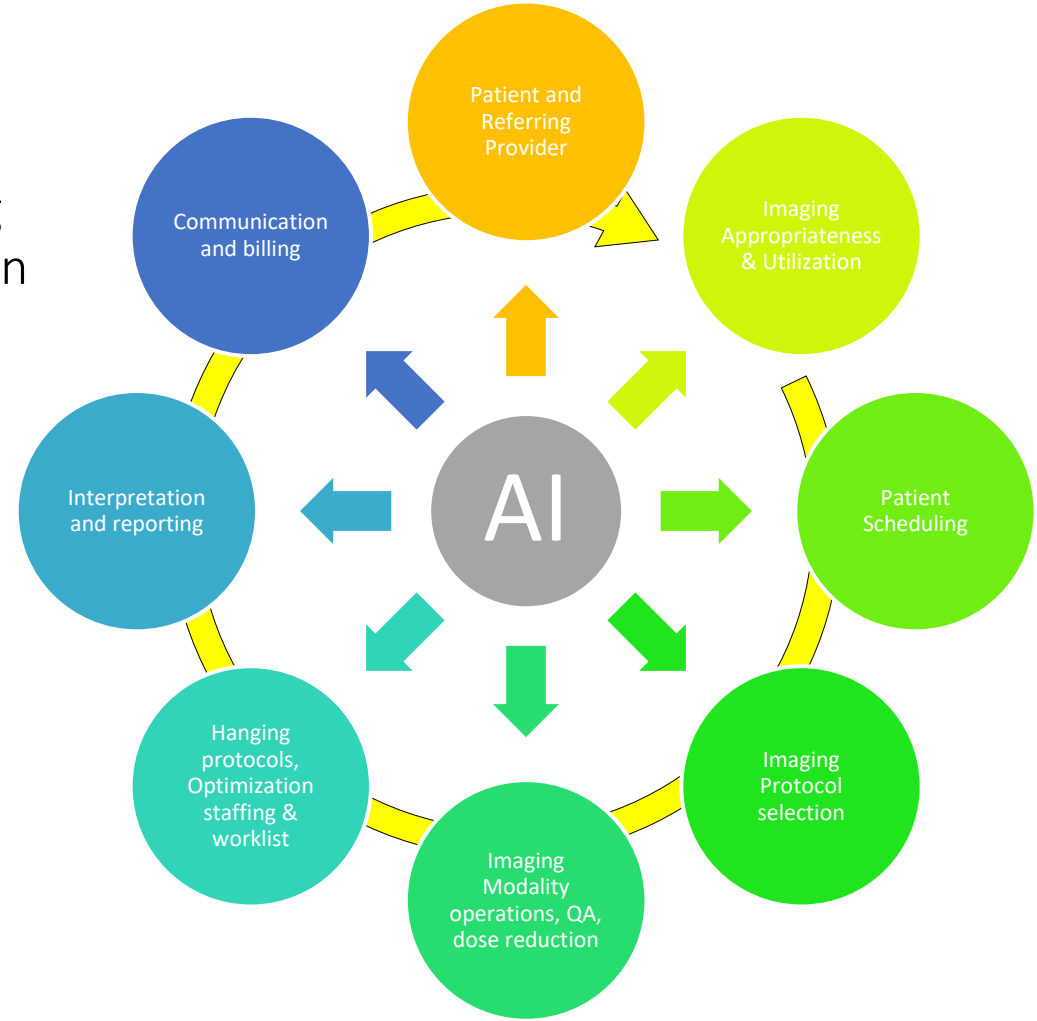
Konstantinos C. Siontis, Peter A. Noseworthy, Zachi I. Attia & Paul A. Friedman

Nature Reviews Cardiology **18**, 465–478 (2021) | Cite this article

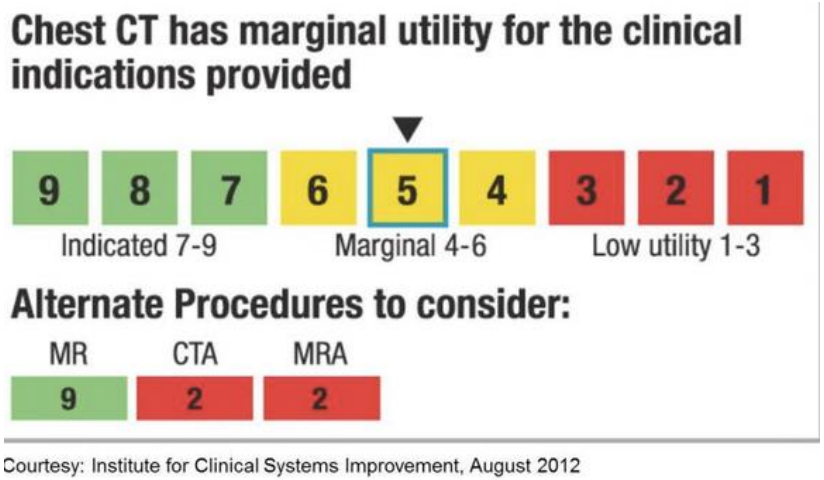
- Detect asymptomatic left ventricular dysfunction, silent atrial fibrillation, hypertrophic cardiomyopathy and an individual's age, sex and race on the basis of the ECG alone.
- Detect other cardiac conditions, such as aortic valve stenosis and amyloid heart disease, are in active development.
- Standard 12-lead ECG or to data obtained from single-lead or multilead mobile or wearable ECG technologies.



AI Imaging Value Chain



Radiology Order Decision Support



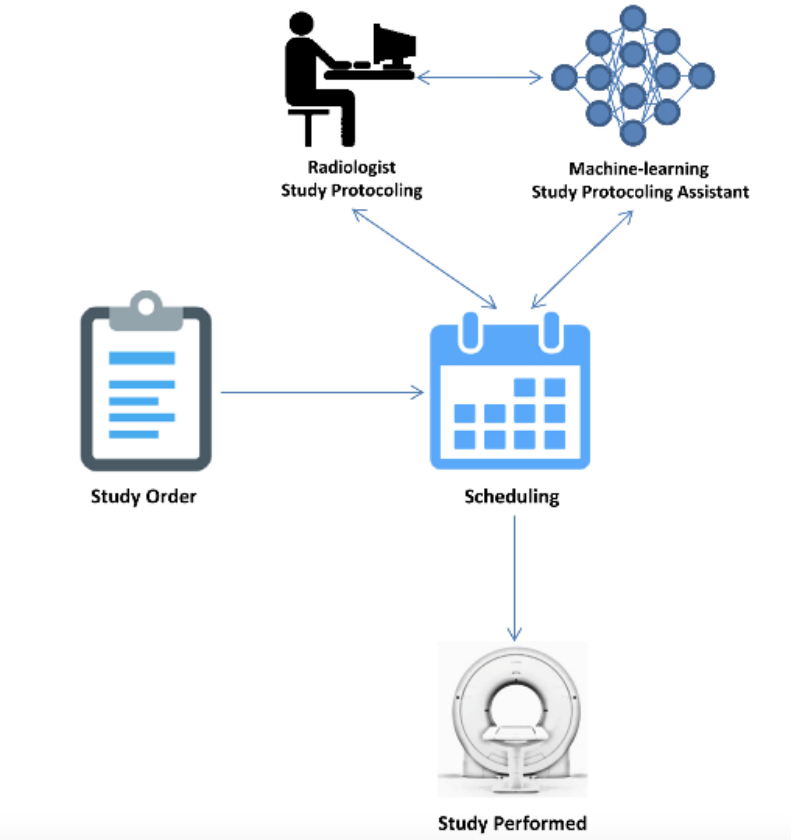
Current: Rule-based

Future?: AI-based

"1% of CT chest exams in those with similar clinical parameters will be positive for PE"

Imaging Protocol selection

Study Protocoling



MR Neuro/ENT To Protocol: 523	14
MRI MSK To Protocol: 176	1
MR Body To Protocol: 65	
CT Neuro/ENT To Protocol: 110	6
CT MSK To Protocol: 12	
CT Body To Protocol: 23	4
CT Cardiothoracic to Protocol: 116	5
CT Body/Chest All Unfiltered to Protocol: 37	9

Imaging Modality operations, QA, dose reduction

200x Low Dose PET

Ground Truth **Low-dose** **DL Recon**

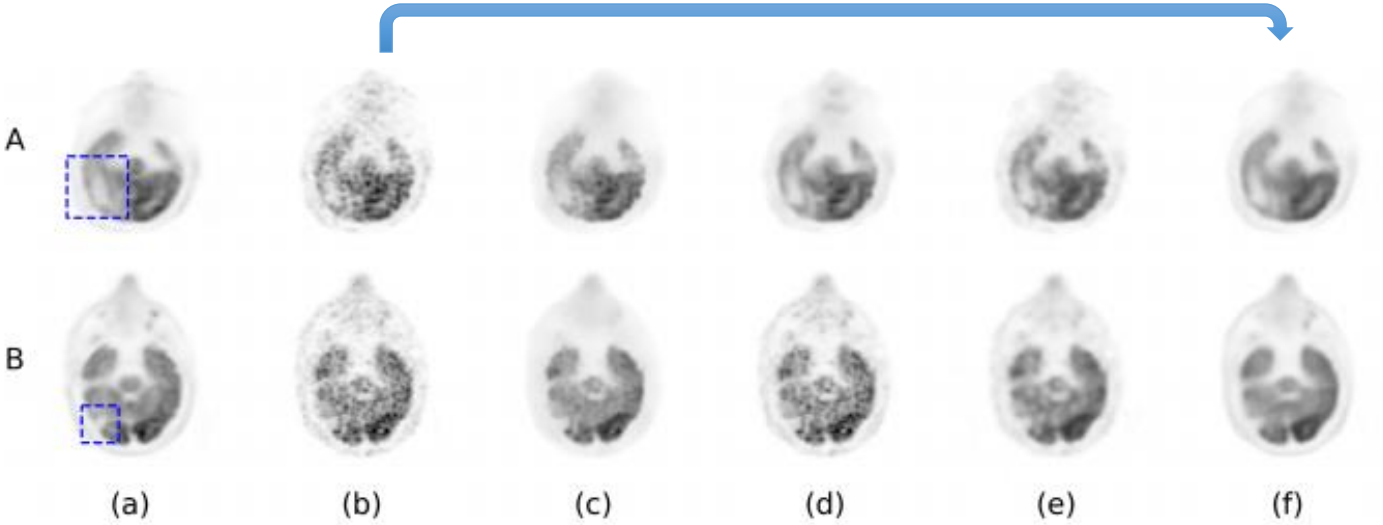
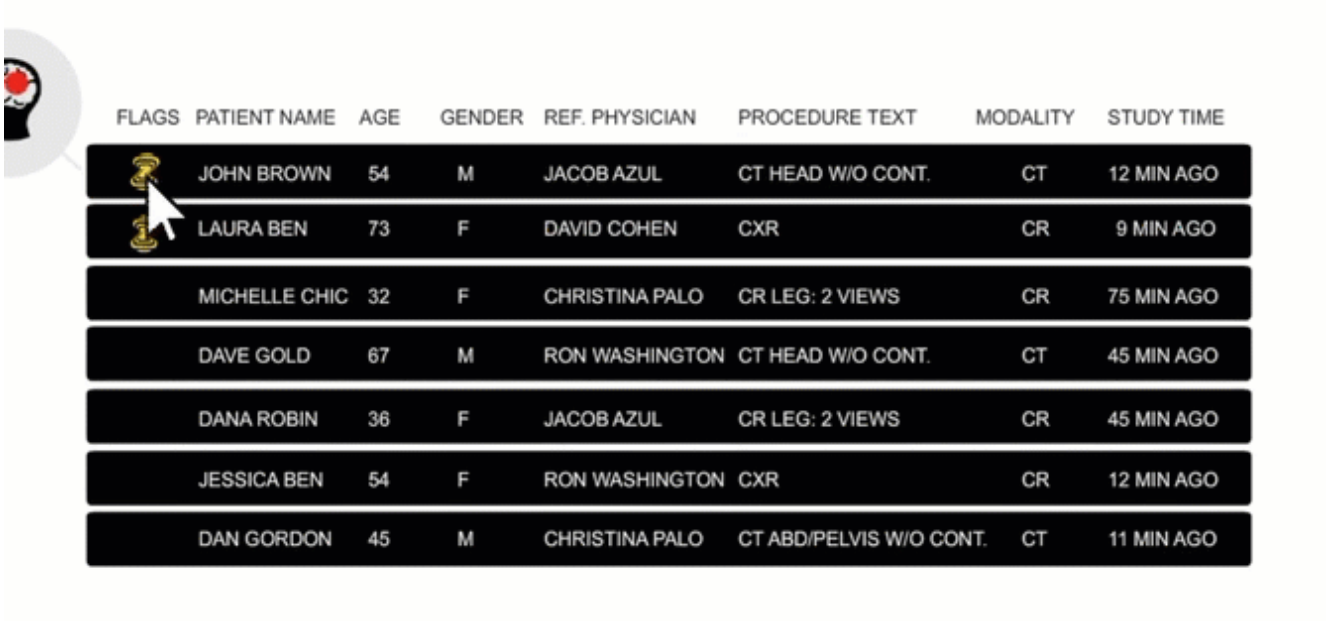


Fig. 6. Results from different methods for comparison. (a) standard-dose, (b) low-dose, (c) NLM, (d)BM3D, (e) AC-Net, and (f) proposed.

Worklist Triage

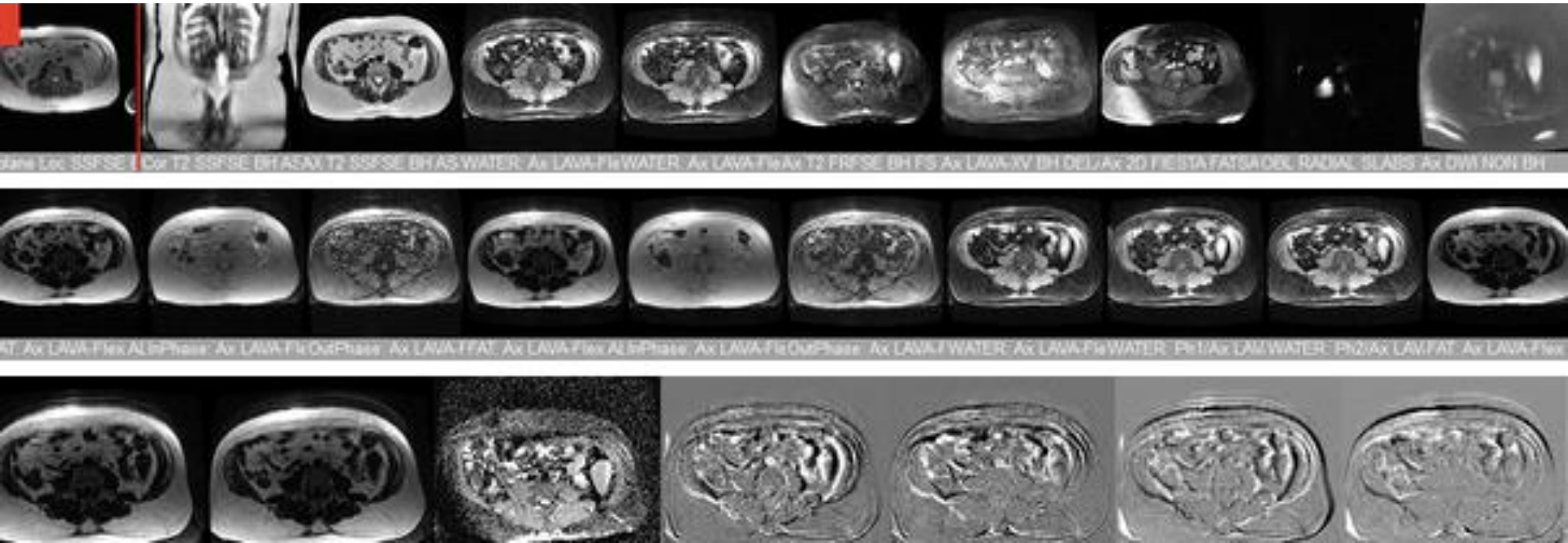


FLAGS	PATIENT NAME	AGE	GENDER	REF. PHYSICIAN	PROCEDURE TEXT	MODALITY	STUDY TIME
	JOHN BROWN	54	M	JACOB AZUL	CT HEAD W/O CONT.	CT	12 MIN AGO
	LAURA BEN	73	F	DAVID COHEN	CXR	CR	9 MIN AGO
	MICHELLE CHIC	32	F	CHRISTINA PALO	CR LEG: 2 VIEWS	CR	75 MIN AGO
	DAVE GOLD	67	M	RON WASHINGTON	CT HEAD W/O CONT.	CT	45 MIN AGO
	DANA ROBIN	36	F	JACOB AZUL	CR LEG: 2 VIEWS	CR	45 MIN AGO
	JESSICA BEN	54	F	RON WASHINGTON	CXR	CR	12 MIN AGO
	DAN GORDON	45	M	CHRISTINA PALO	CT ABD/PELVIS W/O CONT.	CT	11 MIN AGO

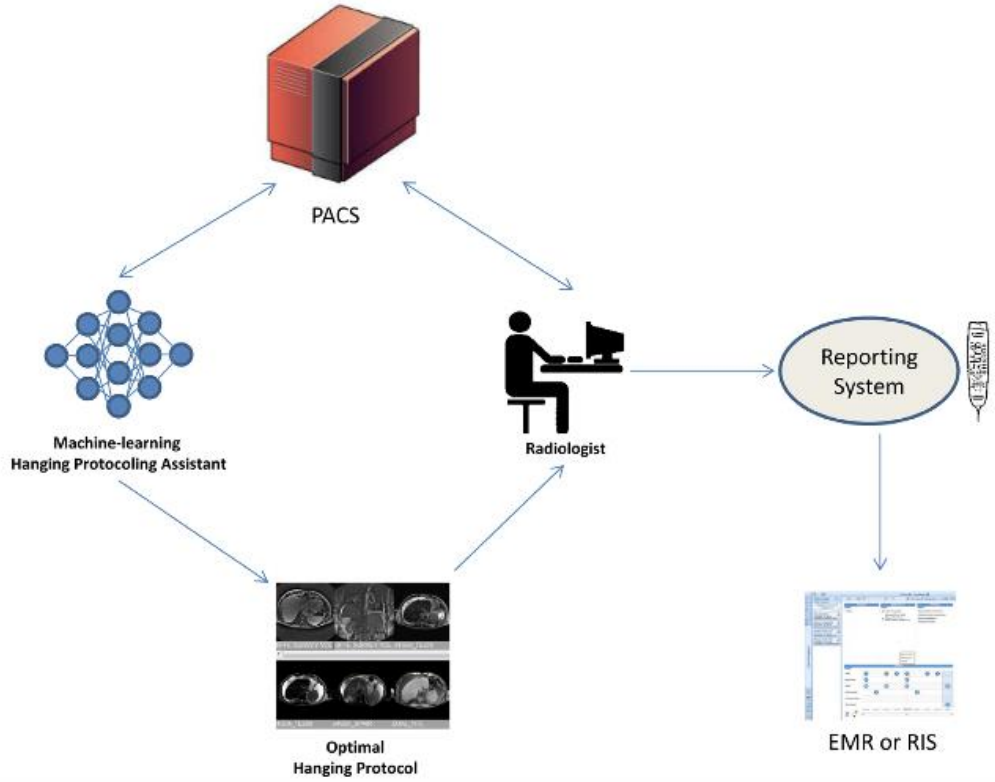
Hanging protocols, Optimization staffing & worklist

Hanging Protocols

Hanging protocols, Optimization staffing & worklist



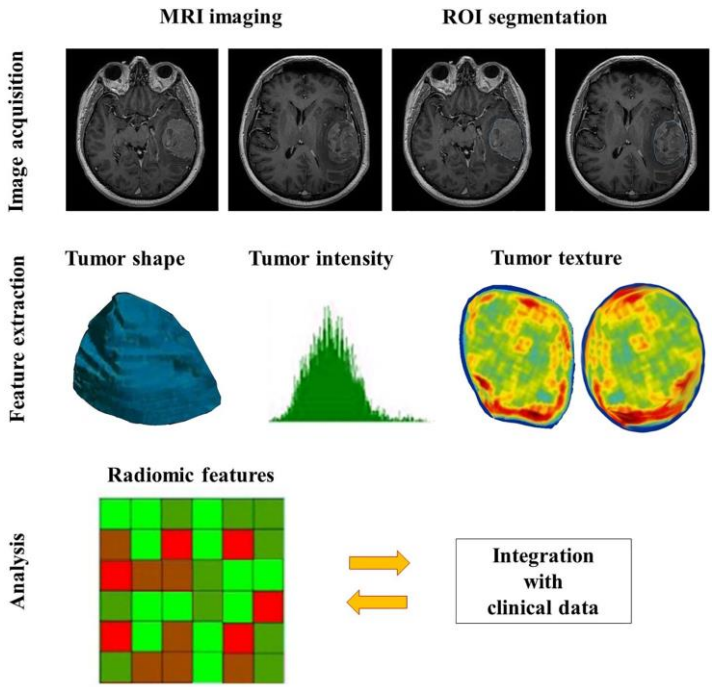
ML: Aid Hanging Protocols



Hanging protocols, Optimization staffing & worklist

Lakhani P, Prater AB, Hutson RK, Andriole KP, Dreyer KJ, Morey J, Prevedello LM, Clark TJ, Geis JR, Itri JN, Hawkins CM. Machine learning in radiology: applications beyond image interpretation. Journal of the American College of Radiology. 2018 Feb 1;15(2):350-9.

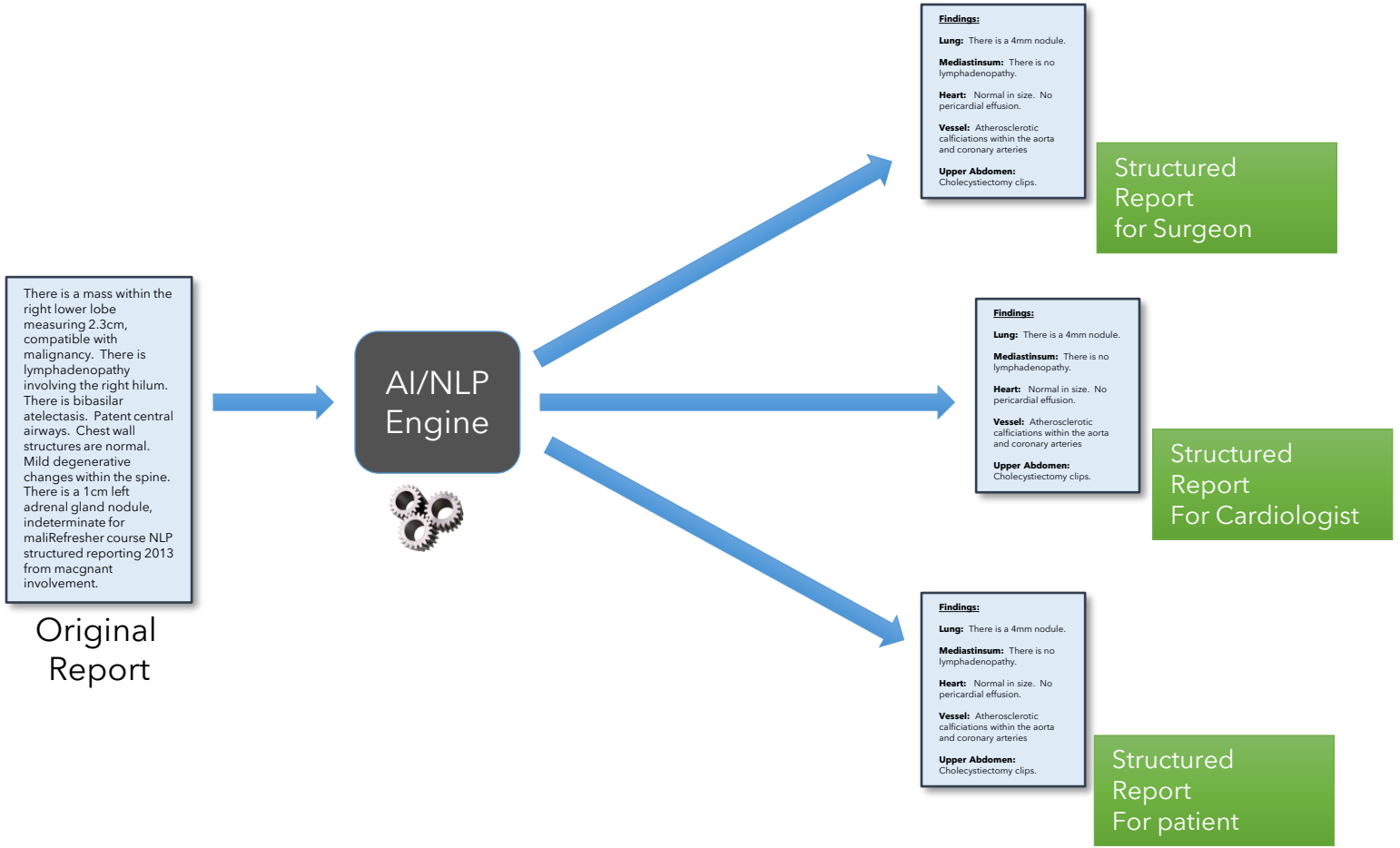
Radiology and Radiomics (pixel biopsy)



- Radiomics (as applied to radiology) is a field of medical study that aims to extract a large number of quantitative features from medical images using data characterization algorithms.
- Data is assessed for improved decision support.

Interpretation and reporting

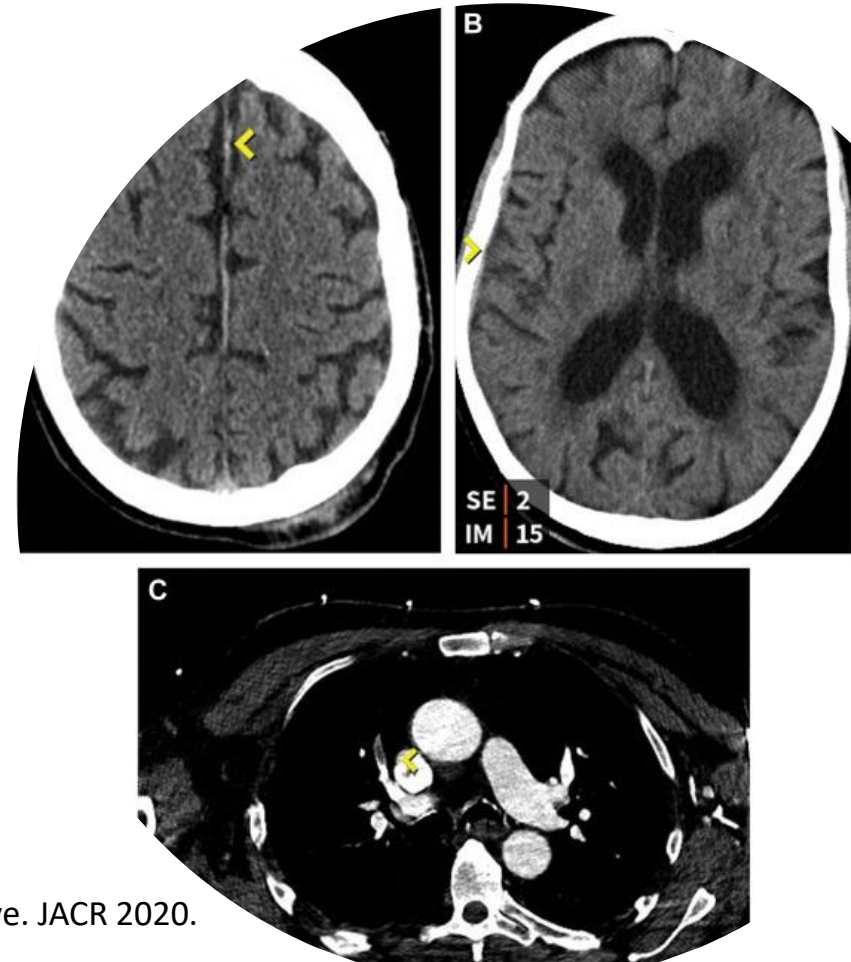
Translation of Radiology Reports



Interpretation and reporting

Peer Learning and Teaching Files

- 4.3% of pulmonary emboli in CT pulmonary angiogram studies and 2.4% of intracranial hemorrhages in noncontrast head CT studies were not mentioned by the radiologist.
- Collect these missed cases and utilize them for peer learning by identifying trends and presenting the cases and insights to our radiologists in monthly peer learning conferences.
- Instead of having radiologists evaluate a 3% to 5% sample of studies as is typical with traditional peer review, AI algorithms could review nearly 100% of prior examinations providing a far greater sample of cases for radiologist education.



Kottler. Artificial Intelligence: A Private Practice Perspective. JACR 2020.

Curt Langlotz
@curtlanglotz

Will #AI replace radiologists? The answer is NO. But rads who use #AI will replace rads who don't @RSNAInformatics @SIIM_Tweets

12:55 AM · Feb 8, 2017 · Hootsuite

88 Retweets **13** Quote Tweets **113** Likes

Neurology/Stroke

Viz.ai Granted Medicare New Technology Add-on Payment

Viz.ai, the leading AI stroke platform, is the first AI software to receive a Medicare New Technology Add-on Payment, helping hospitals adopt advanced technology to improve stroke care

September 2020

- In the US, stroke is the **number one cause of long-term disability** but is a treatable condition if identified early enough.

- CMS has granted Viz.ai the first New Technology Add-on Payment (NTAP) for artificial intelligence software. NTAP, part of the CMS Inpatient Prospective Payment System (IPPS), was set up to support the adoption of cutting-edge technologies that have demonstrated substantial clinical improvement and ensure early availability to Medicare patients.

- Viz LVO has been granted a New Technology Add on Payment of up to $1,040 per use in patients with suspected strokes.

Viz VIEW

Mobile DICOM Image Viewer

Viz LVO

A.I. Powered Detection of Suspected LVOs

Viz.ai

Viz CTP

Automated CTP Analysis

Viz HUB

HIPAA-Compliant Messaging

Regulation and Oversight of AI in Medicine

AMA Passes First Policy Recommendations on Augmented Intelligence

For immediate release: Jun 14, 2018

Artificial vs Augmented

Artificial intelligence constitutes a host of computational methods that produce systems that perform tasks normally requiring human intelligence.

In healthcare, a more appropriate term is **augmented intelligence**, reflecting the enhanced capabilities of human clinical decision-making when coupled with these computational methods and systems.

Augmented Intelligence

Original Investigation | Imaging

March 2, 2020

Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms

Thomas Schaffter, PhD[1]; Diana S. M. Buist, PhD, MPH[2]; Christoph I. Lee, MD, MS[3]; et al

» Author Affiliations | Article Information

JAMA Netw Open. 2020;3(3):e200265. doi:10.1001/jamanetworkopen.2020.0265

AMA AI Policy

Leverage its ongoing engagement in digital health and other priority areas for improving **patient outcomes** and **physicians' professional satisfaction** to help set priorities for health care AI.

Identify opportunities to **integrate** the perspective of practicing physicians into the development, design, validation and implementation of health care AI.

AMA AI Policy

Promote development of thoughtfully designed, high-quality, **clinically validated** health care AI	Is designed and evaluated in keeping with **best practices** in user-centered design, particularly for physicians and other members of the health care team
	Is **transparent**
	Conforms to leading standards for **reproducibility**
	Identifies and takes steps to address **bias** and avoids introducing or exacerbating health care **disparities** including when testing or deploying new AI tools on vulnerable populations
	Safeguards patients' and other individuals' **privacy** interests and preserves the security and integrity of personal information

AMA AI Policy

Encourage **education** for patients, physicians, medical students, other health care professionals, and health administrators to promote greater understanding of the promise and limitations of health care AI

Explore the **legal** implications of health care AI, such as issues of liability or intellectual property, and advocate for appropriate professional and governmental oversight for safe, effective, and equitable use of and access to health care AI

Version 2 - For Public Discussion

IEEE
Advancing Technology for Humanity

IEEE Ethically Aligned Design

- General Principles 20-32
- **Embedding Values into Autonomous Intelligent Systems** 33-54
- Methodologies to Guide Ethical Research and Design 55-72
- **Safety and Beneficence of Artificial General Intelligence** (AGI) and Artificial Superintelligence (ASI) 73-82
- **Personal Data and Individual Access Control** 83-112
- Reframing Autonomous Weapons Systems 113-130
- Economics/Humanitarian Issues 131-145
- Law 146-161

Embedding Values into Autonomous and Intelligent Systems

- Society **has not established universal standards or guiding principles** for embedding human values and norms into autonomous and intelligent systems (A/IS) today.
- If systems are instilled with increasing autonomy in making decisions and manipulating their environment, it is essential that they are **designed to adopt, learn, and follow the norms and values of the community they serve**.
- Actions should be **transparent** in signaling their norm compliance and, if needed, they must be able to explain their actions.

Embedding Values into Autonomous and Intelligent Systems

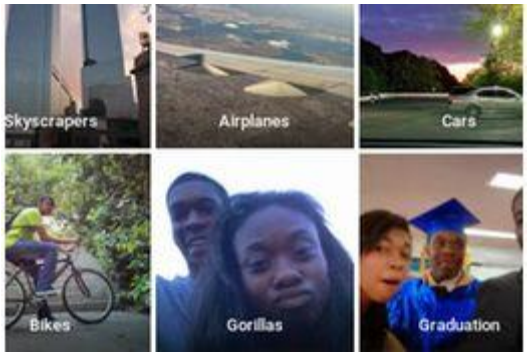
- Not all norms of a target community apply equally to human and artificial agents (one size does not fit all)
- A/IS can have **biases** that disadvantage specific groups
  - Biases may still emerge from imperfections in the norm identification process itself, from **unrepresentative training sets** for machine learning systems, and from programmers' and designers' **unconscious assumptions**
  - Unanticipated or undetected biases should be further reduced by **including members of diverse social groups** in both the planning and evaluation of A/IS and integrating community outreach into the evaluation process
  - Ensure A/IS works with different races, ethnicities, genders, ages, body shapes, or to people who use wheelchairs or prosthetics

THE WALL STREET JOURNAL.

DIGITS

Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms

By Alistair Barr
Updated July 1, 2015 3:41 pm ET



Safety and Beneficence of Artificial Intelligence

- As AI systems become more capable, as measured by the ability to optimize more complex objective functions with greater autonomy across a wider variety of domains, **unanticipated or unintended behavior** becomes increasingly dangerous.
- Retrofitting safety into future more generally capable AI systems may be difficult.
- Researchers and developers will confront a progressively more complex set of ethical and technical safety issues in the development and deployment of increasingly autonomous and capable AI systems.

The New York Times

U.S. Will Investigate Tesla's Autopilot System Over Crashes With Emergency Vehicles

It will be the broadest look yet at Tesla's assisted-driving technology. The National Highway Traffic Safety Administration has the authority to force a recall or require new safety features.

A Tesla Model S crashed into a fire engine on Interstate 405 in Culver City, Calif., in 2018. A government report said the driver of the Tesla had been using the car's Autopilot system. KCBS-TV, via Associated Press

By Neal E. Boudette and Niraj Chokshi

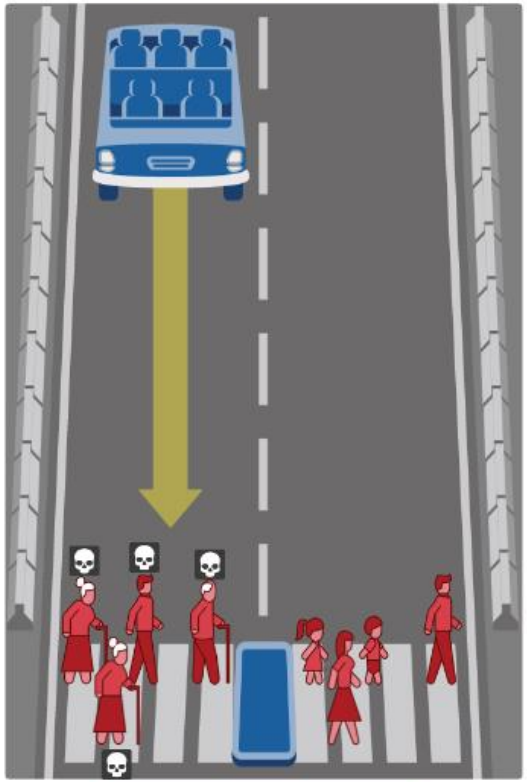
Published Aug. 16, 2021 Updated Sept. 1, 2021

What should the self-driving car do?

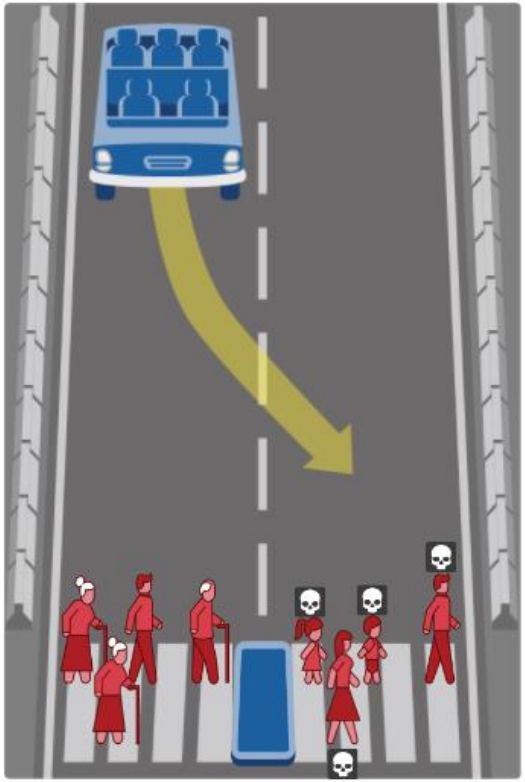
In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in ...

Dead:

- 2 elderly women
- 1 man
- 1 elderly man



Hide Description



Hide Description

In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane. This will result in ...

Dead:

- 1 girl
- 1 boy
- 1 man
- 1 woman

- AI must learn to diagnose disease on large data sets, and if that data doesn't include enough patients from a particular background, it **won't be as reliable** for them.

- Some facial recognition programs **incorrectly classify** less than 1 percent of light-skinned men but more than one-third of dark-skinned women
  - **What happens when we rely on such algorithms to diagnose melanoma on light versus dark skin?**

- Most widely used cardiovascular risk score, developed using data from mostly white male patients, can be **less precise** for women and minorities.

The New York Times

OPINION

A.I. Could Worsen Health Disparities

In a health system riddled with inequity, we risk making dangerous biases automated and invisible.

Jan. 31, 2019

Jenice Kim

By Dhruv Khullar
Dr. Khullar is an assistant professor of health care policy and research.

AI Could Worsen Health Disparities

AI is trained on **real-world data**, it risks incorporating, entrenching and perpetuating the economic and social biases that contribute to health disparities in the first place.

- AI programs used to help judges predict which criminals are most likely to reoffend have shown troubling racial biases, as have those designed to help child protective services decide which calls require further investigation.

In medicine, unchecked AI could create self-fulfilling prophesies that confirm our **pre-existing biases**, especially when used for conditions with complex trade-offs and high degrees of uncertainty.

- Poorer patients do worse after organ transplantation or after receiving chemotherapy for end-stage cancer, machine learning algorithms may conclude such patients are less likely to benefit from further treatment — and recommend against it.

AI Could Worsen Health Disparities

- Neutral AI has the potential to worsen disparities if its implementation has **disproportionate effects for certain groups**.
    - Consider a program that helps doctors decide whether a patient should go home or to a rehab facility after knee surgery (a decision imbued with uncertainty but has real consequences: Evidence suggests discharge to an institution is associated with higher costs and higher risk of readmission).
    - If an algorithm incorporates residence in a low-income neighborhood as a marker for poor social support, it may recommend minority patients go to nursing facilities instead of receive home-based physical therapy.
    - A program designed to maximize efficiency or lower medical costs might discourage operating on those patients altogether.

Safety and Beneficence of Artificial Intelligence

Future AI systems may have the capacity to **impact the world** on the scale of the agricultural or industrial revolutions.

Personal Data and Individual Access Control

- **Autonomous and Intelligent systems** (A/IS) are developing faster than the supporting standards and regulation required for transparency and societal protections can keep pace.
- Impact of these systems on society is **direct** and considerable.
- A/IS require data to fuel learning and inform automatic decision-making.
- Personal data, or personally identifiable information, known as PII is defined as any data that can be reasonably linked to an individual based on their unique physical, digital, or virtual identity.

Privacy

Ethical considerations regarding data are often focused largely on issues of privacy.

What rights should a person have to keep certain information to themselves or have input into how it is shared?

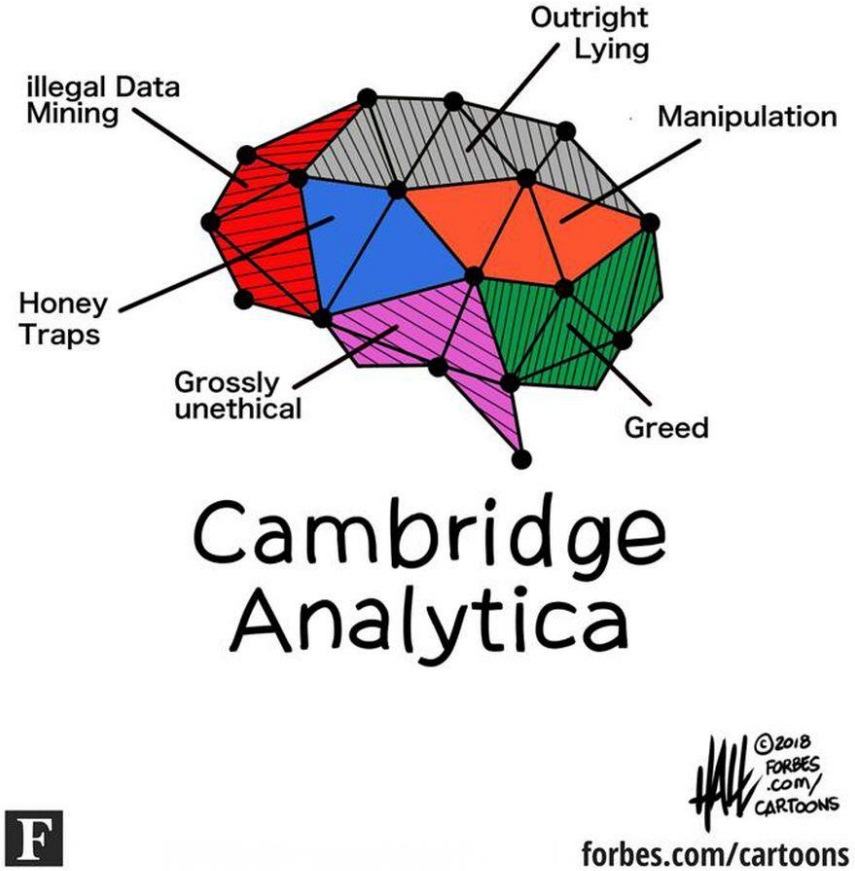
Individuals lack clarity around how to access, organize, and share their data to ensure unintended consequences are not encountered.

Privacy

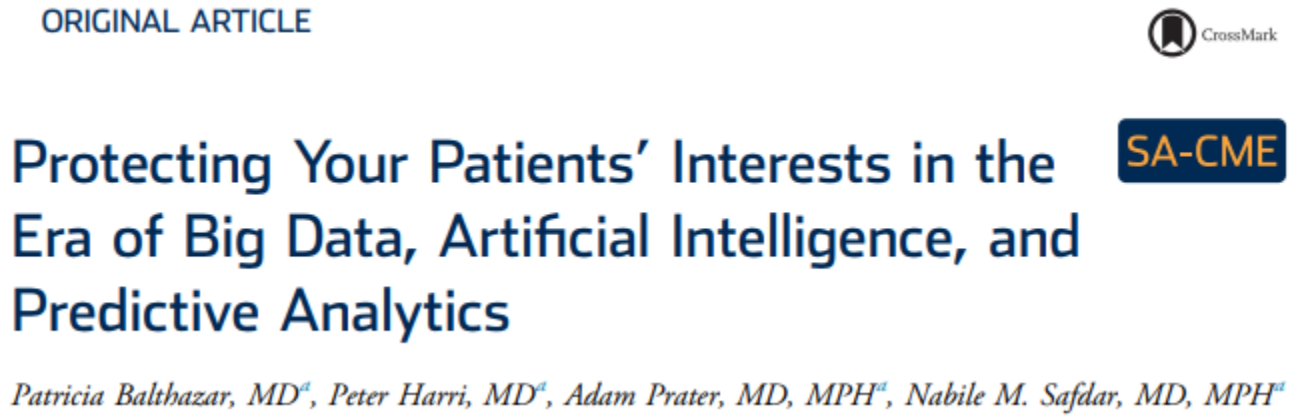
Fundamental need for people to have the right to define access and provide informed consent with respect to the use of their personal data (as they do in the physical world).

Individuals require mechanisms to help curate their unique identity and personal data in conjunction with policies and practices that make them explicitly aware of consequences resulting from the bundling or resale of their personal information and life experiences.

What could go wrong?



JACR


ORIGINAL ARTICLE
CrossMark
SA-CME
Protecting Your Patients' Interests in the Era of Big Data, Artificial Intelligence, and Predictive Analytics
Patricia Balthazar, MD[a], Peter Harri, MD[a], Adam Prater, MD, MPH[a], Nabile M. Safdar, MD, MPH[a]

Privacy and Confidentiality

How do we keep data-driven insights about sensitive health issues **confidential**?

How do institutions **prevent the reidentification** of individuals from joining of data sets?

What is your **obligation** to notify a patient or subject of a health risk or propensity identified using big data or machine learning techniques?

Ownership of Data and Subsequently Developed Products

- Can patient data be **reused** for developing and validating advanced analytic methods?
- Can they be **shared or sold** for this purpose?
- If an app is developed and validated using patient data, should the app be sold for profit?

Informed Consent

What mechanisms are in place to exclude the data of individuals who do opt out?

What mechanisms are in place to allow patients to donate all their medical data for research?

Opting In

SYNC FOR SCIENCE

Helping patients share EHR data with researchers

What is Sync for Science (S4S)?

Sync for Science (S4S) is a national collaboration among electronic health record (EHR) vendors – including Allscripts, Cerner, eClinicalWorks, and Epic – and the National Institutes of Health (NIH), the Office of the National Coordinator for Health IT (ONC), and Harvard Medical School's Department of Biomedical Informatics.

Eric Topol ✓
@EricTopol

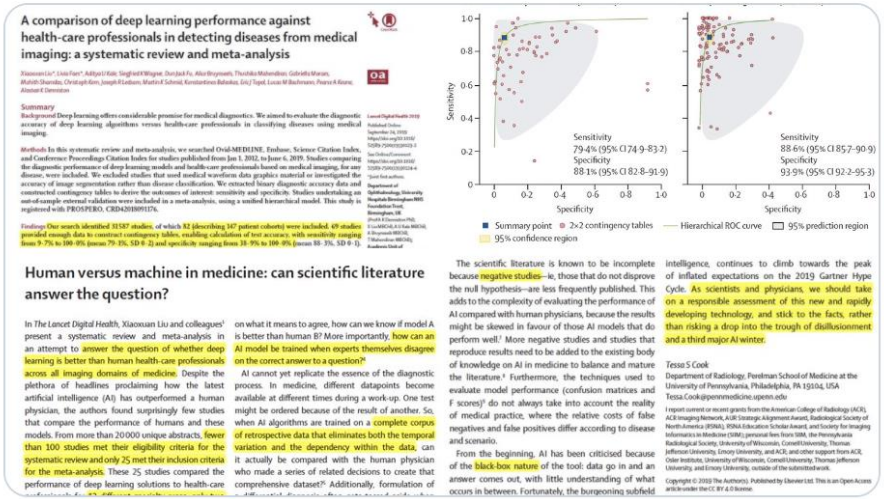
physician-scientist, author, editor

La Jolla, CA

stsiweb.org

Eric Topol ✓
@EricTopol

The 1st systematic review of deep learning #AI for medical imaging @LancetDigitalH shows how early we are in the science & validation by @DrXiaoLiu @FaesLiv @A_U_Kale @sktywagner @DunJackFu @Denniston_Ophth @pearsekeane me, et al; + @asset25's ★ editorial thelancet.com/journals/landi...



A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis

Summary

Background Deep learning offers considerable promise for medical diagnostics. We aimed to evaluate the diagnostic accuracy of deep learning algorithms versus health-care professionals in classifying diseases using medical imaging.

Methods In this systematic review and meta-analysis, we searched Ovid-MEDLINE, Embase, Science Citation Index, and Conference Proceedings Citation Index for studies published from Jan 1, 2012, to June 6, 2019. Studies comparing the diagnostic performance of deep learning models and health-care professionals based on medical imaging, for any disease, were included. We excluded studies that used medical waveform data graphics material or investigated the accuracy of image segmentation rather than disease classification. We extracted binary diagnostic accuracy data and constructed contingency tables to derive the outcomes of interest: sensitivity and specificity. Studies undertaking an out-of-sample external validation were included in a meta-analysis, using a unified hierarchical model. This study is registered with PROSPERO, CRD42018091176.

Findings Our search identified 31 587 studies, of which 82 (describing 147 patient cohorts) were included. 69 studies provided enough data to construct contingency tables, enabling calculation of test accuracy, with sensitivity ranging from 9·7% to 100·0% (mean 79·1%, SD 0·2) and specificity ranging from 38·9% to 100·0% (mean 88·3%, SD 0·1).

Sensitivity
79·4% (95% CI 74·9–83·2)
Specificity
88·1% (95% CI 82·8–91·9)

Sensitivity
88·6% (95% CI 85·7–90·9)
Specificity
93·9% (95% CI 92·2–95·3)

Summary point · 2×2 contingency tables · Hierarchical ROC curve · 95% prediction region · 95% confidence region

Human versus machine in medicine: can scientific literature answer the question?

In *The Lancet Digital Health*, Xiaoxuan Liu and colleagues[1] present a systematic review and meta-analysis in an attempt to answer the question of whether deep learning is better than human health-care professionals across all imaging domains of medicine. Despite the plethora of headlines proclaiming how the latest artificial intelligence (AI) has outperformed a human physician, the authors found surprisingly few studies that compare the performance of humans and these models. From more than 20 000 unique abstracts, fewer than 100 studies met their eligibility criteria for the systematic review and only 25 met their inclusion criteria for the meta-analysis. These 25 studies compared the performance of deep learning solutions to health-care

on what it means to agree, how can we know if model A is better than human B? More importantly, how can an AI model be trained when experts themselves disagree on the correct answer to a question?[2]

AI cannot yet replicate the essence of the diagnostic process. In medicine, different datapoints become available at different times during a work-up. One test might be ordered because of the result of another. So, when AI algorithms are trained on a complete corpus of retrospective data that eliminates both the temporal variation and the dependency within the data, can it actually be compared with the human physician who made a series of related decisions to create that comprehensive dataset?[3] Additionally, formulation of

The scientific literature is known to be incomplete because negative studies—ie, those that do not disprove the null hypothesis—are less frequently published. This adds to the complexity of evaluating the performance of AI compared with human physicians, because the results might be skewed in favour of those AI models that do perform well.[3] More negative studies and studies that reproduce results need to be added to the existing body of knowledge on AI in medicine to balance and mature the literature.[4] Furthermore, the techniques used to evaluate model performance (confusion matrices and F scores)[5] do not always take into account the reality of medical practice, where the relative costs of false negatives and false positives differ according to disease and scenario.

From the beginning, AI has been criticised because of the black-box nature of the tool: data go in and an answer comes out, with little understanding of what occurs in between. Fortunately, the burgeoning subfield

intelligence, continues to climb towards the peak of inflated expectations on the 2019 Gartner Hype Cycle. As scientists and physicians, we should take on a responsible assessment of this new and rapidly developing technology, and stick to the facts, rather than risking a drop into the trough of disillusionment and a third major AI winter.

Tessa S Cook
Department of Radiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
Tessa.Cook@pennmedicine.upenn.edu

I report current or recent grants from the American College of Radiology (ACR), ACR Imaging Network, AUR Strategic Alignment Award, Radiological Society of North America (RSNA), RSNA Education Scholar Award, and Society for Imaging Informatics in Medicine (SIIM); personal fees from SIIM, the Pennsylvania Radiological Society, University of Wisconsin, Cornell University, Thomas Jefferson University, Emory University, and ACR; and other support from ACR, Osler Institute, University of Wisconsin, Cornell University, Thomas Jefferson University, and Emory University, outside of the submitted work.

3:36 PM · Sep 24, 2019 · Twitter Web App

280 Retweets **18** Quote Tweets **438** Likes

A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis

Xiaoxuan Liu*, Livia Faes*, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, Alastair K Denniston

Summary

Background Deep learning offers considerable promise for medical diagnostics. We aimed to evaluate the diagnostic accuracy of deep learning algorithms versus health-care professionals in classifying diseases using medical imaging.

Methods In this systematic review and meta-analysis, we searched Ovid-MEDLINE, Embase, Science Citation Index, and Conference Proceedings Citation Index for studies published from Jan 1, 2012, to June 6, 2019. Studies comparing the diagnostic performance of deep learning models and health-care professionals based on medical imaging, for any disease, were included. We excluded studies that used medical waveform data graphics material or investigated the accuracy of image segmentation rather than disease classification. We extracted binary diagnostic accuracy data and constructed contingency tables to derive the outcomes of interest: sensitivity and specificity. Studies undertaking an out-of-sample external validation were included in a meta-analysis, using a unified hierarchical model. This study is registered with PROSPERO, CRD42018091176.

Findings Our search identified 31 587 studies, of which 82 (describing 147 patient cohorts) were included. 69 studies provided enough data to construct contingency tables, enabling calculation of test accuracy, with sensitivity ranging from 9·7% to 100·0% (mean 79·1%, SD 0·2) and specificity ranging from 38·9% to 100·0% (mean 88·3%, SD 0·1). An out-of-sample external validation was done in 25 studies, of which 14 made the comparison between deep learning models and health-care professionals in the same sample. Comparison of the performance between health-care professionals in these 14 studies, when restricting the analysis to the contingency table for each study reporting the highest accuracy, found a pooled sensitivity of 87·0% (95% CI 83·0–90·2) for deep learning models and 86·4% (79·9–91·0) for health-care professionals, and a pooled specificity of 92·5% (95% CI 85·1–96·4) for deep learning models and 90·5% (80·6–95·7) for health-care professionals.

Interpretation Our review found the diagnostic performance of deep learning models to be equivalent to that of health-care professionals. However, a major finding of the review is that few studies presented externally validated results or compared the performance of deep learning models and health-care professionals using the same sample. Additionally, poor reporting is prevalent in deep learning studies, which limits reliable interpretation of the reported diagnostic accuracy. New reporting standards that address specific challenges of deep learning could improve future studies, enabling greater confidence in the results of future evaluations of this promising technology.

Funding None.

Lancet Digital Health 2019

Published Online
September 24, 2019
https://doi.org/10.1016/S2589-7500(19)30123-2

See Online/Comment
https://doi.org/10.1016/S2589-7500(19)30124-4

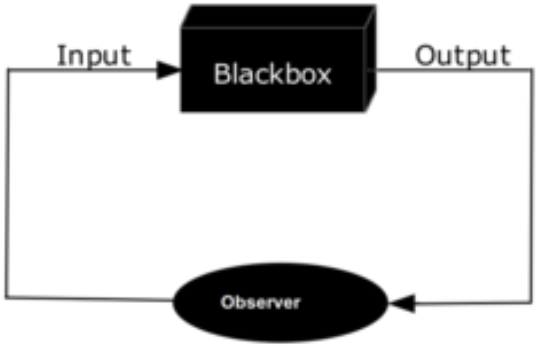
*Joint first authors.

Department of Ophthalmology, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK (Prof A K Denniston PhD, X Liu MBChB, A U Kale MBChB, A Bruynseels MBChB, T Mahendiran MBChB); **Academic Unit of Ophthalmology, Institute of Inflammation & Ageing, College of Medical and Dental Sciences** (X Liu, Prof A K Denniston, M Shamdas MBBS) **and Centre for Patient Reported Outcome Research, Institute of Applied Health Research** Prof A K Denniston), **University of Birmingham, Birmingham, UK; Medical Retina Department, Moorfields Eye Hospital NHS Foundation Trust, London, UK** (X Liu, L Faes MD, D J Fu PhD, G Moraes MD, C Kern MD, K Balaskas MD); **Eye Clinic, Cantonal Hospital of Lucerne, Lucerne, Switzerland** (L Faes, M K Schmid MD); **NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK** (S K Wagner BMBCh, K Balaskas, P A Keane MD, Prof A K Denniston); **University Eye Hospital, Ludwig Maximilian University of Munich, Munich, Germany** (C Kern); **DeepMind, London, UK** (J R Ledsam MBChB); **Scripps Research Translational Institute, La Jolla, California**

Introduction

The first paper indexed in MEDLINE with the MeSH term "artificial intelligence" (AI) dates back to 1951, when Fletcher described a tortoise robot in the seminal paper "Matter with mind; a neurological research robot".[1] Today, more than 16 000 peer-reviewed scientific papers are published in the AI field each year, with countless more in the lay press.[2] The application of AI has already started to transform daily life through applications such as photo captioning, speech recognition, natural language translation, robotics, and advances in self-driving cars.[3–9] Many people anticipate similar success in the health sphere, particularly in diagnostics, and some have suggested that AI applications will even replace whole medical disciplines or create new roles for doctors to fulfil, such as "information specialists".[10–12]

Medical imaging is one of the most valuable sources of diagnostic information but is dependent on human interpretation and subject to increasing resource challenges. The need for, and availability of, diagnostic images is rapidly exceeding the capacity of available specialists, particularly in low-income and middle-income



Black Box

- In science, computing, and engineering, a black box is a device, system or object which can be viewed in terms of its inputs and outputs (or transfer characteristics), without any knowledge of its internal workings.

- Its implementation is “opaque”.

AI “Black Box”

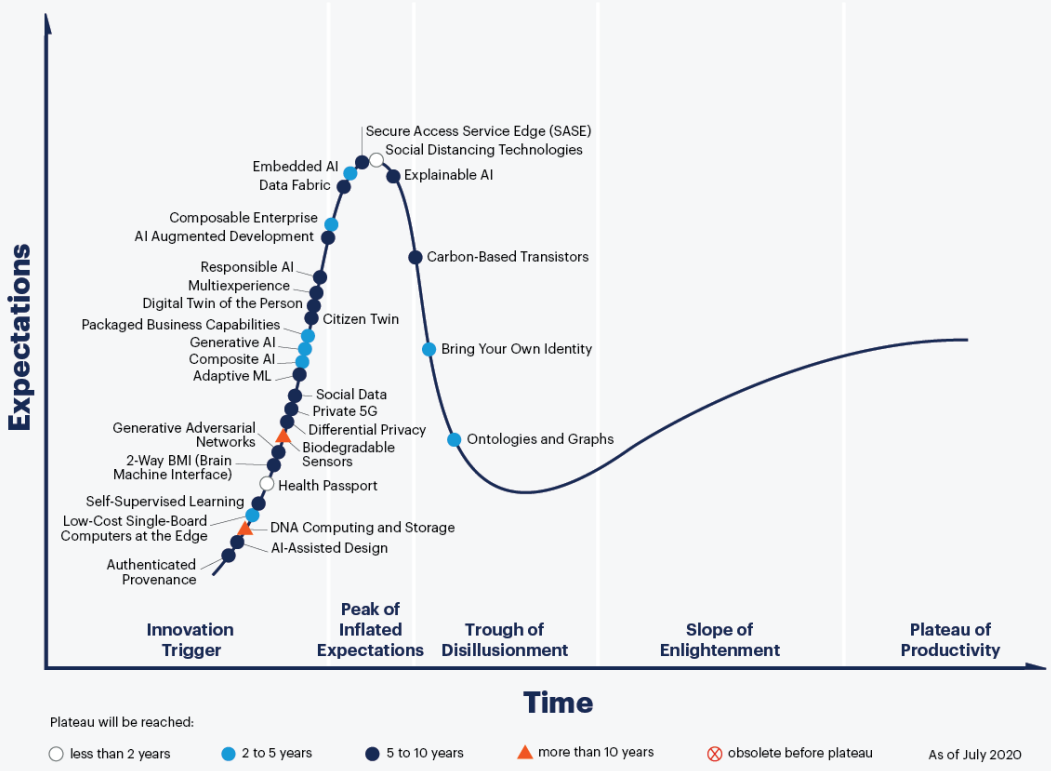
How do we know that the results of artificial intelligence algorithms are valid?

Were the data sets with which they were developed representative?

How can we defend the results of an algorithm directly affecting a patient’s health care, if no provider could completely comprehend how the algorithm reached its conclusion?

What happens if the algorithm makes a mistake or leads to an adverse patient event or irreversible damage?

Hype Cycle for Emerging Technologies, 2020



gartner.com/SmarterWithGartner

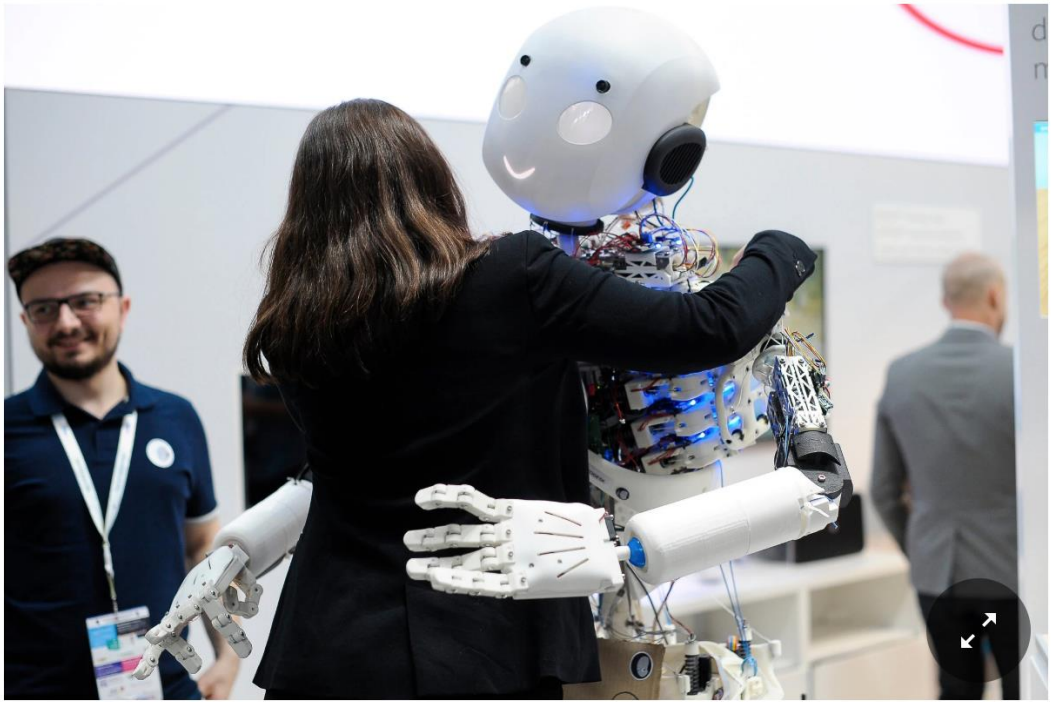
Gartner

Attacking the Black Box

Robust Physical-World Attacks on Machine Learning Models, Evtimov et al. https://arxiv.org/abs/1707.08945

The New York Times

Warnings of a Dark Side to A.I. in Health Care



Scientists worry that with just tiny tweaks to data, neural networks can be fooled into committing “adversarial attacks” that mislead rather than help.

Joan Cros/NurPhoto, via Getty Images

Adversarial attacks

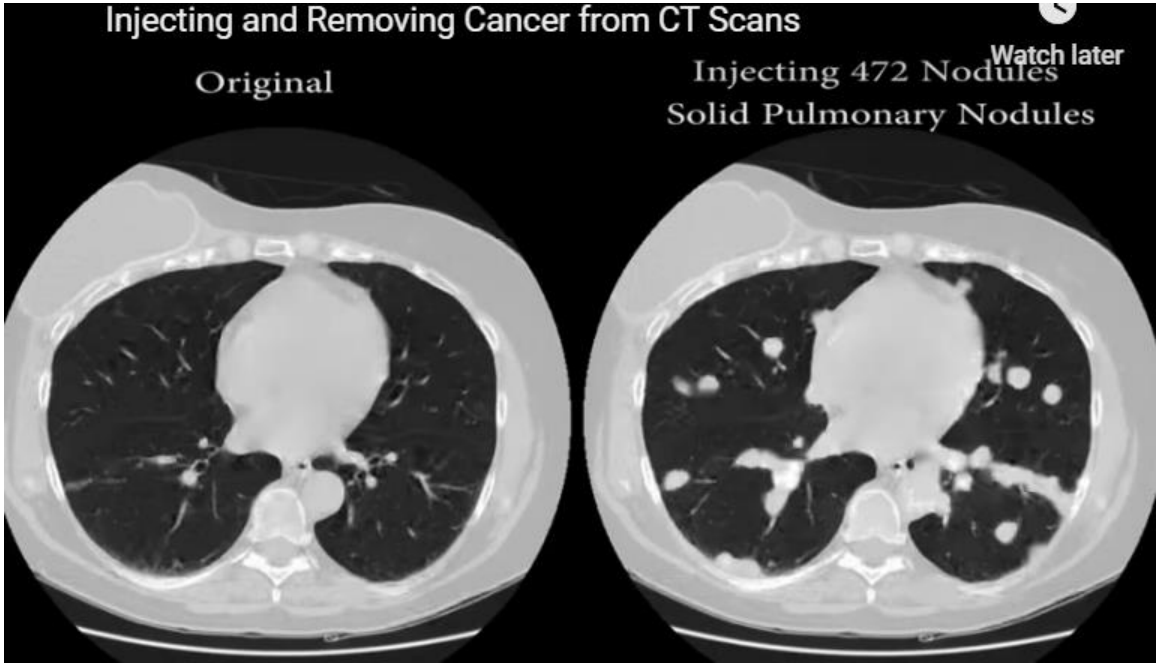
- **Manipulations** that can change the behavior of AI systems using tiny pieces of digital data.
    - By changing a few pixels on a lung scan, for instance, someone could fool an AI system into seeing an illness that is not really there, or not seeing one that is.
- Doctors, hospitals and other organizations could manipulate the AI in billing or insurance software in an effort to maximize the money coming their way.
- “The inherent ambiguity in medical information, coupled with often-competing financial incentives, allows for high-stakes decisions to swing on very subtle bits of information”

Manipulation

- Researchers demonstrated that, by changing a small number of pixels in an image of a benign skin lesion, a diagnostic AI system could be tricked into identifying the lesion as malignant.
    - Simply rotating the image could also have the same effect.
- Small changes to written descriptions of a patient's condition also could alter an AI diagnosis: "Alcohol abuse" could produce a different diagnosis than "alcohol dependence," and "lumbago" could produce a different diagnosis than "back pain."

Technology

Hospital viruses: Fake cancerous nodes in CT scans, created by malware, trick radiologists



https://www.extremetech.com/extreme/288968-deepfake-malware-can-trick-radiologists-into-believing-you-have-cancer

Deepfake

Malware they created would let attackers automatically add realistic, malignant-seeming growths to CT or MRI scans before radiologists and doctors examine them

Or it could remove real cancerous nodules and lesions without detection, leading to misdiagnosis and possibly a failure to treat patients who need critical and timely care

AI and the Physician

How does the AI algorithm influence the performance of the physician?

Does Physician + AI outperform just the Physician?

What is considered the “ground truth”?

How will the AI model be displayed?

Will the AI model learn over time?

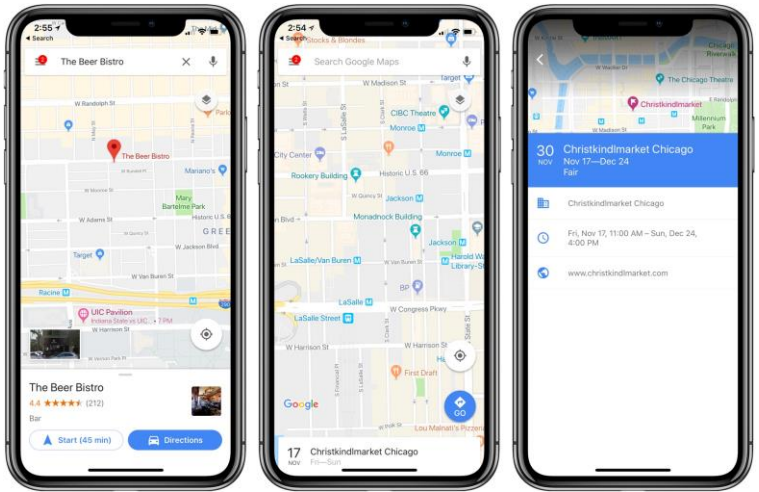
AI and the Physician (automation bias)

Article | Open Access | Published: 19 February 2021

Do as AI say: susceptibility in deployment of clinical decision-aids

Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J. Berkowitz, Eva Lermer, Joseph F. Coughlin, John V. Guttag, Errol Colak & Marzyeh Ghassemi

npj Digital Medicine **4**, Article number: 31 (2021) | Cite this article



- Physicians received chest X-rays and diagnostic advice, some of which was inaccurate, and were asked to evaluate advice quality and make diagnoses.
- All advice was generated by human experts, but some was labeled as coming from an AI system.
- As a group, radiologists rated advice as lower quality when it appeared to come from an AI system; physicians with less task-expertise did not.
- **Diagnostic accuracy was significantly worse** when participants received inaccurate advice, regardless of the purported source.
- Important considerations for how advice, AI and non-AI, should be deployed in clinical environments.

Liability

This Issue | Views 35,788 | Citations 54 | Altmetric 457 | Comments

Viewpoint

October 4, 2019

More

Potential Liability for Physicians Using Artificial Intelligence

W. Nicholson Price II, JD, PhD[1]; Sara Gerke, Dipl-Jur Univ[2]; I. Glenn Cohen, JD[3]

Author Affiliations | Article Information

JAMA. 2019;322(18):1765-1766. doi:10.1001/jama.2019.15064

Current law shields physicians from liability as long as they follow the **standard of care**, the "safest" way to use medical AI from a liability perspective is as a **confirmatory tool** to support existing decision-making processes, rather than as a source of ways to improve care.

Liability

This Issue | Views 35,788 | Citations 54 | Altmetric 457 | Comments

Viewpoint

October 4, 2019

More

Potential Liability for Physicians Using Artificial Intelligence

W. Nicholson Price II, JD, PhD[1]; Sara Gerke, Dipl-Jur Univ[2]; I. Glenn Cohen, JD[3]

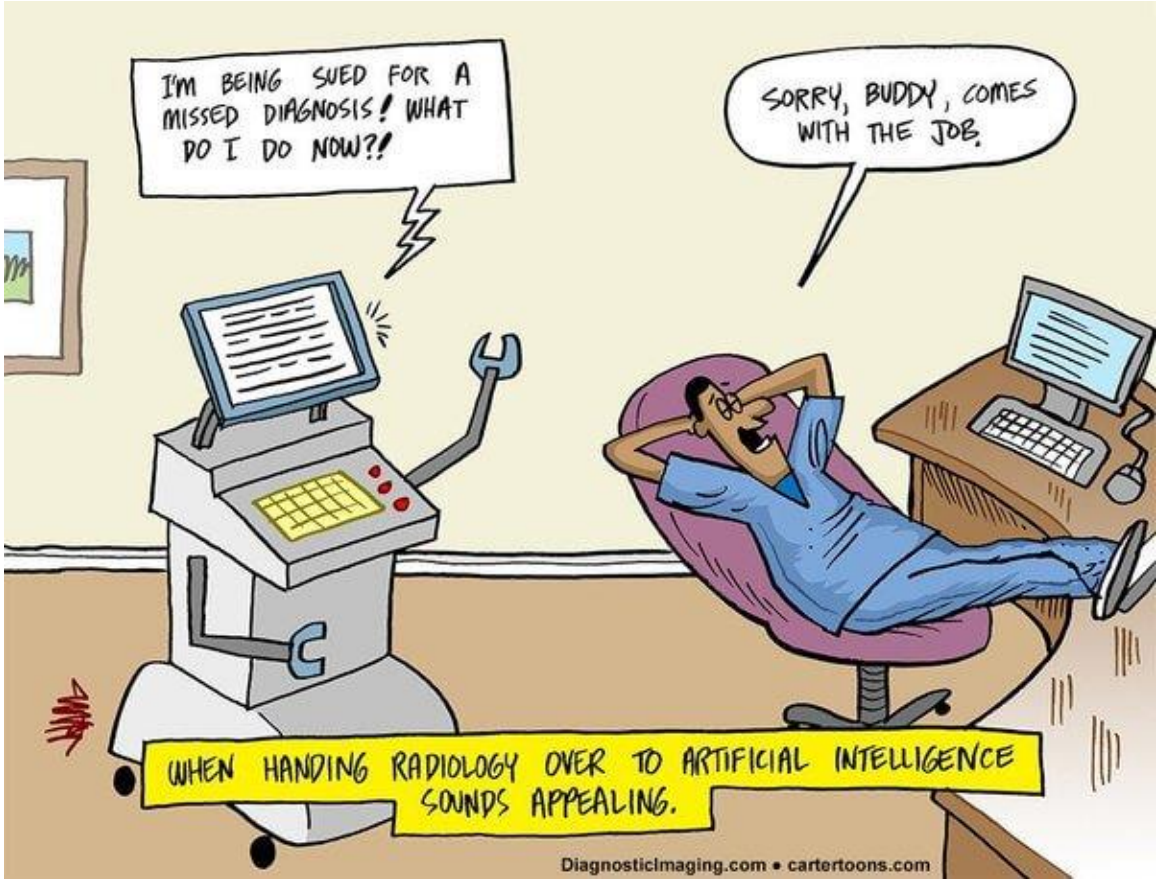
Author Affiliations | Article Information

JAMA. 2019;322(18):1765-1766. doi:10.1001/jama.2019.15064

- In the future, some medical **AIs will perform better** than even the best physicians.

- Because threat of liability encourages physicians to meet and follow the standard of care, they may reject such recommendations and thus fail to realize the full value of AI, in some cases to patients' detriment.

Scenario	AI recommendation	AI accuracy	Physician action	Patient outcome	Legal outcome (probable)
1	Standard of care	Correct	Follows	Good	No injury and no liability
2			Rejects	Bad	Injury and liability
3		Incorrect (standard of care is incorrect)	Follows	Bad	Injury but no liability
4			Rejects	Good	No injury and no liability
5	Nonstandard care	Correct (standard of care is incorrect)	Follows	Good	No injury and no liability
6			Rejects	Bad	*Injury but no liability*
7		Incorrect	Follows	Bad	*Injury and liability*
8			Rejects	Good	No injury and no liability


I'M BEING SUED FOR A MISSED DIAGNOSIS! WHAT DO I DO NOW?!
SORRY, BUDDY, COMES WITH THE JOB.
WHEN HANDING RADIOLOGY OVER TO ARTIFICIAL INTELLIGENCE SOUNDS APPEALING.
DiagnosticImaging.com • cartertoons.com

Regulating AI

Perspective | Open Access | Published: 07 April 2020

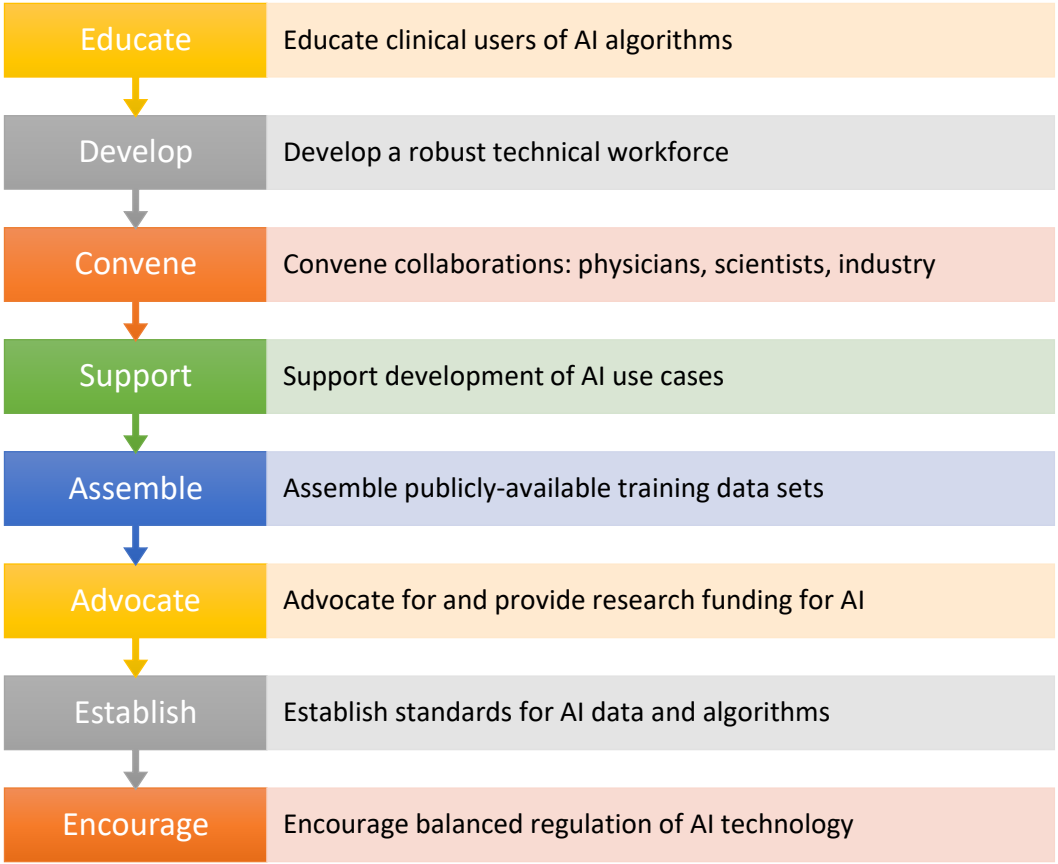
The need for a system view to regulate artificial intelligence/machine learning-based software as medical device

Sara Gerke, Boris Babic, Theodoros Evgeniou & I. Glenn Cohen

npj Digital Medicine **3**, Article number: 53 (2020) | Cite this article

- Which medical AI/ML-based products should be reviewed by regulators?
- What **evidence** should be required to permit marketing for AI/ML-based software as a medical device (SaMD)?
- How can we ensure the safety and effectiveness of AI/ML-based SaMD that may change over time as they are applied to new data?
- Regulators like the FDA need to **widen their scope** from evaluating medical AI/ML-based products to assessing systems.

Responsibilities of Medical Societies and Organizations



Responsibilities of Medical Societies and Organizations

Encourage	Encourage professional organizations to take active steps to evaluate **practice-specific algorithms**
Review	Review by the FDA will provide some quality assurance, but societies will be well placed to provide additional guidelines to evaluate AI products at implementation and to evaluate AI recommendations for individual patients.
Guide	As societies guide the standard of care for specific interventions through practice guidelines, they can guide practices for adopting and using medical AI reliably, safely, and effectively.

Physicians' Role

Deeply involved in the development, validation, and implementation of big data analytics, artificial intelligence, and personalized medicine in medicine.

Learn how to better use and interpret AI algorithms, including in what situations medical AI should be applied and how much confidence should be placed in an algorithmic recommendation.

Fiduciary responsibility for the well-being of their patients, as affirmed in the Hippocratic oath, rendering them professionally responsible for securing the interest of their patients.

Physicians' Role

Ensure that administrative efforts to develop and deploy algorithms reflect what is truly needed in clinical care.

When external AI products are procured, physicians should advocate for safeguards to ensure that such products are rigorously vetted before procurement, just as with other novel medical devices.

Malpractice Coverage and Org Liability

Check carefully with their malpractice insurer to determine how the insurer covers the use of medical AI in practice.

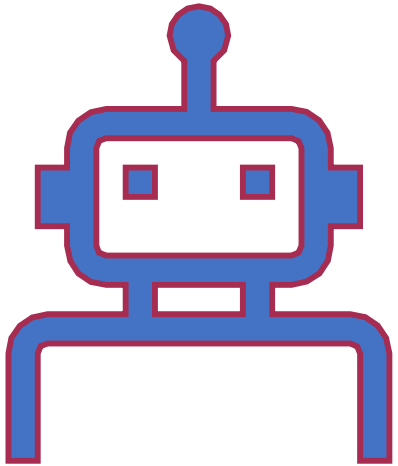
- Is care that relies on AI recommendations covered the same as care without such recommendations, or does the insurer treat such practices differently?
- Are practices different for more opaque algorithms that provide little or no reasoning?

Problem becomes far more complex with the recognition that physician liability is only one piece of a larger ecosystem of liability.

- Hospital systems that purchase and implement medical AI, makers of medical AI, and potentially even payers could all face liability.

Parting Advice and Summary

- AI is a **powerful** tool with many applications that can help physicians in many diagnostic tasks.
- Integrating AI models holds promise for **improving** healthcare delivery and patient outcomes.
- More research needs to be done regarding the **evaluation** of AI in a clinical setting, including its **impact** on workflow and value of services.
- No matter how AI is implemented in the workflow, the physicians will have an important role in ensuring accuracy, safety and quality of the algorithms.



Thank you