Research Ethics for AI in Health Applications

BAC's 20th Anniversary Virtual Public Conference, 17th June 2021

Pavitra Krishnaswamy
Senior Scientist, Machine Intellection Department
Deputy Head, Healthcare & MedTech Division

A*STAR Institute for Infocomm Research

I²R Research Capabilities

Research Dept.

- Visual Intelligence
- CyberSecurity
- Communication & Networks
- Machine Intellection
- Robotics and Autonomous Systems
- Aural & Language Intelligence
- Signal Processing, RF & Optical

Research Division

- Safety & Security
- Satellite, Aviation & Maritime
- Intelligent Transportation Solutions
- Healthcare & Medtech
- Advanced Manufacturing & Engineering
- Sustainable Built Environment

Outline

Opportunities: AI in Health Applications

Data Access and Use: Challenges & Case Studies

Translation and Deployment: Challenges & Case Studies

Summary and Discussion

OPPORTUNITIES
AI in Health Applications

Access to Care Difficult

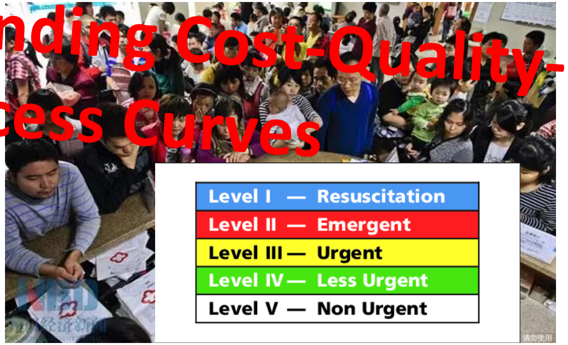
Reactive to Disease

Acute Care Heavy

Made for the Average Person

Access to Care Difficult

Bending Cost-Quality-Access Curves



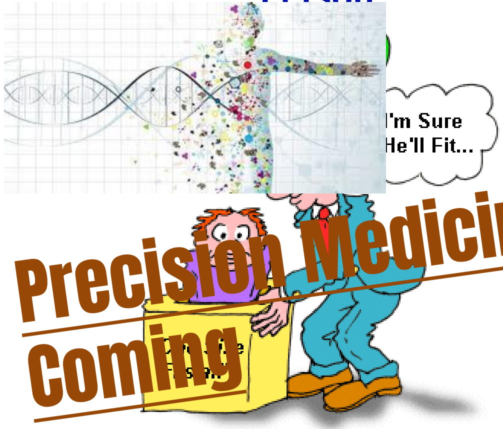
Reactive to Disease

From Reactive to Proactive

Acute Care Heavy

Beyond Hospital to Home

Made for the Average Person



Precision Medicine is Coming

Two Types of Challenges

Access and Use of Health Data for R&D

R&D in AI for health applications relies heavily on large health or clinical datasets. However, the sensitive and private nature of these datasets introduces several ethical and legal challenges for data access, use and governance.

Evaluation of AI Solutions for Deployment

Translation of AI technology for clinical and public health applications requires systematic evaluation and quality control. However, this requires increasing robustness and transparency of AI models for end user buy-in and regulatory oversight.

DATA ACCESS AND USE
Challenges and Solutions

R&D Requires Data from Large Numbers

Article | Open Access | Published: 08 May 2018

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar, Eyal Oren, [...]Jeffrey Dean

npj Digital Medicine **1**, Article number: 18 (2018) | Cite this article

Abstract

Predictive modeling with electronic health record (EHR) data is anticipated to drive personalized medicine and improve healthcare quality. Constructing predictive statistical models typically requires extraction of curated predictor variables from normalized EHR data, a labor-intensive process that discards the vast majority of information in each patient's record. We propose a representation of patients' entire raw EHR records based on the Fast Healthcare Interoperability Resources (FHIR) format. We demonstrate that deep learning methods using this representation are capable of accurately predicting multiple medical events from multiple centers without site-specific data harmonization. We validated our approach using de-identified EHR data from two US academic medical centers with 216,221 adult patients hospitalized for at least 24 h. In the sequential format we propose, this volume of EHR data unrolled into a total of 46,864,534,945 data points, including clinical notes. Deep learning models achieved high accuracy for tasks such as predicting: in-hospital mortality (area under the receiver operator curve [AUROC] across sites 0.93–0.94), 30-day unplanned readmission (AUROC 0.75–0.76), prolonged length of stay (AUROC 0.85–0.86), and all of a patient's final discharge diagnoses (frequency-weighted AUROC 0.90). These models outperformed traditional, clinically-used predictive models in all cases. We believe that this approach can be used to create accurate and scalable predictions for a variety of clinical scenarios. In a case study of a particular prediction, we demonstrate that neural networks can be used to identify relevant information from the patient's chart.

Letter | Published: 17 August 2020

Wearable-device-measured physical activity and future health risk

Tessa Strain, Katrien Wijndaele, Paddy C. Dempsey, Stephen J. Sharp, Matthew Pearce, Justin Jeon, Tim Lindsay, Nick Wareham & Søren Brage

Nature Medicine **26**, 1385–1391 (2020) | Cite this article

Abstract

Use of wearable devices that monitor physical activity is projected to increase more than fivefold per half-decade[1]. We investigated how device-based physical activity energy expenditure (PAEE) and different intensity profiles were associated with all-cause mortality. We used a network harmonization approach to map dominant-wrist acceleration to PAEE in 96,476 UK Biobank participants (mean age 62 years, 56% female). We also calculated the fraction of PAEE accumulated from moderate-to-vigorous-intensity physical activity (MVPA). Over the median 3.1-year follow-up period (302,526 person-years), 732 deaths were recorded. Higher PAEE was associated with a lower hazard of all-cause mortality for a constant fraction of MVPA (for example, 21% (95% confidence interval 4–35%) lower hazard for 20 versus 15 $kJ\,kg^{-1}\,d^{-1}$ PAEE with 10% from MVPA). Similarly, a higher MVPA fraction was associated with a lower hazard when PAEE remained constant (for example, 30% (8–47%) lower hazard when 20% versus 10% of a fixed 15 $kJ\,kg^{-1}\,d^{-1}$ PAEE volume was from MVPA). Our results show that higher volumes of PAEE are associated with reduced mortality rates, and achieving the same volume through higher-intensity activity is associated with greater reductions than through lower-intensity activity. The linkage of device-measured activity to energy expenditure creates a framework for using wearables for personalized prevention.

Retrospective Data of Large Numbers -> Lack of Consent

Article | Open Access | Published: 08 May 2018

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar, Eyal Oren, [...]Jeffrey Dean

npj Digital Medicine **1**, Article number: 18 (2018) | Cite this article

Datasets

We included EHR data from the University of California, San Francisco (UCSF) from 2012 to 2016, and the University of Chicago Medicine (UCM) from 2009 to 2016. We refer to each health system as Hospital A and Hospital B. All EHRs were de-identified, except that dates of service were maintained in the UCM dataset. Both datasets contained patient demographics, provider orders, diagnoses, procedures, medications, laboratory values, vital signs, and flowsheet data, which represent all other structured data elements (e.g., nursing flowsheets), from all inpatient and outpatient encounters. The UCM dataset additionally contained de-identified, free-text medical notes. Each dataset was kept in an encrypted, access-controlled, and audited sandbox.

Ethics review and institutional review boards approved the study with waiver of informed consent or exemption at each institution.

Letter | Published: 17 August 2020

Wearable-device-measured physical activity and future health risk

Tessa Strain, Katrien Wijndaele, Paddy C. Dempsey, Stephen J. Sharp, Matthew Pearce, Justin Jeon, Tim Lindsay, Nick Wareham & Søren Brage

Nature Medicine **26**, 1385–1391 (2020) | Cite this article

Abstract

Use of wearable devices that monitor physical activity is projected to increase more than fivefold per half-decade[1]. We investigated how device-based physical activity energy expenditure (PAEE) and different intensity profiles were associated with all-cause mortality. We used a network harmonization approach to map dominant-wrist acceleration to PAEE in 96,476 UK Biobank participants (mean age 62 years, 56% female). We also calculated the fraction of PAEE accumulated from moderate-to-vigorous-intensity physical activity (MVPA). Over the median 3.1-year follow-up period (302,526 person-years), 732 deaths were recorded. Higher PAEE was associated with a lower hazard of all-cause mortality for a constant fraction of MVPA (for example, 21% (95% confidence interval 4–35%) lower hazard for 20 versus 15 kJ $kg^{-1} d^{-1}$ PAEE with 10% from MVPA). Similarly, a higher MVPA fraction was associated with a lower hazard when PAEE remained constant (for example, 30% (8–47%) lower hazard when 20% versus 10% of a fixed 15 kJ $kg^{-1} d^{-1}$ PAEE volume was from MVPA). Our results show that higher volumes of PAEE are associated with reduced mortality rates, and achieving the same volume through higher-intensity activity is associated with greater reductions than through lower-intensity activity. The linkage of device-measured activity to energy expenditure creates a framework for using wearables for personalized prevention.

More than Just Numbers…

Article | Open Access | Published: 10 October 2018

Genome-wide association studies of brain imaging phenotypes in UK Biobank

Lloyd T. Elliott, Kevin Sharp, Fidel Alfaro-Almagro, Sinan Shi, Karla L. Miller, Gwenaëlle Douaud, Jonathan Marchini & Stephen M. Smith

Nature **562**, 210–216 (2018) | Cite this article

Abstract

The genetic architecture of brain structure and function is largely unknown. To investigate this, we carried out genome-wide association studies of 3,144 functional and structural brain imaging phenotypes from UK Biobank (discovery dataset 8,428 subjects). Here we show that many of these phenotypes are heritable. We identify 148 clusters of associations between single nucleotide polymorphisms and imaging phenotypes that replicate at $P < 0.05$, when we would expect 21 to replicate by chance. Notable significant, interpretable associations include: iron transport and storage genes, related to magnetic susceptibility of subcortical brain tissue; extracellular matrix and epidermal growth factor genes, associated with white matter micro-structure and lesions; genes that regulate mid-line axon development, associated with organization of the pontine crossing tract; and overall 17 genes involved in development, pathway signalling and plasticity. Our results provide insights into the genetic architecture of the brain that are relevant to neurological and psychiatric disorders, brain development and ageing.

Article | Published: 01 January 2020

International evaluation of an AI system for breast cancer screening

Scott Mayer McKinney, Marcin Sieniek, […]Shravya Shetty

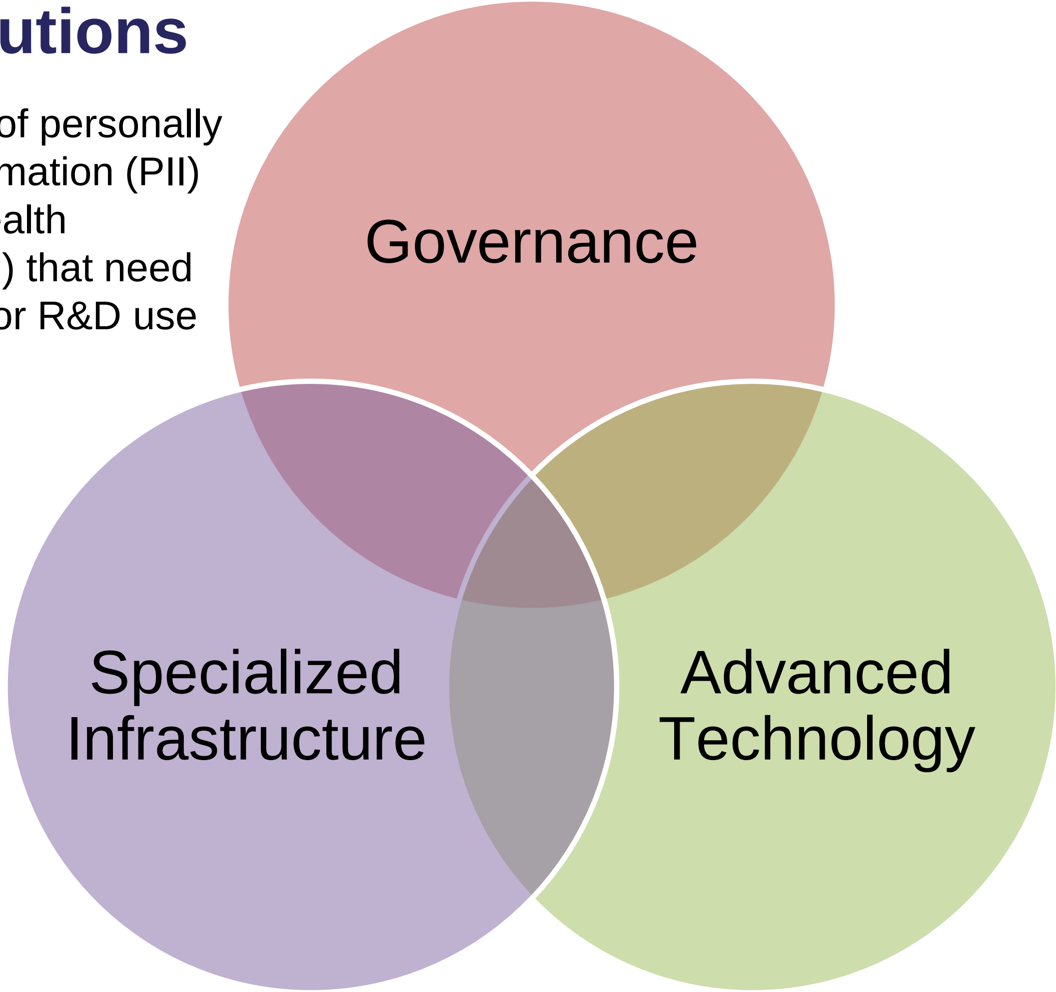
Nature **577**, 89–94 (2020) | Cite this article

Abstract

Screening mammography aims to identify breast cancer at earlier stages of the disease, when treatment can be more successful[1]. Despite the existence of screening programmes worldwide, the interpretation of mammograms is affected by high rates of false positives and false negatives[2]. Here we present an artificial intelligence (AI) system that is capable of surpassing human experts in breast cancer prediction. To assess its performance in the clinical setting, we curated a large representative dataset from the UK and a large enriched dataset from the USA. We show an absolute reduction of 5.7% and 1.2% (USA and UK) in false positives and 9.4% and 2.7% in false negatives. We provide evidence of the ability of the system to generalize from the UK to the USA. In an independent study of six radiologists, the AI system outperformed all of the human readers: the area under the receiver operating characteristic curve (AUC-ROC) for the AI system was greater than the AUC-ROC for the average radiologist by an absolute margin of 11.5%. We ran a simulation in which the AI system participated in the double-reading process that is used in the UK, and found that the AI system maintained non-inferior performance and reduced the workload of the second reader by 88%. This robust assessment of the AI system paves the way for clinical trials to improve the accuracy and efficiency of breast cancer screening.

Plausible Solutions

Enlist the types of personally identifiable information (PII) and sensitive health information (SHI) that need to be removed for R&D use



Capabilities to learn without exposing PII or SHI (even if present) beyond native institution where they reside already

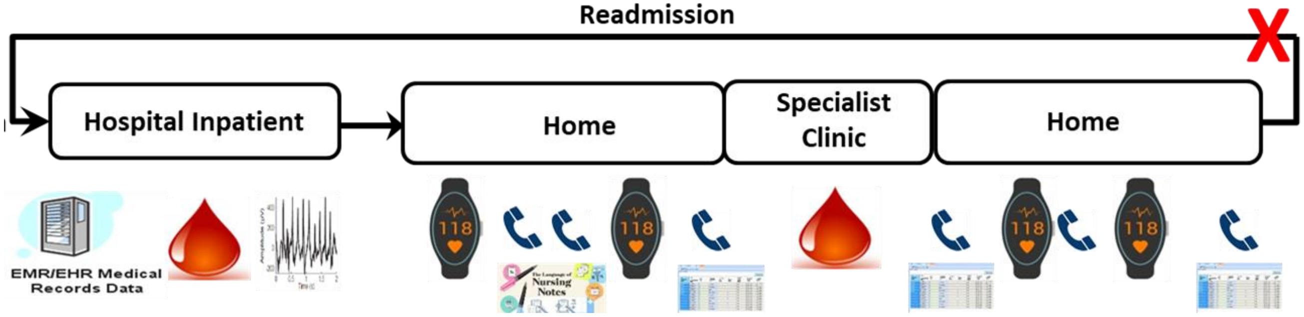
Infrastructure to remove PII, SHI, and all links to metadata that might enable re-identification

Plausible Solutions

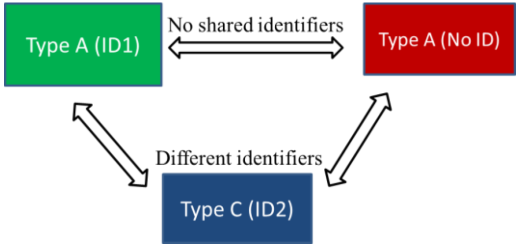
	Application	Specifics	Data Type(s)	Type of Solution
1	Allied Health	Implicit Identifiers General Consent for R&D	Electronic Health Record Tele-monitoring & Lifestyle Data Nurse-Patient Conversations	Specialized Infrastructure: Anonymization & Crosslinking
2	Precision Medicine	Anonymized Consent Maybe Present/Waived	Electronic Health Record Consumer Health Data Genomics	Specialized Infrastructure: Remote Access
3	Screening	De-identified Consent Waived	Medical Imaging Scans	Advanced Technology: Federated Learning

Specialized Infrastructure

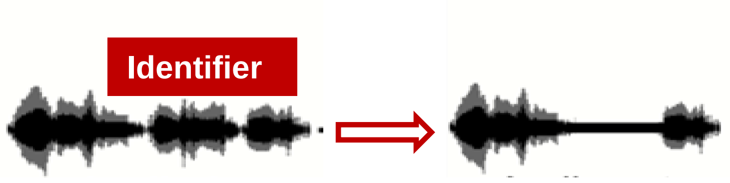
1. Risk Stratification in the Community: Anonymize and Crosslink Multimodal Data



Multiple data modalities from multiple fragmented systems	Data not readily anonymized	Complexity of specialist unstructured data type



Crosslinking Fragmented Data

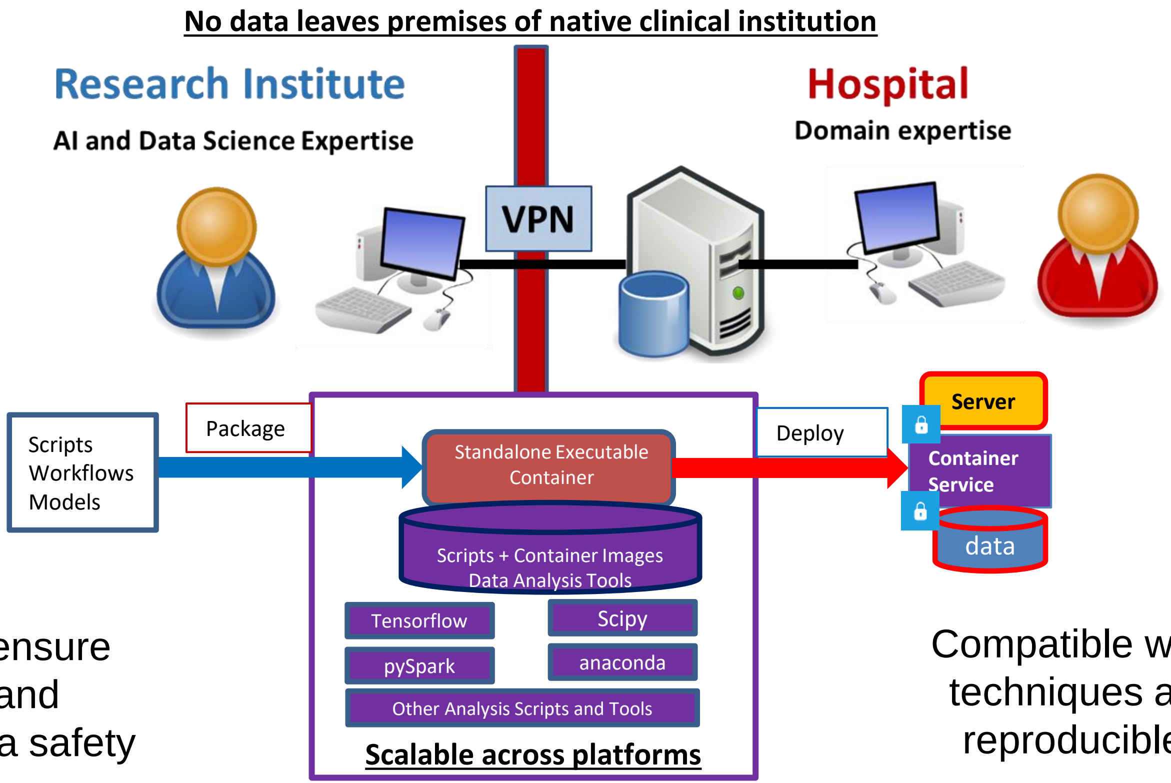


Infrastructure to remove PII and SHI from specialized data types

Role and work site segregation between data preparation and analysis teams

Specialized Infrastructure

2. Possibilities for Precision Medicine: Move Analysis to Data



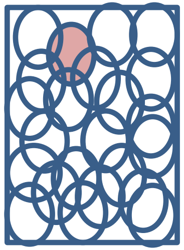
Containers ensure isolation and maximize data safety

Compatible with latest AI techniques and enable reproducible analysis

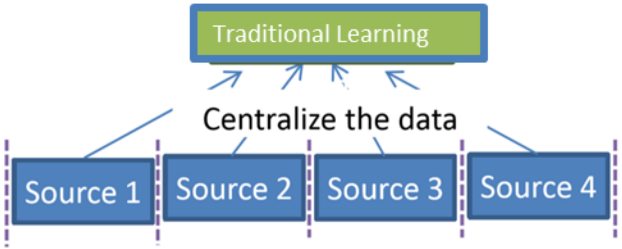
Advanced Technology

3. Screening: Learn in a Federated Manner

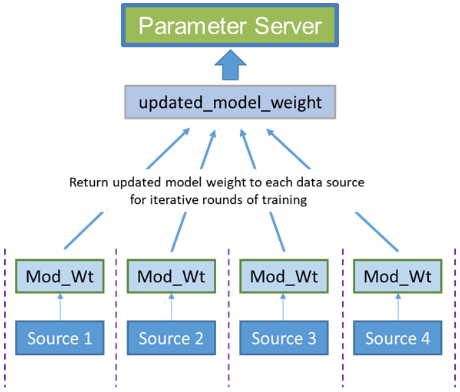
Screening conditions with low prevalence: paucity of positive samples



Centralization of Patient Data Beyond Provider Site Infeasible



Federated Learning



Move Models Not Data
Protect Biometric Information that Can Remain After Anonymization
Delineate Ownership of Data vs Models
Open Question: How to Control Quality?

TRANSLATION AND DEPLOYMENT: Challenges & Solutions

Translation: Potential vs. Reality

ORIGINAL ARTICLE | ARTICLES IN PRESS

2020 ACR Data Science Institute Artificial Intelligence Survey

Bibb Allen • Sheela Agarwal • Laura Coombs • Keith Dreyer • Christoph Wald

Published: April 20, 2021 • DOI: https://doi.org/10.1016/j.jacr.2021.04.002

Key Words

Artificial intelligence in clinical practice • artificial intelligence survey • barriers to implementation of artificial intelligence • market penetrance of artificial intelligence

Purpose

The ACR Data Science Institute conducted its first annual survey of ACR members to understand how radiologists are using artificial intelligence (AI) in clinical practice and to provide a baseline for monitoring trends in AI use over time.

Methods

The ACR Data Science Institute sent a brief electronic survey to all ACR members via e-mail. Invitees were asked for demographic information about their practice and if and how they were currently using AI as part of their clinical work. They were also asked to evaluate the performance of AI models in their practices and to assess future needs.

Results

Approximately 30% of radiologists are currently using AI as part of their practice. Large practices were more likely to use AI than smaller ones, and of those using AI in clinical practice, most were using AI to enhance interpretation, most commonly detection of intracranial hemorrhage, pulmonary emboli, and mammographic abnormalities. Of practices not currently using AI, 20% plan to purchase AI tools in the next 1 to 5 years.

Conclusion

The survey results indicate a modest penetrance of AI in clinical practice. Information from the survey will help researchers and industry develop AI tools that will enhance radiological practice and improve quality and efficiency in patient care.

FDA CLEARED AI PRODUCTS

56 FDA CLEARED AI ALGORITHMS AS OF JULY 2020

ACR DSI AI SURVEY

CURRENT AND POTENTIAL MARKET PENETRANCE

Total AI Users	Total AI Algorithms	AI per Rad (Surveyed)	Algorithms in Use (Projected)	AI per Rad Maximum (Est)	Total AI Market Opportunity	Market Penetration	Average Rads per Group (Est)	Total AI Sales Opportunities
315	389	1.2	10,881	20	560,000	2%	10	56,000

Most Popular AI Models	Total Surveyed Use	Percent of Total AI Used	Estimated Total Market Use	Estimated Total Sales To Date
Self-Developed AI	38	9.8%	1,063	N/A
Mammography (Screening)	35	9.0%	979	98
CT Chest (Embolism)	25	6.4%	699	70
MR Brain (Analytics)	23	5.9%	643	64
CT Brain (Hemorrhage)	22	5.7%	615	62

Despite availability of regulatory cleared products that match clinical needs, overall penetration of AI in clinical practice is very low

Why Low Penetrance into Clinical Use?

Clinical Impact

- See no benefit
- Concerned that it will decrease their productivity

Performance

- Low trust: inconsistent AI performance, algorithm biases, arbitrary variation
- Lack of evaluation in the clinical workflow and rigorous performance standards (local to workflow and institution)

Business

- Cannot justify the expense/lack of reimbursement models
- System integration challenges
- Decision to implement requires many stakeholders

Emerging Standards for Validation of AI Models

SPIRIT-AI

Reporting Guidelines for Clinical Trial Protocols for Interventions Involving Artificial Intelligence

The SPIRIT-AI Extension

CONSORT-AI

Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence

The CONSORT-AI Extension

The SPIRIT-AI and CONSORT-AI initiative is an international collaborative effort to improve the transparency and completeness of reporting of clinical trials evaluating interventions involving artificial intelligence (AI). SPIRIT-AI stands for Standard Protocol Items: Recommendations for Interventional Trials - Artificial Intelligence and CONSORT-AI stands for (Consolidated Standards of Reporting Trials - Artificial Intelligence).

The SPIRIT-AI and CONSORT-AI statements are extensions to the SPIRIT 2013 and CONSORT 2010 reporting guidelines for clinical trial protocols and clinical trial reports, respectively. The two extensions build upon existing recommendations to address considerations specific to AI health interventions.

They are intended to help promote transparency and completeness for clinical trial protocols for AI interventions and assist editors and peer-reviewers, as well as the general readership, to understand, interpret and critically appraise clinical trial protocols and reports.

The SPIRIT-AI and CONSORT-AI initiative was announced in *Nature Medicine* in October 2019. The two statements were developed simultaneously through international multi-stakeholder consensus and published in *Nature Medicine*, the *British Medical Journal* and the *Lancet Digital Health* in September 2020. The guidelines are supported by, and developed according to guidelines set out by, the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) network. The use of the guidelines are endorsed by key medical journals for the reporting of clinical trials for AI interventions.

- **Documentation of inputs and outputs**
- **How the algorithm makes recommendations and fits into a clinical pathway**
- **Error Analysis: Sources of bias and generalisability**
- **Statements on algorithm ownership and access**

https://www.clinical-trials.ai/

So Where is the Gap?

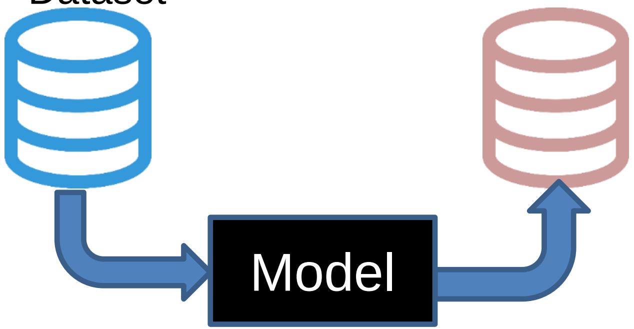
Controlled Clinical Trial

vs

Data, Workflow and System Challenges in Ongoing Practice

Development Dataset

Inference Dataset

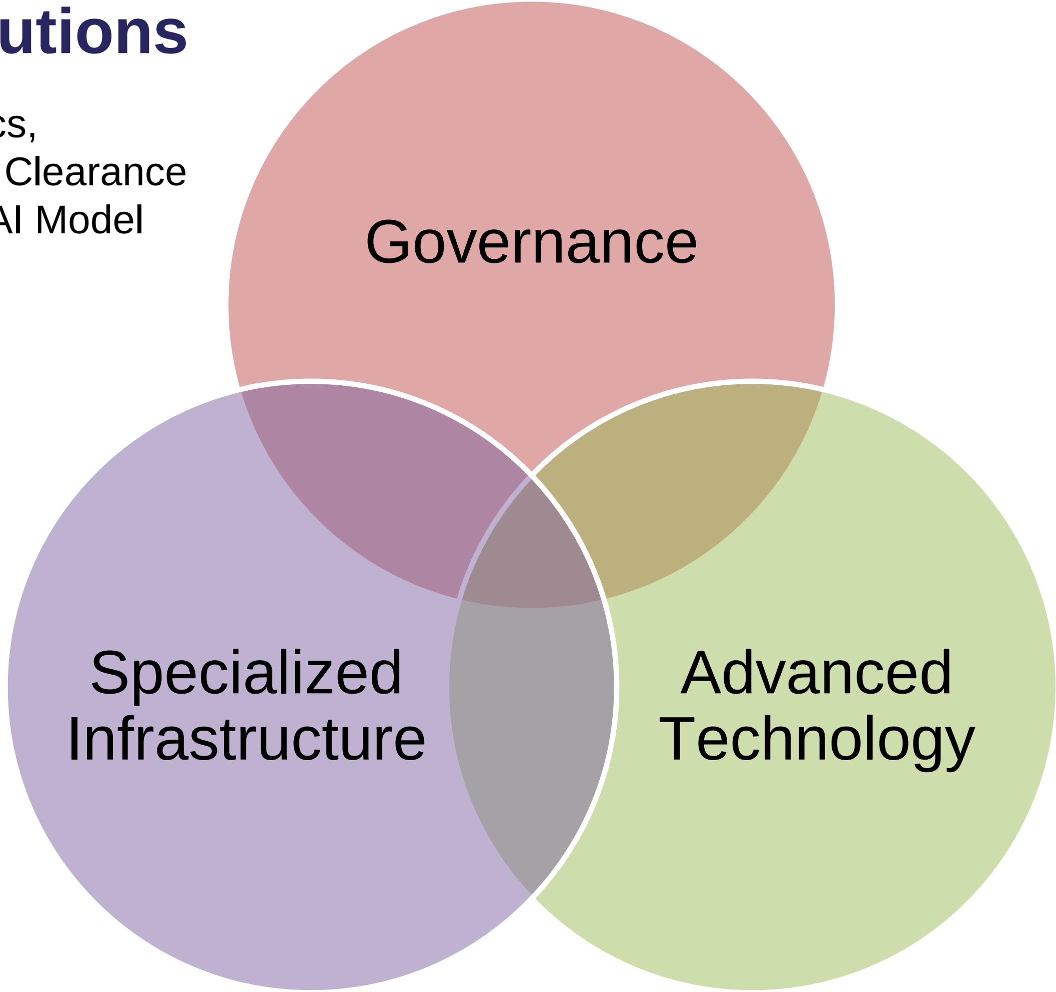


Mechanisms to Detect and Adapt to Dataset Drift, Non-Stationarity, Domain Shift

- Observational Data Quality Limited – Uncertainty Measures and Audit Tools Needed
- Running Models on Patient Data Poses Exposure Risks – Privacy Preserving Inference Tools Needed

Plausible Solutions

Standards, Ethics, Regulation, and Clearance Procedures for AI Model Deployment



Capabilities to enhance trust: quantify uncertainty, preserve privacy, and audit AI solutions in routine practice

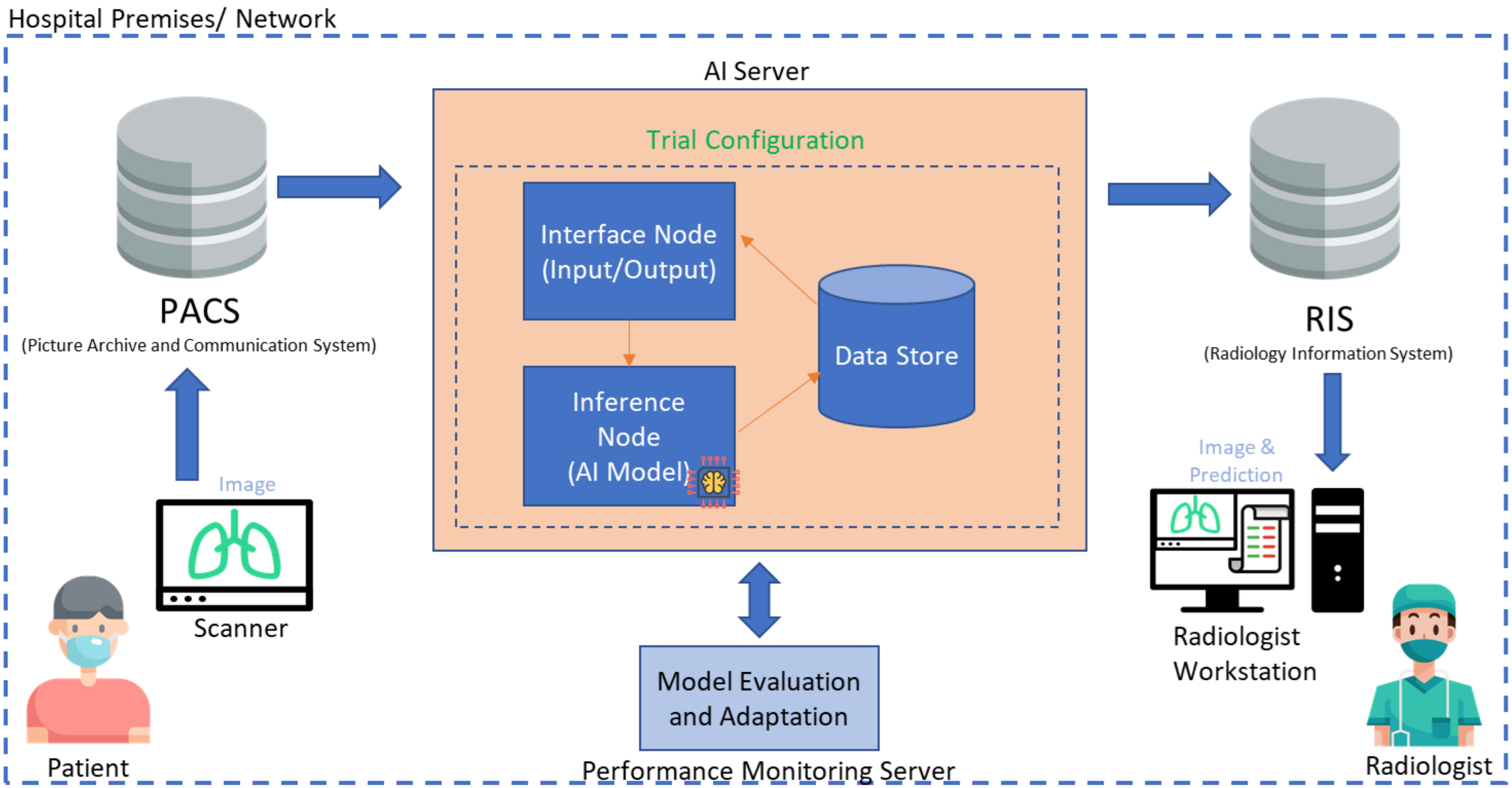
Infrastructure to monitor performance, assess and adapt to drifts/domain shift

Plausible Solutions

	Application	Specifics	Data Type(s)	Type of Solution
1	Clinical Decision Support	De-identified Data for Workflow Solutions	Medical Imaging Scans and Reports	Specialized Infrastructure for Evaluation
2	Allied Health	De-identified Data for Workflow Solutions	Electronic Health Record Tele-monitoring & Lifestyle Data	Advanced Technology: Uncertainty Quantification
3	Screening	Anonymized but Contains Biometric Information	Medical Imaging Scans	Advanced Technology: Encrypted Deep Learning Inference

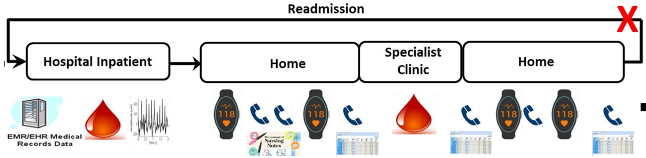
Specialized Infrastructure

1. Clinical Decision Support: Continuous Evaluation in Practice

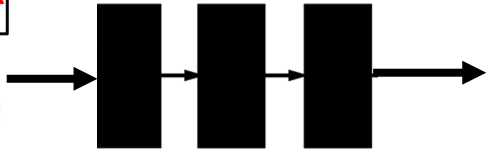


Advanced Technology

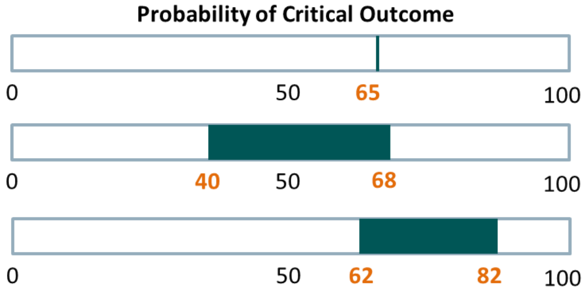
2. Risk Stratification in the Community: Uncertainty Quantification



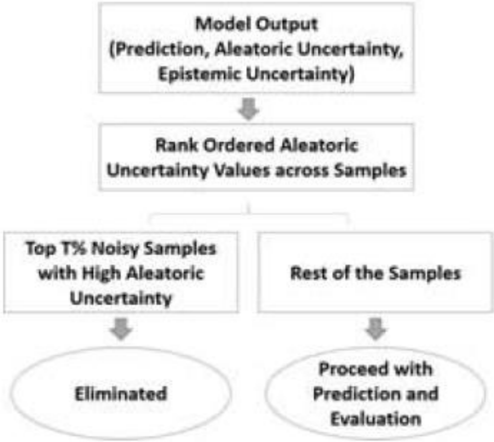
High Risk



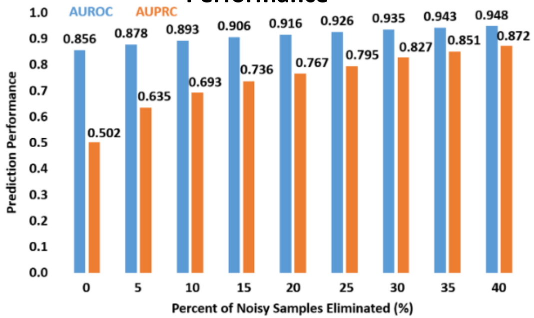
Black box



The error bars matter!



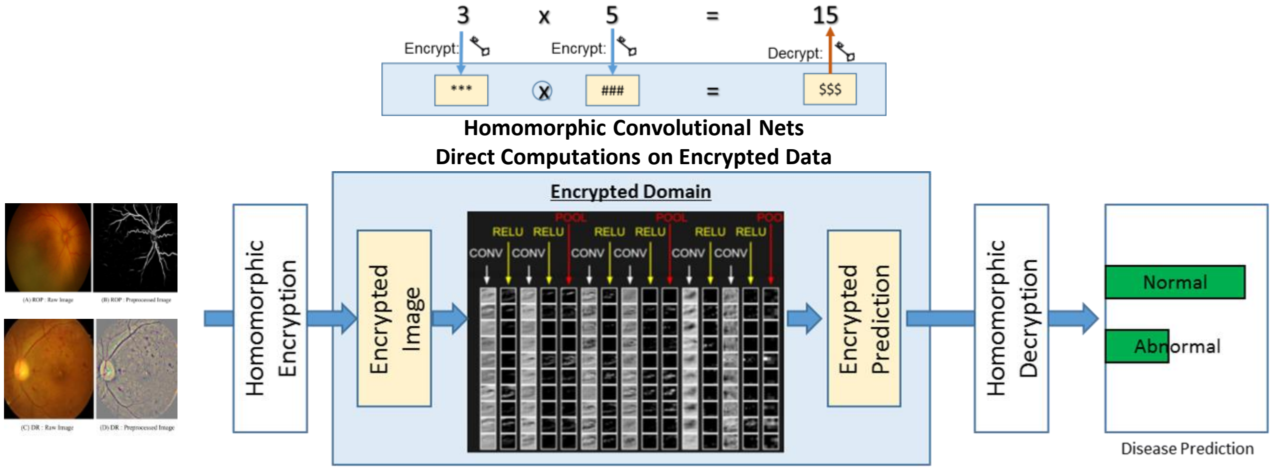
Data and Model Quality Characterization for Improved Predictive Performance



Bayesian Learning for Characterization of Data and Model Uncertainty (AAAI 2020)

Advanced Technology

3. Screening: Deep Learning Meets Homomorphic Encryption



The Holy Grail: Privacy preserving Inference on Encrypted Biomedical Data

Challenges: Slow, High Computational Resource Demands, Reduces Accuracy

Biomedical Applications: Large anonymization load, scarce computational resources at point of care, high data dimensionality

CaRENets – Compact and Resource Efficient Homomorphic CNN (Privacy in ML, NeurIPS 2019)

Approach	HE Encryption Parameters			Application 1: 96 x 96 grey scale images			Application 2: 256 x 256 RGB images		
[redacted]									
CaRENets (CPU)	14	660	>80 bit	2.94	994.9	2.90	17.2	6004	61.88
[redacted]									

Maintains accuracy within 3%, Improves latency by > 32X, memory usage by > 45X, message size by > 5000x

SUMMARY

Acknowledgments

Collaborators

Khin Mi Mi Aung
Lux Anantharaman
Nancy Chen
Savitha Ramasamy
Vijay Chandrasekhar
Bee Yong Mong
Chow Wai Leng
Jayashree Kalpathy-Cramer
Lim Weng Khong
Michael Chiang
Oh Hong Choon
Oliver Nickalls
Patrick Tan
Sonia Maria Davila Dominguez
Steven Wong

Team

Ahmad Al Badawi
Balagopal Unnikrishnan
Benedict Hing
Fatemeh Fahimi
Feri Guretno
Guo Yang
Helen Erdt
Ivan Ho Mien
Jin Chao
Liu Zhengyuan
Lim Jia Hui Hazel
Milashini Nambiar
Nur Farah Binte Suhaimi
Sabriel Koh
Siti Umairah Md Salleh

References: Works Presented

- A Vital Signs Tele-monitoring Program Improves the Dynamic Prediction of Readmission Risk in Patients with Heart Failure, *AMIA 2020* **[Informatics Outside the Clinics: New Approaches for Chronic Disease Management]**
- Fast Prototyping a Dialogue Comprehension System for Nurse-Patient Conversations on Symptom Monitoring, *NAACL 2019* **[Oral presentation; Feasibility for automated comprehension of symptoms in multi-turn spoken patient-nurse conversations]**
- Evaluation of Federated Deep Learning Approaches on Multi-centre Clinical Images, *ICBME, 2019* **[Realistic multi-centre evaluation for federated learning]**
- Uncertainty Characterization for Predictive Analytics with Clinical Time Series Data, *Health Intelligence, AAAI 2019* **[Methods to quantify data quality and inform prediction with clinical time series]**
- Uncertainty Modeling for Machine Comprehension Systems using Efficient Bayesian Neural Networks, *Industry Track, COLING 2020* [**Bayesian Methods for Interpretable and Active Learning on Dialogue Data**]
- Unsupervised Deep Learning for Automated Fundus Image Quality Assessment in Retinopathy of Prematurity, *ICBME, 2019* **[First-in-class report on unsupervised deep learning based medical image quality assessment]**
- CareNets: Efficient Homomorphic CNN for High Resolution Images, *Privacy in ML, Conference on Neural Information Processing Systems (NeurIPS)*, 2019 **[Most resource-efficient deep learning inference on encrypted high resolution medical images, First GPU results]**

Summary and Discussion

- ❑ Confluence of governance, infrastructure and advanced technology developments needed to address research ethics and regulatory challenges

- ❑ What kinds of set ups can institutions working in the AI for Health domain be encouraged to invest in (remote data access, confidential labs, certified personnel)?

- ❑ What kinds of environments can deployment sites be mandated to develop (on site model updates in trial environments?

- ❑ How do we start developing evidence for emerging privacy and trust technologies to make them part of the ethics and regulatory ecosystem?

Thank you

pavitrak@i2r.a-star.edu.sg

A*STAR
Institute for
Infocomm Research