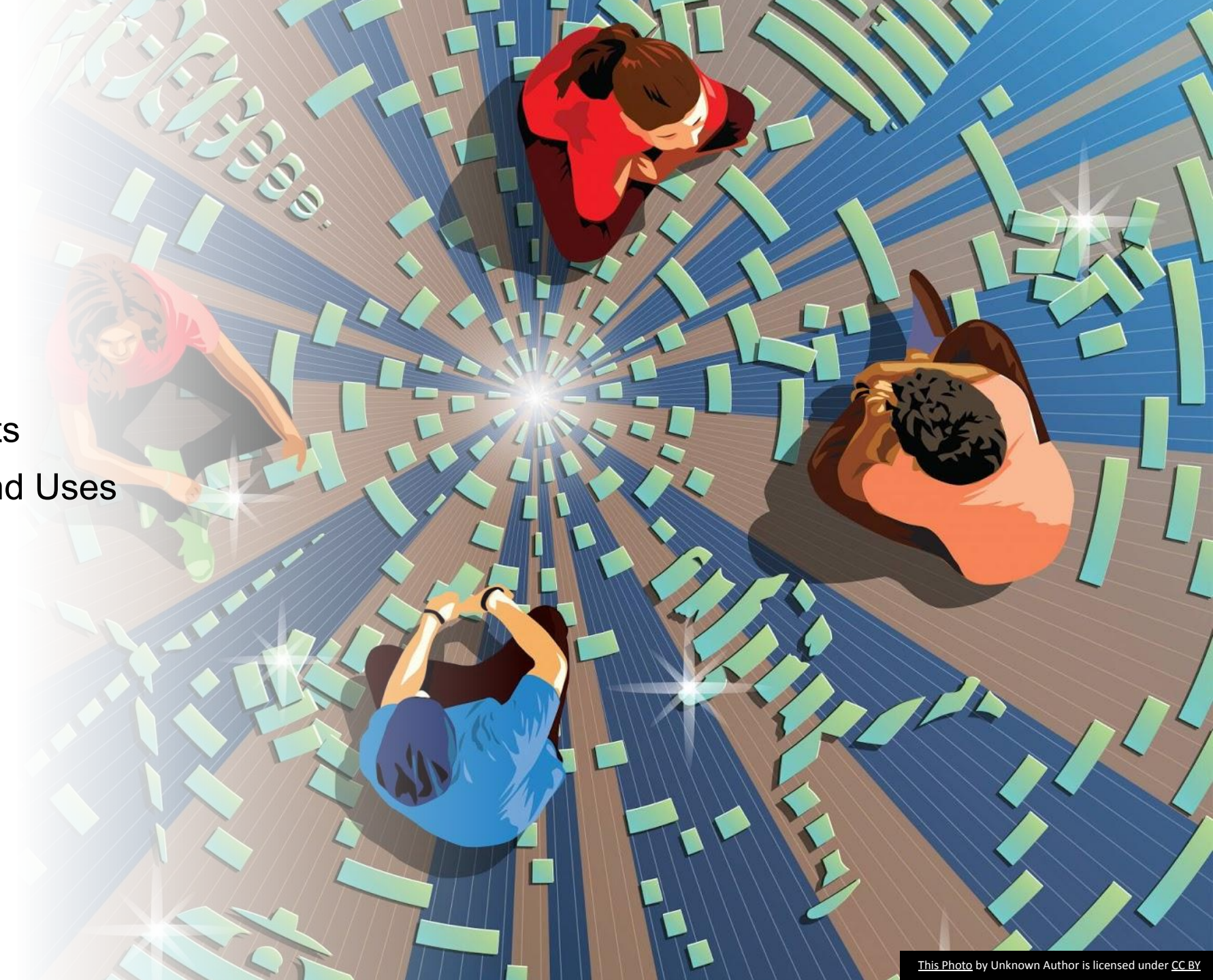
# *Comprehensive AI Evaluation Framework Applied to COVID-Related AI Studies*

Associate Professor Sandeep Reddy,
School of Medicine, Deakin University

INTELLIGENT HEALTH 2022

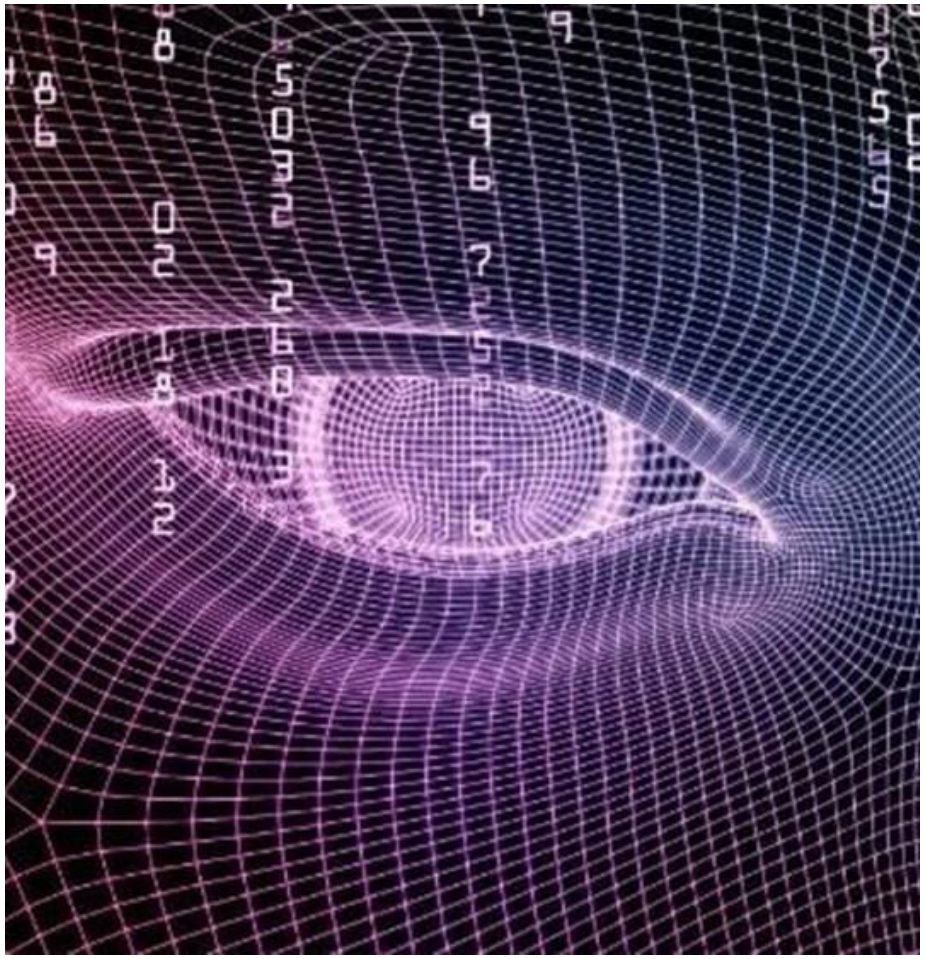
# Outline

- Context
- TEHAI and components
- Scoring Mechanism and Uses
- Application
- Review Process
- Findings
- Learning
- Discussion



# Context

- Progress in artificial intelligence (AI) has opened new opportunities
- However, in limited assessments that have taken place so far, it has been found AI systems have fallen short of their translational goals
- This is because many AI systems have intrinsic inadequacies that don't get assessed until after deployment
- Utilising and integrating AI systems in clinical settings can be potentially expensive and disruptive
- Therefore, a rigorous evaluation that assesses AI systems early and at various stages of their deployment can support or contradict the use of a specific AI tool



# Context

- Currently available evaluation frameworks generally focus on the reporting and regulatory aspects
- It is evident there is an absence of an evaluation framework that assesses various stages of development, deployment, integration and adoption of AI systems
- Dependence on disparate evaluation frameworks to assess different aspects and phases of AI systems is unrealistic
- Also, currently available evaluation and reporting frameworks fall short in adequately assessing the functional, utility, and ethical aspects of the models

# Translational Evaluation of Healthcare AI (TEHAI)

- To address the gap in currently available evaluation frameworks, an international team of medical researchers and data scientists was constituted to develop TEHAI
- Following constitution of the team, we considered the evaluation and research principles that would inform the development of the framework
- Based on these principles and a critical review of related literature including frameworks and guidelines, the project team identified the key components developed the initial version of the TEHAI over a period of six months.

# Translational Evaluation of Healthcare AI (TEHAI)

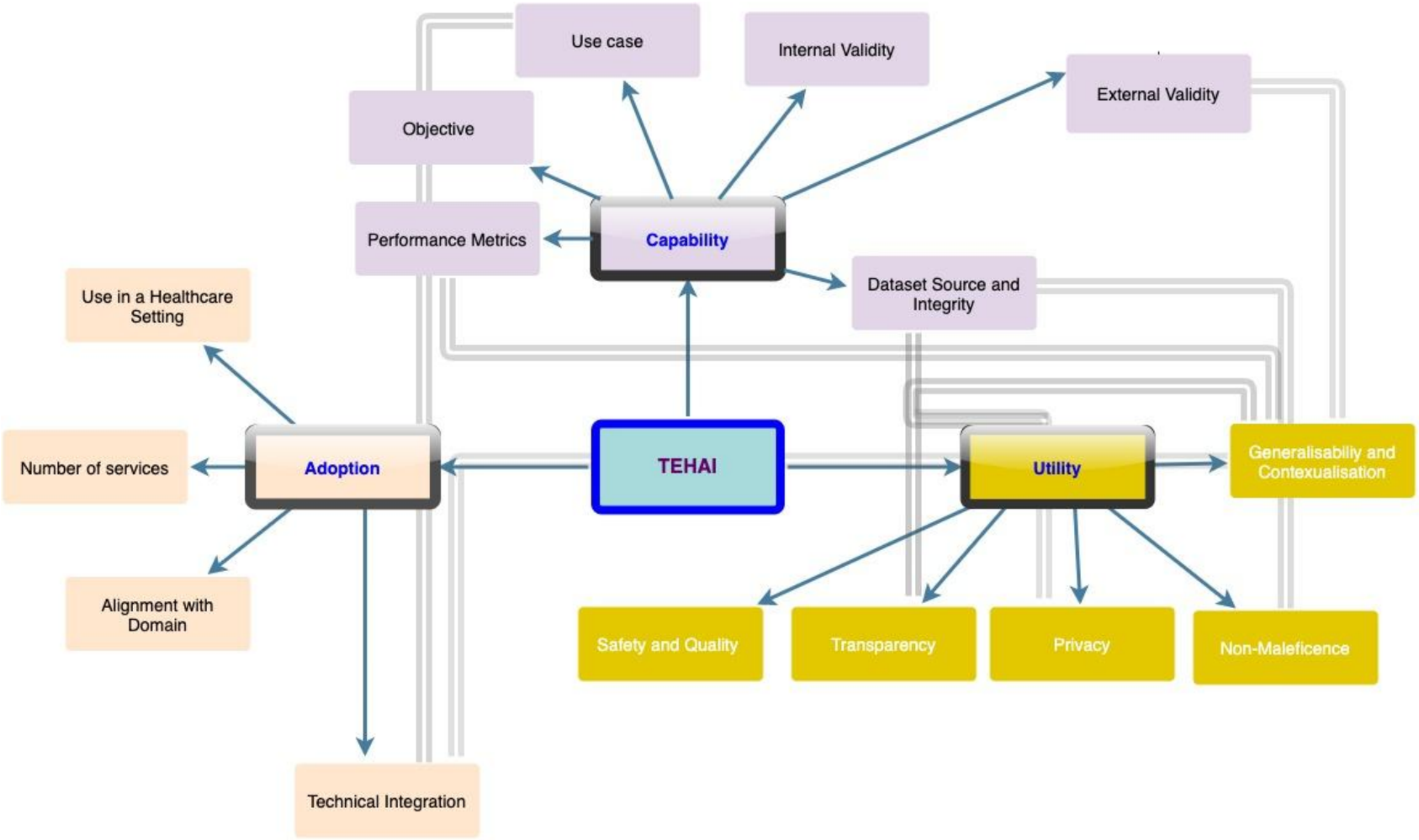
- To provide a layer of independent review before finalization of TEHAI, the draft consensus framework was then reviewed by an international panel

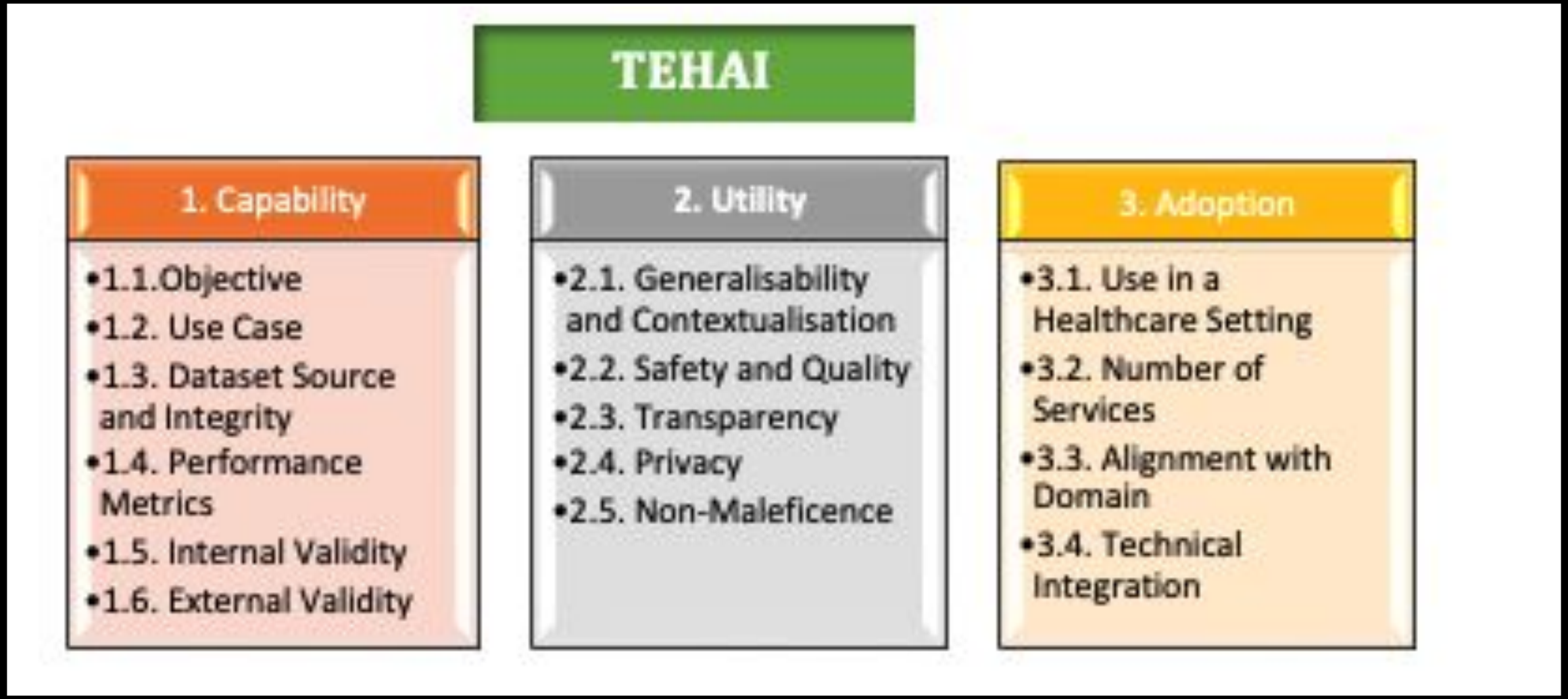
- The eight-member international panel had expertise in medicine, data science, healthcare policy, biomedical research and healthcare commissioning, and were drawn from the United Kingdom, United States of America and New Zealand

- The panel members were provided the framework and documentation and after, meetings were convened with panel members to receive their feedback.

- Following collation of the feedback from the expert panel, TEHAI was refined to incorporate panel members feedback and was then finalised

# TEHAI


Use case
Internal Validity
External Validity
Objective
Performance Metrics
Capability
Dataset Source and Integrity
Use in a Healthcare Setting
Number of services
Adoption
TEHAI
Utility
Generalisability and Contexualisation
Alignment with Domain
Safety and Quality
Transparency
Privacy
Non-Maleficence
Technical Integration


TEHAI
1. Capability
•1.1.Objective
•1.2. Use Case
•1.3. Dataset Source and Integrity
•1.4. Performance Metrics
•1.5. Internal Validity
•1.6. External Validity
2. Utility
•2.1. Generalisability and Contextualisation
•2.2. Safety and Quality
•2.3. Transparency
•2.4. Privacy
•2.5. Non-Maleficence
3. Adoption
•3.1. Use in a Healthcare Setting
•3.2. Number of Services
•3.3. Alignment with Domain
•3.4. Technical Integration

# TEHAI

# TEHAI Components

**Capability:** This component assesses the intrinsic technical capability of the AI system to perform its expected purpose, by reviewing key aspects as to how the AI system was developed

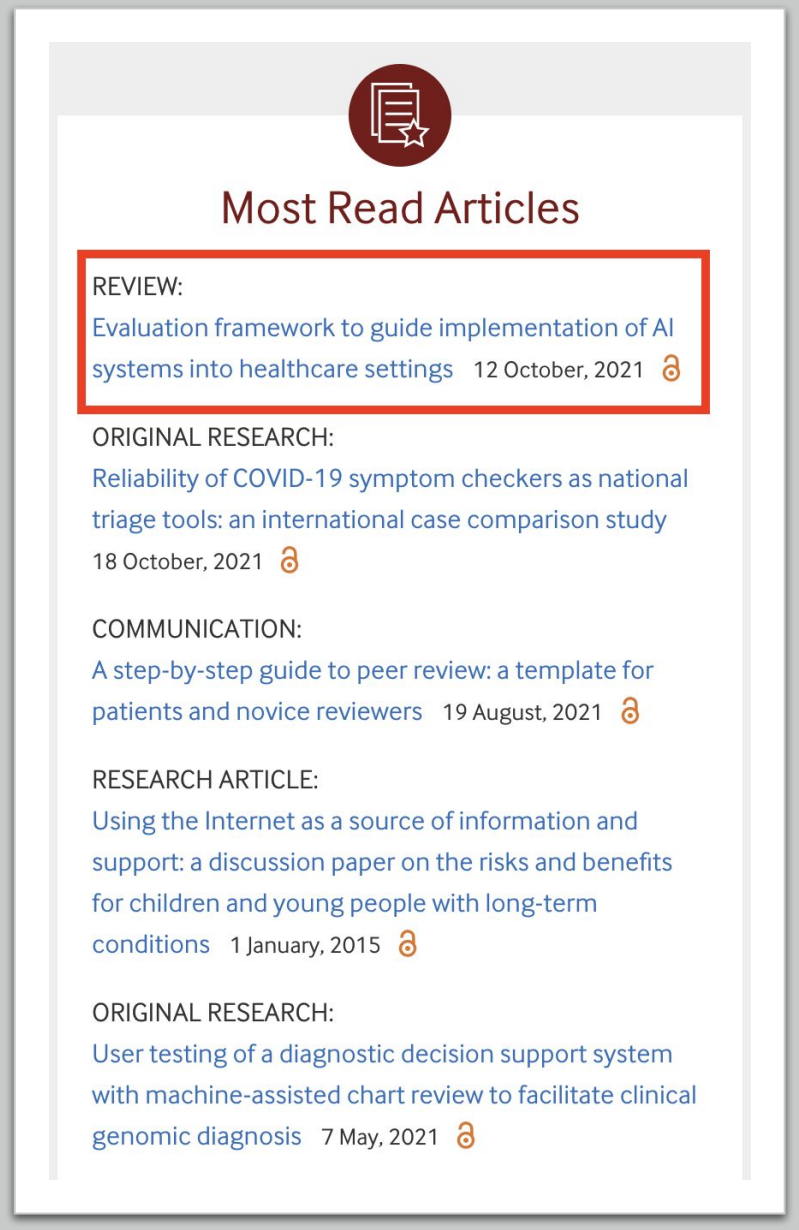
**Utility:** This component evaluates the usability of the AI system across different dimensions including the contextual relevance, and safety and ethical considerations. It also assesses the efficiency of the system

**Adoption:** This component appraises translational value by evaluating key elements that demonstrate the adoption of the model in real life settings

# TEHAI Scoring

| Component | Sub-component | Initial Score | Weight | Subcomponent Score= Initial Score x Weight | |
|---|---|---|---|---|---|
| Capability | *Objective of Study* | 0-3 | 10 | Weight 5 | 0-9<br>10-14<br>15 and above |
| | *Dataset Source and Integrity* | | 10 | | |
| | *Internal Validity* | | 10 | | |
| | *External Validity* | | 10 | | |
| | *Performance Metrics* | | 10 | | |
| | *Use Case* | | 5 | | |
| | | | | | |
| Utility | *Generalizability and Contextualisation* | 0-3 | 10 | Weight 10 | 0-19<br>20-29<br>30 and above |
| | *Safety and Quality* | | 10 | | |
| | *Transparency* | | 10 | | |
| | *Privacy* | | 10 | | |
| | *Non-Maleficence* | | 10 | | |
| | | | | | |
| Adoption | *Use in a Healthcare Setting* | 0-3 | 10 | | |
| | *Technical Integration* | | 10 | | |
| | *Number of Services* | | 5 | | |
| | *Alignment with Domain* | | 5 | | |

# TEHAI usability




# Publication

Reddy, S et al. (2021). Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health & Care Informatics* 2021;**28:**e100444.

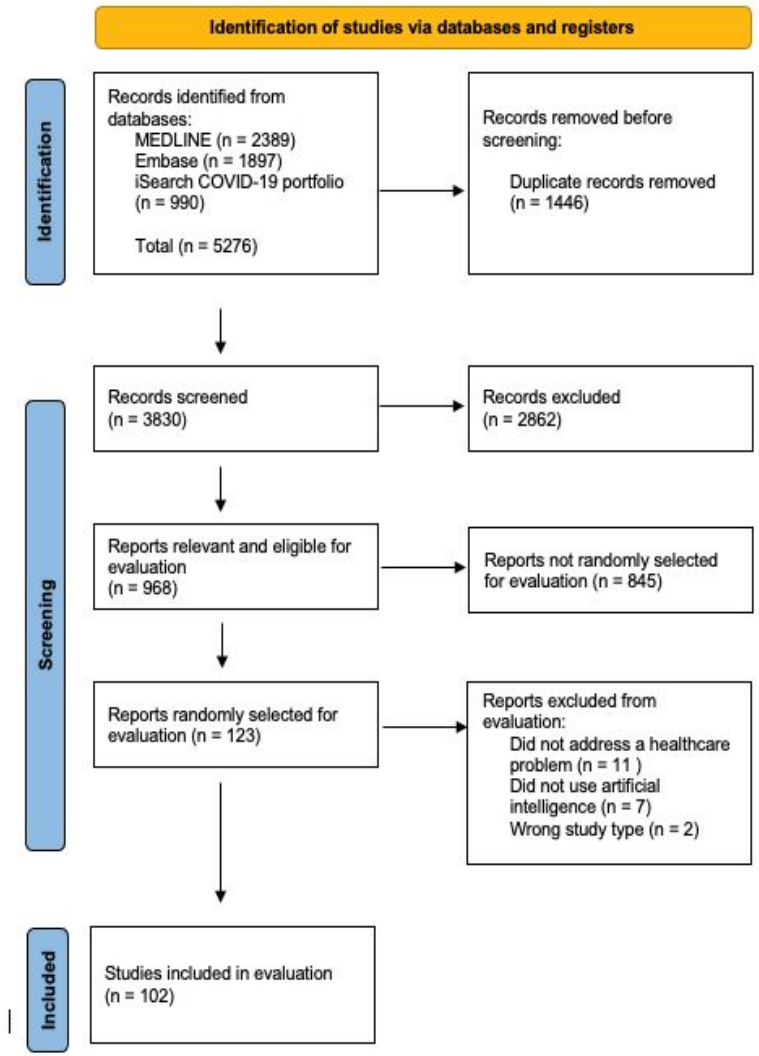
# Application

- The emergence of the COVID-19 pandemic has resulted in several papers outlining the utility of AI in tackling various aspects of the disease like diagnosis, treatment, and surveillance

- Some recent reviews have outlined how most of these studies or the AI applications presented in these studies have shown minimal value for clinical care

- To further assess the translational gaps of the COVID-19 AI studies, we decided to apply TEHAI to COVID-19 AI studies

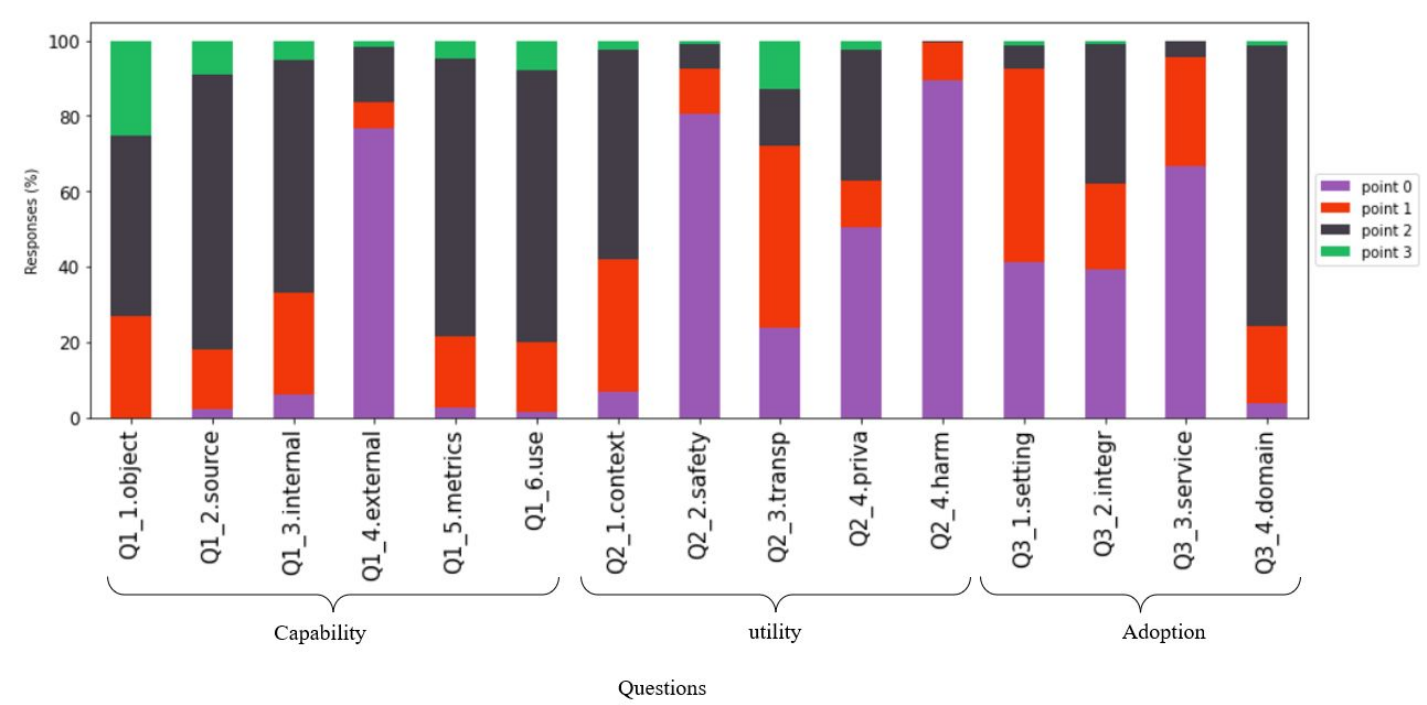
# Systematic Review



# Evaluation Process

- Evaluation and data extraction was conducted using Covidence software
- A combined data extraction / quality assessment template based on the TEHAI framework was created in Covidence to facilitate this
- Reviewer roles were randomly assigned across the evaluation team
- Each paper was viewed by two reviewers who independently evaluated the paper against the elements of the TEHAI framework and extracted relevant data
- Evaluation scores and extracted data from each reviewer were compared by a third reviewer for agreement. This third reviewer also resolved any discrepancies

# Findings




- 10 Reviewers reviewed a total of 102 manuscripts
- On raw scores, “capability” scored the highest compared with “utility” and “adoption”

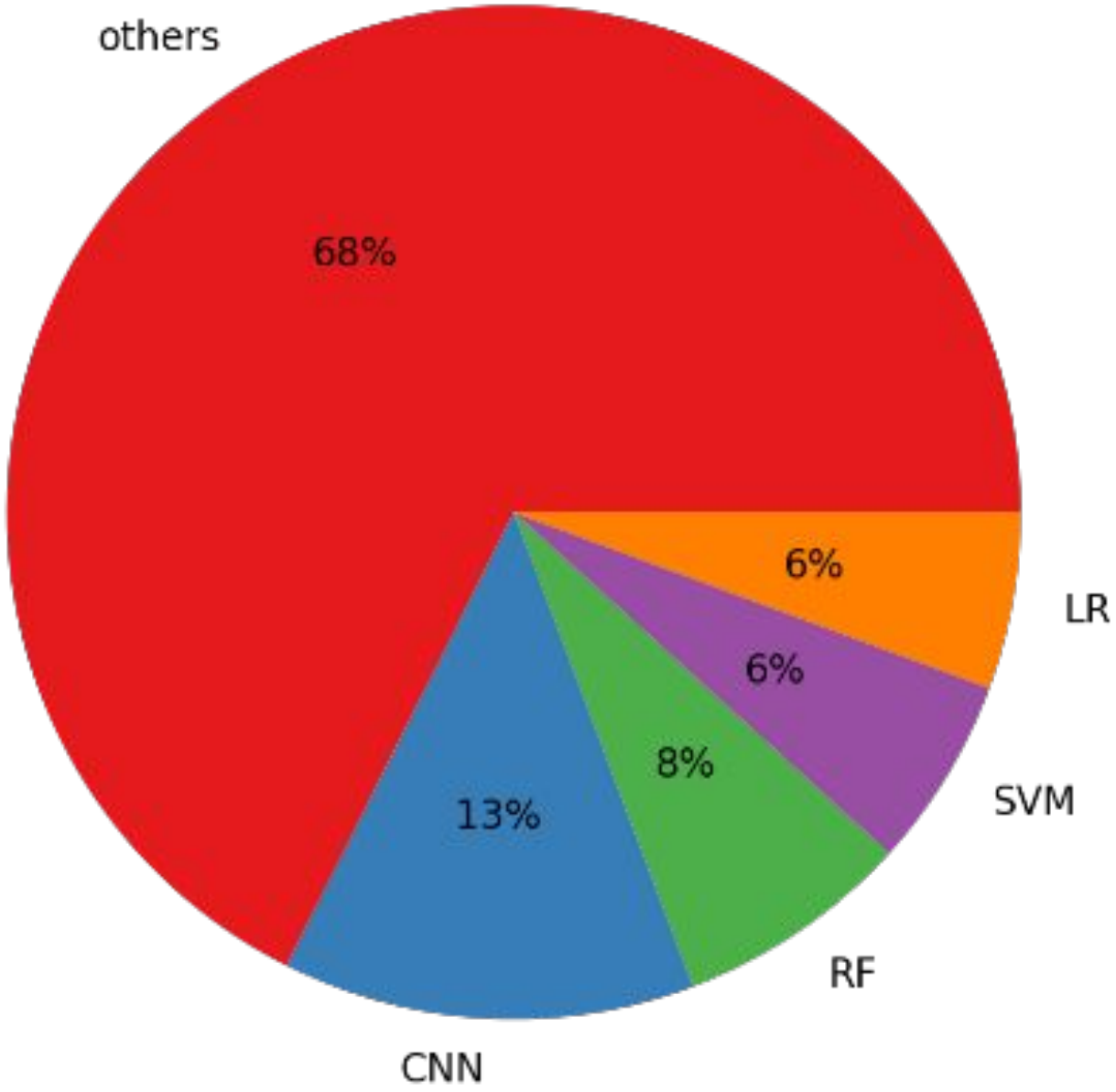
# Findings

| Component | 2-point average | 3-point average |
|---|---|---|
| **Capability** | 53 percent | 8 percent |
| **Utility** | 20 percent | 4 percent |
| **Adoption** | 27 percent | 1 percent |

# Findings

- We extracted the machine learning models used in COVID-19 studies and found that Convolutional Neural Network (CNN) was the most used machine learning model followed by the Random Forest (RF) algorithm

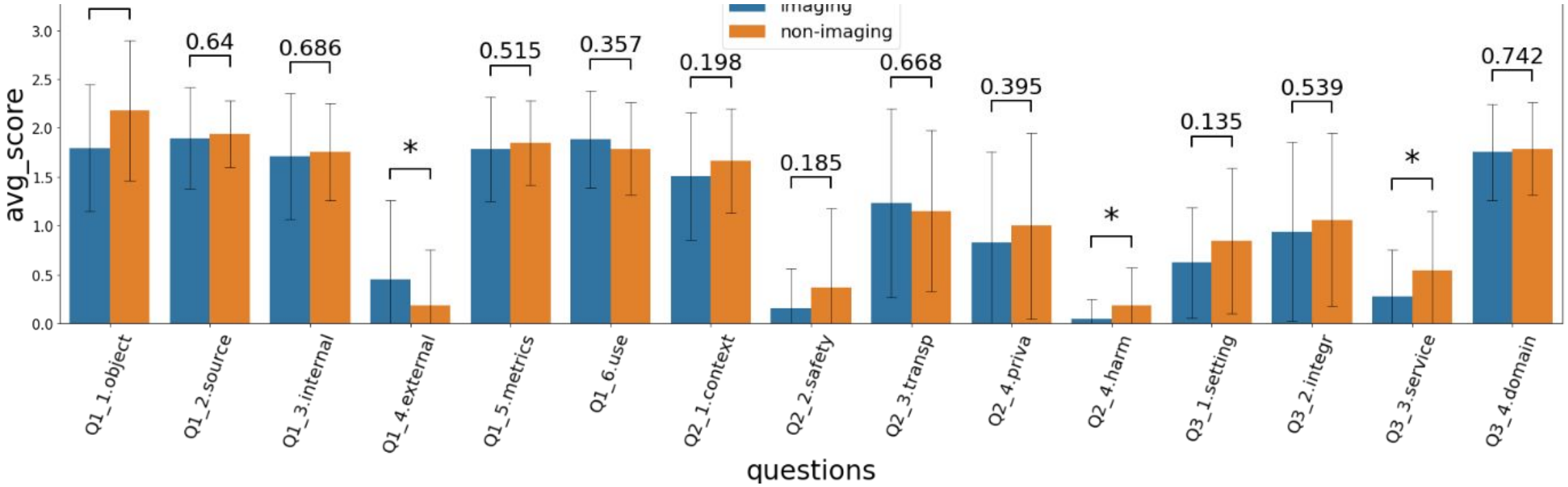
- It shows that automatic feature extraction based on deep learning models are becoming more prevalent in the clinical data analysis



# Findings

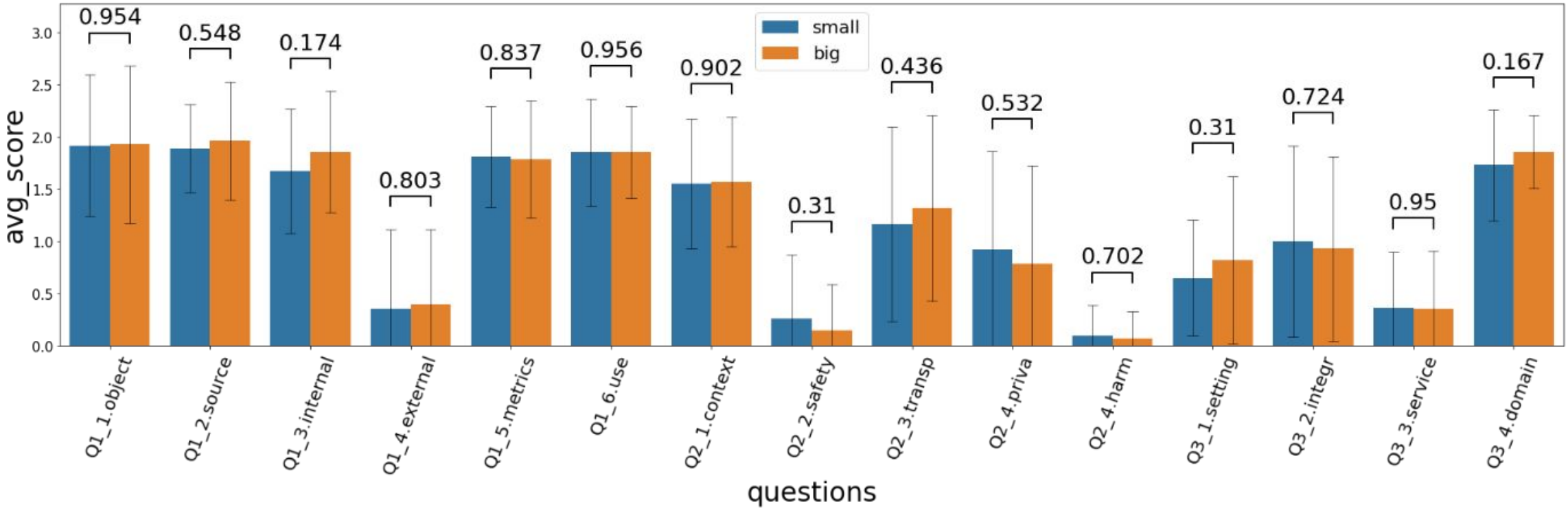
- Based on the average score and variability of data per question for both non-imaging and imaging studies , Non-imaging studies scored better than imaging studies in objective, safety and non-maleficence and number of services

- While imaging studies scored higher than non-imaging studies external validity

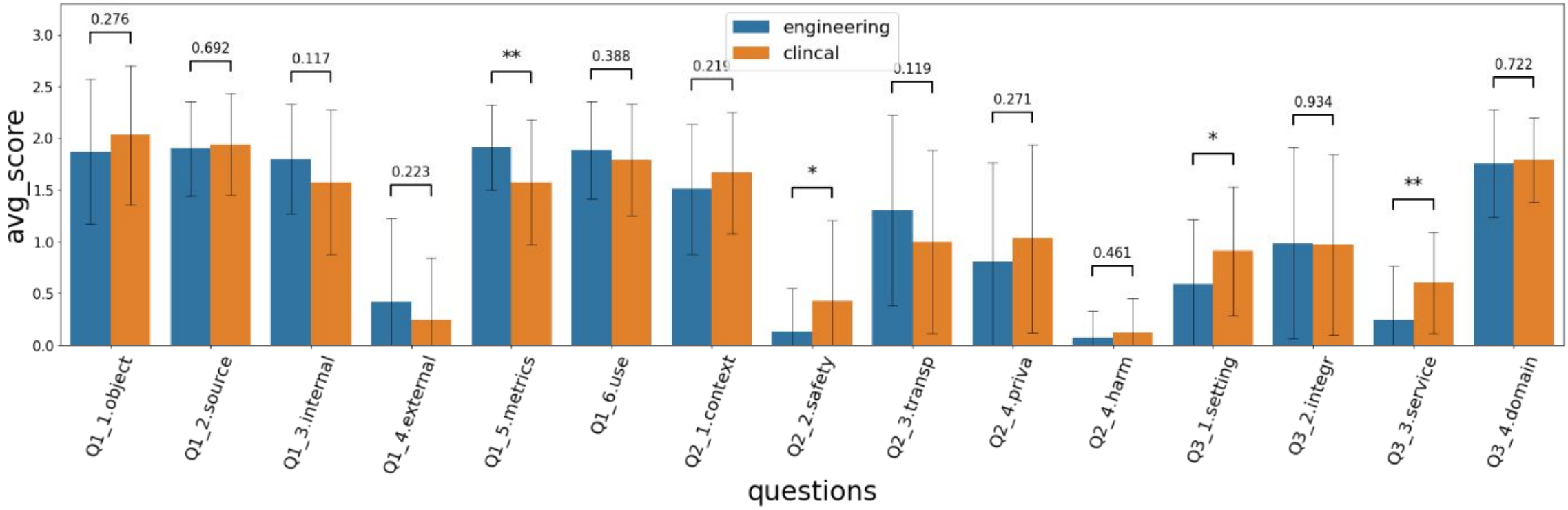


# Findings

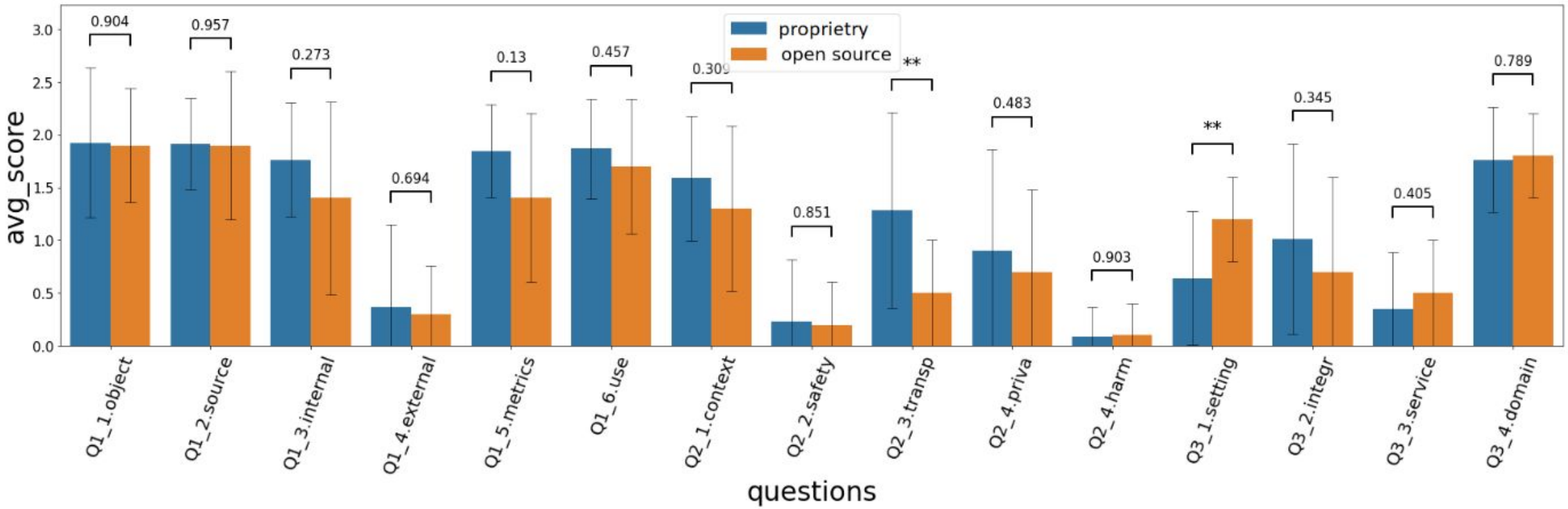
- The average score, variability of score and P values for both big and small datasets.
- The P-values with less than 0.01 and 0.001 are presented with * and ** respectively



# Findings


engineering
clincal
avg_score
questions
0.276
0.692
0.117
**
0.388
0.219
*
0.119
0.271
0.461
*
0.934
**
0.722
0.223
Q1_1.object
Q1_2.source
Q1_3.internal
Q1_4.external
Q1_5.metrics
Q1_6.use
Q2_1.context
Q2_2.safety
Q2_3.transp
Q2_4.priva
Q2_4.harm
Q3_1.setting
Q3_2.integr
Q3_3.service
Q3_4.domain

# Findings


proprietry
open source
avg_score
questions
0.904
0.957
0.273
0.694
0.13
0.457
0.309
0.851
**
0.483
0.903
**
0.345
0.405
0.789
Q1_1.object
Q1_2.source
Q1_3.internal
Q1_4.external
Q1_5.metrics
Q1_6.use
Q2_1.context
Q2_2.safety
Q2_3.transp
Q2_4.priva
Q2_4.harm
Q3_1.setting
Q3_2.integr
Q3_3.service
Q3_4.domain

| Study | Author | Capability | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Objective of Study | Dataset Source and Integrity | Internal Validity | External Validity | Performance Metrics | Use Case | Final Score |
| **COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images.** | **Afshar et al 2020** | 30 | 30 | 20 | 30 | 20 | 10 | **140** |
| **Development and evaluation of an artificial intelligence system for COVID-19 diagnosis.** | **Jin et al 2020** | 30 | 20 | 30 | 20 | 20 | 10 | **130** |
| **Development and external validation of a prediction risk model for short-term mortality among hospitalized U.S. COVID-19 patients: A proposal for the COVID-AID risk tool.** | **Hajifathalian et al 2020** | 30 | 20 | 20 | 20 | 20 | 10 | **120** |
| **4S-DT: Self Supervised Super Sample Decomposition for Transfer learning with application to COVID-19 detection** | **Abbas et al 2020** | 10 | 30 | 20 | 20 | 20 | 15 | **115** |

| Utility | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Study** | **Author** | **Generalizability and Contextualisation** | **Safety and Quality** | **Transparency** | **Privacy** | **Non-Maleficence** | **Final Score** |
| **Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study.** | **An et al 2020** | 20 | 20 | 30 | 20 | 10 | **100** |
| **Prognostic Modeling of COVID-19 Using Artificial Intelligence in the United Kingdom: Model Development and Validation.** | **Abdulaal et al 2020** | 20 | 0 | 30 | 20 | 10 | **80** |
| **Extracting Possibly Representative COVID-19 Biomarkers from X-ray Images with Deep Learning Approach and Image Data Related to Pulmonary Diseases** | **Apostolopoulos et al 2020** | 30 | 10 | 30 | 10 | 0 | **80** |

| Adoption | | | | | | |
|---|---|---|---|---|---|---|
| **Study** | **Author** | **Use in a Healthcare Setting** | **Technical Integration** | **Number of Services** | **Alignment with Domain** | **Final Score** |
| **Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool.** | **Das et al 2020** | 30 | 20 | 10 | 10 | **70** |
| **Development and external validation of a prediction risk model for short-term mortality among hospitalized U.S. COVID-19 patients: A proposal for the COVID-AID risk tool.** | **Hajifathalian et al 2020** | 30 | 20 | 5 | 10 | **65** |
| **Identification of risk factors for mortality associated with COVID-19.** | **Yu et al 2020** | 20 | 20 | 10 | 10 | **60** |
| **Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography.** | **Zhang et al 2020** | 10 | 20 | 10 | 10 | **50** |
| **Using Artificial Intelligence for COVID-19 Chest X-ray Diagnosis.** | **Borkowski et al 2020** | 10 | 20 | 0 | 10 | **50** |

# Learning

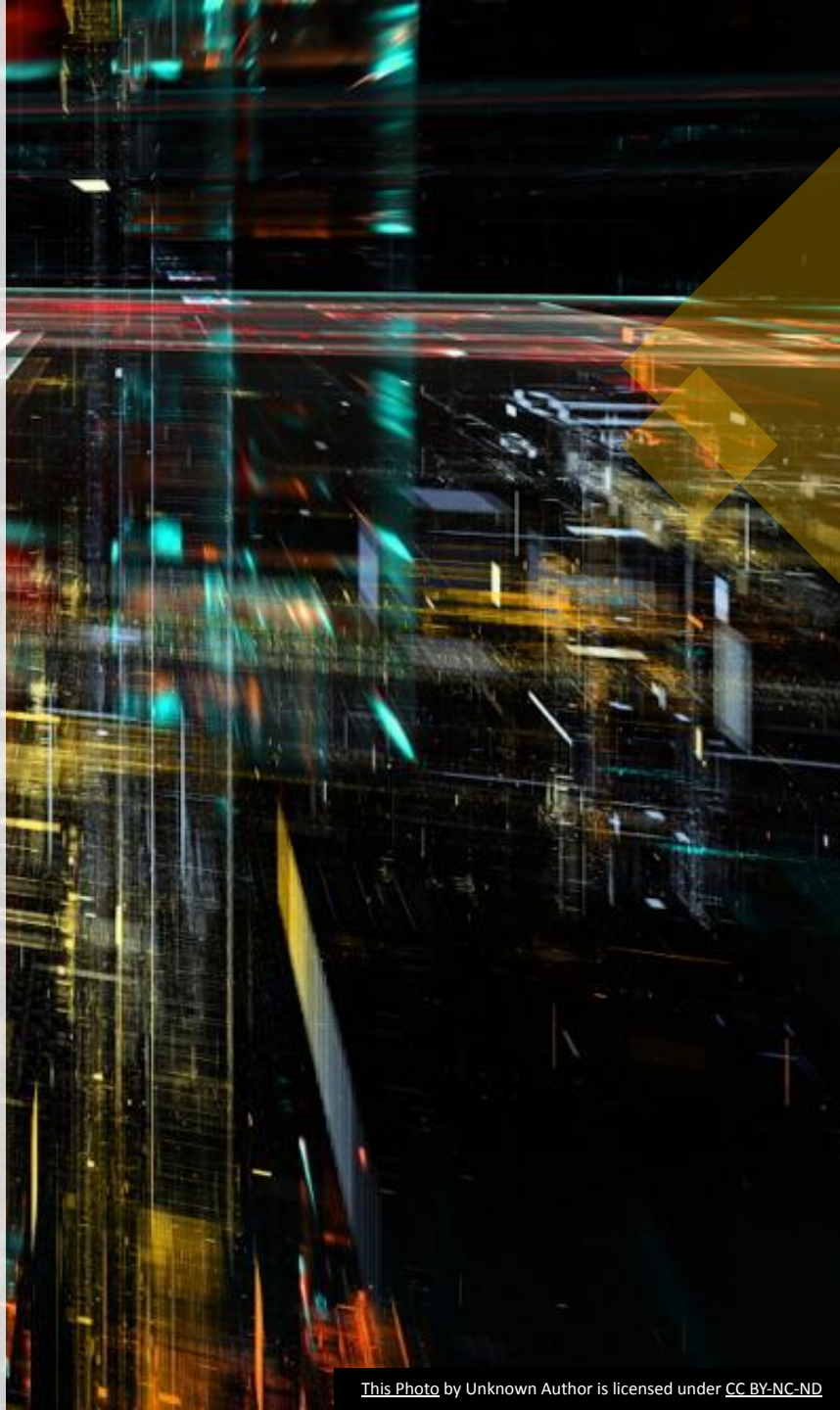
- In TEHAI, equal weighting is placed on both internal and external validity. From a translational point of view, the model performance should extend beyond the test environment and perform well on external datasets
- In relation to this, the top scoring studies in the capability component performed well in these measures
- Afshar et al(17) utilised pre-training and transfer learning of their capsule network model on an external dataset of 94,323 X-ray images
- Jin et al(14) utilised a combination of three medical centres and four public datasets to compile 11,356 CT scans to test their AI model

# Learning

- The ‘utility’ component assesses how safely can the AI model be used in healthcare
- Very few studies scored well across the criteria and scored poorly especially with the safety and quality and non-maleficence subcomponents
- One of the distinguishing aspects of TEHAI framework compared to other evaluation framework is its assessment of how well the AI model is adopted
- This is assessed through the actual use of the AI model in health services or healthcare delivery. Considering many of the COVID-AI models were experimental and the time frames we assessed were short, very few included studies did well in the component
- The top-ranking study in this component was an online COVID-19 mortality prediction model that was deployed as an open-source tool making it highly accessible and adoptable.

# In conclusion

- TEHAI- A comprehensive evaluation framework
- Three main components (Capability, Utility and Adoption) and 15 subcomponents
- Can be used in development, deployment and discernment stages
- Applied to COVID-19 AI studies
- Very few studies have a translational component i.e., did poorly in utility and adoption components
- Therefore, evaluation has to in-built in product/application development cycle



# Discussion/Questions