MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

What Can Go Wrong?

PANEL: The Pitfalls And Dilemmas When Evaluating Artificial Intelligence (AI) Applications In Healthcare Settings

Jack (Yu-chuan) Li, M.D., Ph.D, FACMI, FIAHSI

jack@tmu.edu.tw

President, IMIA

Distinguished Professor, Taipei Medical University, Department of Biomedical Informatics, Taipei, Taiwan

AIDH IMIA

Stephen Hawking

"Our future is a race between the growing power of technology and the wisdom with which we use it."

Type of Failures

- Biased/Incomplete **Training** Data
- Biased/Incomplete **Input** Data
- **"Dataset Shift"** Problems

Biased/Incomplete Training Data

- Machine learning AI – Data is Key
- Biased data → biased engine → biased results
- Incomplete data → leaky engine → low accuracy
- Type of Biases
  - Selection Bias – remove noise along with valid signal
  - Sampling Bias – what kind of patients
  - Labeling Bias – how the data were labeled, who labeled them
  - ….etc.

Biased/Incomplete Input Data

- For a trained ML engine, how and what inputs were fed into it?
- A missing input is NOT a negative input
  Unknown <> Negative
- IBM Watson for Oncology
  - Hit-and-misses
  - Direct EHR copy and paste
  - Many missing/ambiguous input

"Dataset Shift" Problems

N Engl J Med 2021; 385:283-286,
DOI: 10.1056/NEJMc2104626

- **Change in Technology**
  - New devices
  - New IT practice/systems
- **Change in Population Setting**
  - New demographics – ethnicity, locality…
  - New clinical practice setting – specialty change…
  - New treatment/standard of care
  - Change in incidence of diseases/seasonality/new diseases ("black swan")

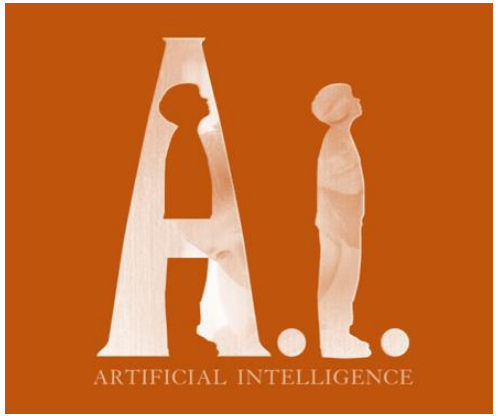
"Dataset Shift" Problems (cont.)

- Change in Behavior
  - New behavioral incentives for clinical practice (sepsis reimbursement…)
  - New patient behavior (Dx on celebrities, major social events…)
  - Changes in clinical practice (new order sets, surgical skin markings…)
  - Changes in clinical nomenclature (competing guidelines, DSM-5…)
  - The use of AI changes behavior (over-reliance, automation bias…)

N Engl J Med 2021; 385:283-286,
DOI: 10.1056/NEJMc2104626

Possible Solutions

- "Bridge2AI" from NIH
- Federated Learning
- FAIR (Findable, Accessible, Interoperable, and Reusable) Data Sets
- A Formal Evaluation Process (Ted)



Thank You.

jack@tmu.edu.tw
https://jackli.cc/

MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Role of Evaluation Throughout the Life Cycle of Medical AI Applications

PANEL: The Pitfalls And Dilemmas When Evaluating Artificial Intelligence (AI) Applications In Healthcare Settings

Edward H. (Ted) Shortliffe

Chair Emeritus & Adjunct Professor

Columbia University Department of Biomedical Informatics, New York City

ted@shortliffe.net

AIDH IMIA

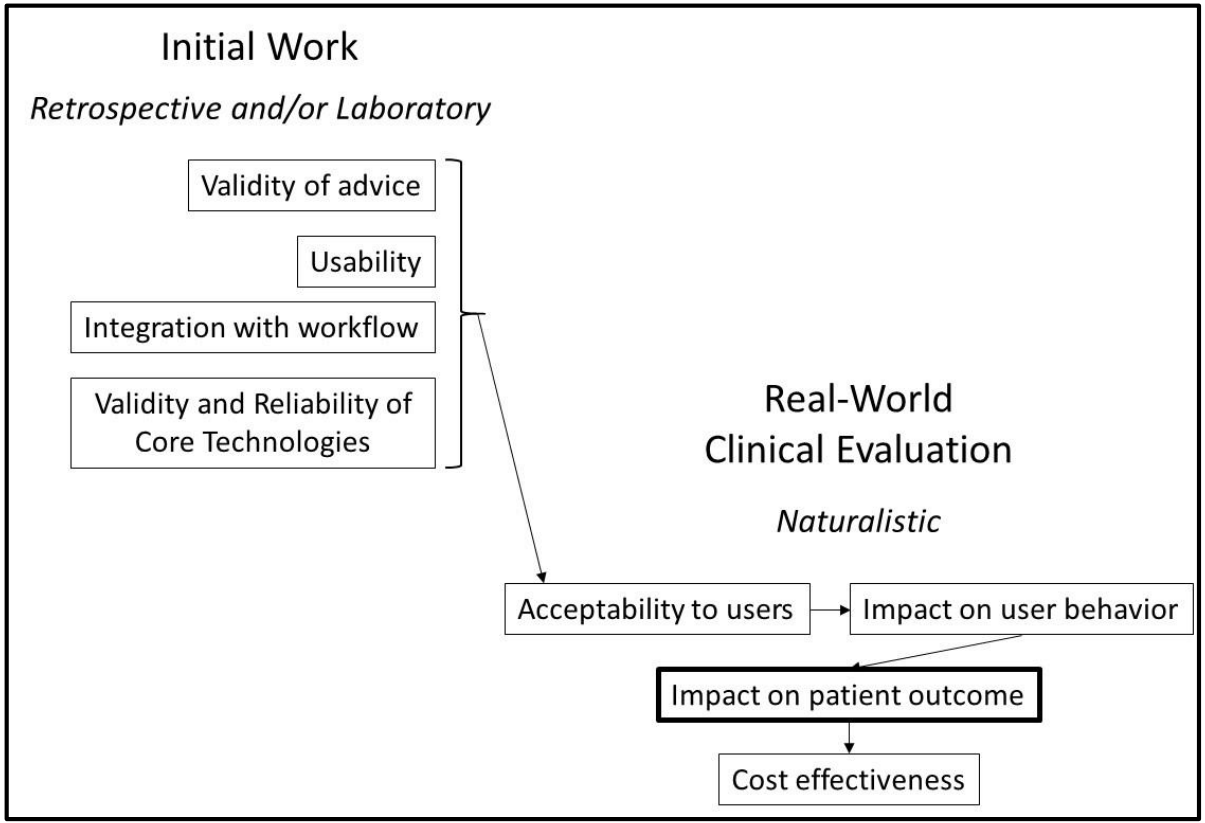
Some Observations Regarding the Evaluation of AI Systems in Health and Medicine

- Tendency to emphasize decision-making performance – often to the exclusion of other key areas
- Issues much discussed in academic circles but too often overlooked by industry in the rush to bring a product to market
- Too often a failure to engage representatives of user community throughout
- Need to emphasize value to anticipated user community

Staged AI or Decision-Support Evaluation

- Demonstrating the quality of key component technologies
- Demonstrating the validity of the advice or interpretation
- Demonstrating acceptability to users
- Demonstrating impact on user behavior
- Demonstrating impact on outcome
- Demonstrating cost effectiveness

Too often overlooked or left for a future study



Staging from Laboratory to Naturalistic Settings -- Iterative Process

Scientific Steps in The Evolution of AI (and other) Informatics Research and Development Efforts

Before:
(definition and design)
- Identify
- Partner
- Analyze
- Motivate
- Create

During:
(execution and assessment)
- Innovate
- Implement
- Assess

After:
(reflection and communication)
- Generalize
- Critique
- Share
- Inspire

Some Implications

- Clinical AI systems (including commercial products) require a time-consuming staged evaluation plan that demonstrates more than quality of advice or interpretation
- Early adopters naturally become partners in evaluation studies, since many questions cannot be answered in the laboratory and must be assessed in actual clinical use environments

MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Thank You

ted@shortliffe.net

https://www.shortliffe.net

MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Science of Cognition and Evaluation of AI System Healthcare Implementation

PANEL: The Pitfalls And Dilemmas When Evaluating Artificial Intelligence (AI) Applications In Healthcare Settings

Vimla L. Patel

vpatel@nyam.net

Senior Research Scientist

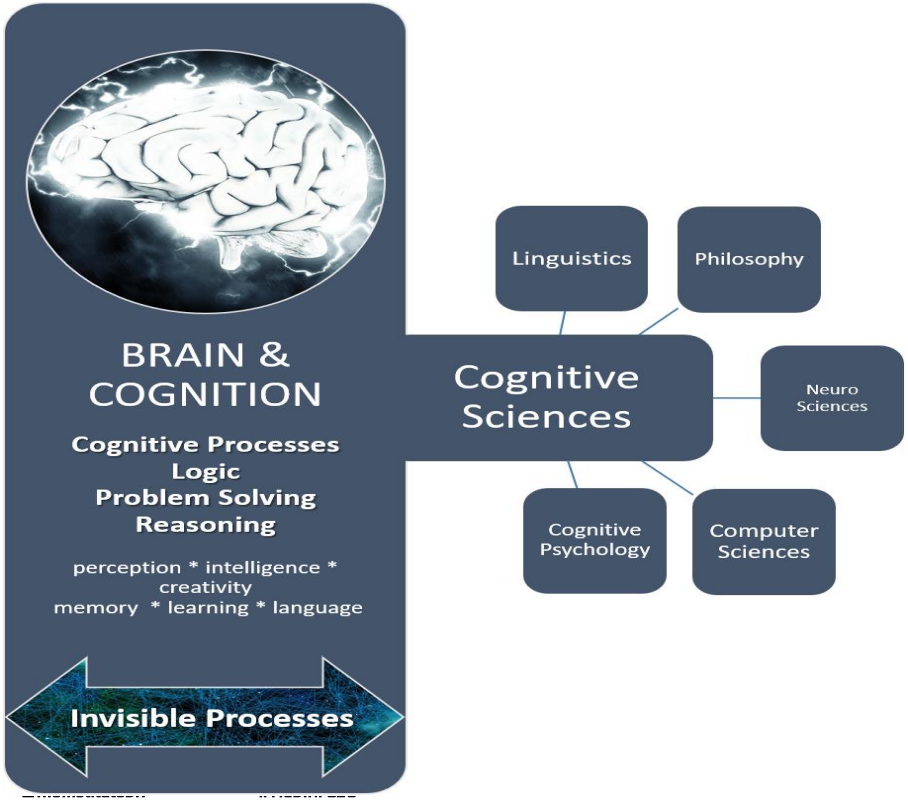
Cognitive Studies in Medicine and Public Health

The New York Academy of Medicine, NY, USA

AIDH IMIA

Cognitive Evaluation of AI Systems in Clinical Environment

- Complexity of clinical environment
- Human expertise in natural clinical practice
- Evaluating AI implementation in healthcare settings requires comprehensive evaluation (including impact on cognition) beyond computational performance
  - Identify challenges of using AI in clinical practice, in which predictability, explainability, transparency, and safety are critical



From: Marisa Tschopp
Human Cognition and Artificial Intelligence — A Plea for Science
April 2018.

https://www.scip.ch/en/?labs.20180419

Understanding What is Underneath

Human Problem Solving and AI

Newell, A., & Simon, H. A. (1972). Human Problem Solving. Englewood Cliffs, NJ: Prentice-Hall.

- Elaborates a comprehensive theory of human problem solving
  - Picture courtesy of Carnegie Mellon University: *Shelf1.library.cmu.edu*

This work provided a foundation for the formal investigation relating human problem solving to research in artificial intelligence.

Methods of evaluation driven by science of cognition (1)

- Use of think-aloud and protocol-analytic methods in investigations of high-level cognition (comprehension, problem solving, decision making)
- In dynamic clinical settings over time, with and without AI support
- Level of analysis using NLP, a framework for formal investigation of symbolic-information processing, before and after AI implementation

Methods of evaluation driven by science of cognition (2)

- Realtime navigation tracking in clinical settings using computational ethnography
  - Leverages automated means (e.g., sensors and log files) for collecting data reflective of real end users' actual, unaltered behavior versus machine-generated behavior in clinical workflow
  - Technology changes they way we think and make decisions
- Identifies challenges in cognition and navigation.

Thank You

vpatel@nyam.org
http://lodhia-patel.net

MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Data bias and fairness impact on AI Applications Evaluation

PANEL: The Pitfalls And Dilemmas When Evaluating Artificial Intelligence (AI) Applications In Healthcare Settings

Usman Iqbal

Senior Clinical Data Consultant and Adj. Assoc. Professor

School of Population Health, Faculty of Medicine and Health, University of New South Wales, Sydney, Australia

Usman.iqbal@unsw.edu.au

AIDH IMIA

Data bias and fairness

- AI systems in healthcare settings are trained on large datasets that may contain biased information.
- If the training data is biased in terms of gender, race, or socioeconomic factors, the AI model may perpetuate these biases.
- This can lead to unequal treatment or disparities in healthcare outcomes.
- It is crucial to address and mitigate biases in the training data to ensure fairness in AI applications.

Data bias and fairness

External Validation and Calibration

- Clinical AI models trained under specific assumptions and contexts may not perform well outside their original setting.
- External validation and re-calibration of models are necessary to understand their utility in different populations.
- It is important to assess model bias and underrepresentation to ensure equitable application of AI findings.

PLOS Digital Health 1(3): e0000022. https://doi.org/10.1371/journal.pdig.0000022

Transparency

- Some AI algorithms, like deep learning neural networks, can be opaque and difficult to interpret.
- This lack of transparency raises concerns about accountability and understanding the AI system's decision-making process.
- Identifying potential errors or biases becomes challenging without transparency.
- Efforts should be made to develop explainable AI models that provide insights into the reasoning behind their decisions.

Ethical Dilemmas

The use of AI in healthcare raises ethical dilemmas that need careful consideration.

- For example, there may be conflicts between individual privacy and the need for data sharing to improve AI models.
- AI systems may also impact human autonomy, such as in decision-making processes.
- Ethical frameworks and guidelines should be developed to navigate these dilemmas and ensure responsible use of AI in healthcare settings.

EBioMedicine https://doi.org/10.1016/j.ebiom.2023.104525

Detect, Prevent, and Mitigate Bias

- Self-motivated and fairness-aware modeling pipeline
- Detailed description of datasets to detect potential bias
- Pre-processing existing data using appropriate sampling methods
- Fairness metrics as bias-monitoring and fairness-evaluation tools
- Representation learning and adversarial thinking to eliminate sensitive information

EBioMedicine https://doi.org/10.1016/j.ebiom.2023.104525

Thank You

usman.iqbal@unsw.edu.au