Benchmarking AI in Healthcare

An industry perspective | Topic Group Symptom Assessment

Clinical Evaluation Working Group Workshop, 14 Oct 2020

Dr Shubhanan Upadhyay
GP/Urgent Care
Clinical Expert, Ada Health

Motivations

Why?

- A global healthcare challenge: Growing shortage of health workers and growing population
  - 7.2 million shortfall in 2013 to 12.9 million by 2035
- Potential for AI solutions to help to close this gap
- Industry perspective: Symptom based decisions support tools
  - “Symptom-Checkers” for supporting patients and health workers
- For safe adoption and trust robust clinical evaluation needed.
- Standardized Benchmarking could fit within a spectrum of clinical evaluation that accounts for the special considerations needed for AI tools

Gaining Trust in healthcare

AI for Health
An ITU Focus Group
In collaboration with WHO

- Clinical Evidence
- SaMD - Post Market Clinical Follow up

- Rapidly iterating/evolving tools

- All peers in this field working on their own:
  - Internal testing and benchmarking
  - Real world evidence gathering

- **But - trust low.** High quality, peer-reviewed prospective clinical studies essential but slow

- ITU/WHO Focus Group is the forum to align industry/those building AI tools with key stakeholders to ensure
  - outputs/metrics answer relevant questions for different stakeholders (health systems, regulators, governments)

- Recognition from all those in our topic group that **we need to work together** to build trust for the whole industry to succeed.

Case Study: TG Symptom

- Needs bottom-up development of frameworks, with alignment and oversight from key stakeholders

- **Companies**
  - *1DOC3, Ada, Babylon, Baidu, Barkibu, Buoy, Deepcare, EQL, Infermedica, Inspired Ideas, Isabel Healthcare, mfine, MyDoctor, Symptify, Visiba Care, Xund.ai, Your.MD*

- **Independent expert contributors (clinical, technical, ethics, anthropology, ontologies)**
  - *Reza Jarral, Thomas Neumark, Muhammad Murhaba, Pritesh Mistry, Alejandro Osornio, Salman Razzaki*

- **Standardised benchmarking framework** for symptom assessment tools:
  - Common inputs (symptoms, attributes)
  - Independently curated data set
    - High quality
    - Representative, global
  - Common outputs (triage advice, condition suggestions, differential diagnoses)
  - General and context relevant metrics answering questions for different stakeholders

- All the technical/clinical/other considerations, nuance needs to be gathered

- **Oversight from WHO, ITU, independent academic/clinical experts, regulators**

Where have we got to so far?

- Collaboration is key

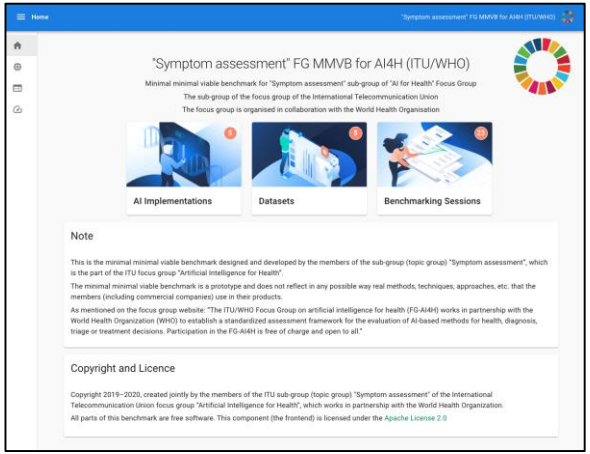
- Challenges
  - Mapping to common ontology
  - Ground truth - variability
  - Different intended uses/users
  - Performance vs Utility - capturing impact/clinical outcomes - clinical metrics

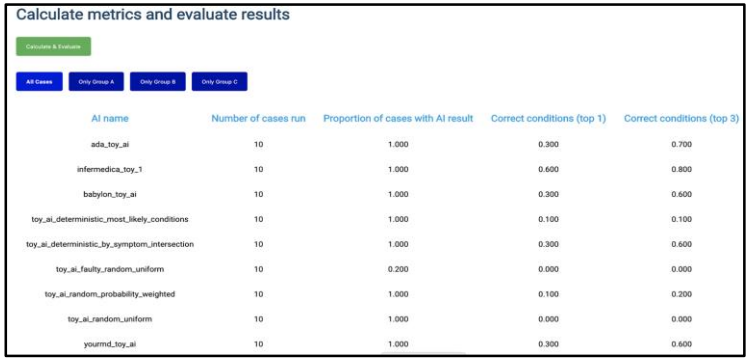
- Online and in person collaboration between tech and clinical experts from each company, plus independent contributors.

- Minimal Minimal Viable Benchmarking
  - A Tool that
    - Allows cases to be created from a model created by clinicians
    - Allows doctors to create labelled test cases in a universal format
    - Has an interface with AI tools and generates **metrics**
    - Allows us to **add complexity**

Where have we got to so far?

- **Meeting A – Geneva, 25-27 September 2018**
  - A-020: Towards a potential AI4H use case "diagnostic self-assessment apps"
- **Meeting C – Lausanne, 22-25 January 2019**
  - Topic Group created
- **Approach:**
  - MMVB 1.0 … MMVB 2.0, etc
- **Meeting F – Zanzibar 2-5 September 2019**
  - First benchmarking with toy AI & toy Data
  - **Minimal Minimal viable Benchmarking**
  - **MMVB 1.0**
- **Meeting G, H, I, J (Sept 2020)**
  - Gradual increase of complexity by adding model detail





AI name	Number of cases run	Proportion of cases with AI result	Correct conditions (top 1)	Correct conditions (top 3)
ada_toy_ai	10	1.000	0.300	0.700
infermedica_toy_1	10	1.000	0.600	0.800
babylon_toy_ai	10	1.000	0.300	0.600
toy_ai_deterministic_most_likely_conditions	10	1.000	0.100	0.100
toy_ai_deterministic_by_symptom_intersection	10	1.000	0.300	0.600
toy_ai_faulty_random_uniform	10	0.200	0.000	0.000
toy_ai_random_probability_weighted	10	1.000	0.100	0.200
toy_ai_random_uniform	10	1.000	0.000	0.000
yourmd_toy_ai	10	1.000	0.300	0.600

What might it mean?

- Transcending old, legacy paradigms of “it needs to be pharma-like intervention RCT” thinking to something more **fit for purpose**

- Meaningful engagement and alignment with
  - Health system and clinical decision makers
  - Regulators and notified bodies (IMRF)

- Helping decision makers assess which tools suitable for their context (part of a holistic 360 assessment)

- **Performance -----> Impact**
  - AI tools are safer and better integrated into health ecosystem, with demonstrable clinical outcomes

- Lots of challenges and complexity to overcome. **Collaboration**

Thank you

Dr Shubhanan Upadhyay
shubs.upadhyay@ada.com