|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-C-025 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | Lausanne, 23-25 January 2019  |
| **DOCUMENT** |
| **Source:** | Your.MD |
| **Title:** | Clinical Evaluation of AI Triage and Risk Awareness in Primary Care Setting |
| **Purpose:** | Discussion |
| **Contact:** | Jonathon Carr-BrownYour.MDUK | Tel: +44 7900 271580Fax: N/AEmail: jcb@your.md |
| **Contact:** | Matteo BerlucchiYour.MDUK | Tel: +44 7867 788348Fax: N/AEmail: matteo@your.md  |

|  |  |
| --- | --- |
| **Abstract** | **Overview**: Ninety per cent of people who seek help from primary care have minor ailments and injuries. The vast majority (>75%) attend primary care because they lack an understanding of the risk they face or the knowledge to care for themselves.**Proposal**: Your.MD wants to help develop a standard testing protocol that can be replicated and applied to other triage and symptom checking services.**Relevance:** AI developed by Your.MD can accurately deliver best and safe next steps advice and practical self-care information by assessing individual risk. This is relevant for LMICs with limited primary care and for reducing pressure on primary care in developed countries.**Impact:** Your.MD’s algorithms can provide high quality triage and self-care advice to reduce costs and transform quality of care. Benchmarking is essential to ensure confidence in the service and to facilitate behaviour change.**Existing Work**: Your.MD was first to launch an AI chat-based triage service and operates globally performing over 500,000 consultations a month in English. Problem is currently addressed by inaccurate linear general triage processes.**Feasibility**: The technology exists but there is no meaningful evidence to support the use of patient-facing symptom checkers.**Data**: Your.MD has well annotated data covering millions of consultations. Your.MD has also developed a proprietary medical model linking hundreds of conditions to thousands of symptoms and influencing factors.**Benchmarking**: There is no standard for triage systems based on AI algorithms. We use both Bayesian and Machine Learning and a ‘gold standard’ is needed to assess the reliability and safety of such systems.**Organisation**: Your.MD has 4 year’s experience in the field, a team of 40 people, won the UNESCO/NetExplo Award and the Horizon 2020 Seal Excellence. |

# 1 Overview

Ninety per cent of people who seek help from primary care have minor ailments and injuries [1]. The vast majority (>75%) attend primary care because they lack an understanding of the risk they face or the knowledge to care for themselves.

In the United Kingdom alone, there are 340 million consultations at the GP every year and the current system is being pushed to do more with less resources. Artificial intelligence has the potential to save national health services time and money, improve patient and health worker experience and deliver safer diagnosis, disposition and management.

Various new models of care have been or are being trialled in primary care with a view to better managing demand. The private sector has seen an expansion of new models which attempt to create disruptive models of care; on demand video consultants (PushDoctor, Babylon), webform based e-consults for specific verticals such as sexual health (Dr Ed) and artificial intelligence-based history and triaging tools (Your.MD and Babylon).

As yet neither video consultations (especially on demand video consultations) nor artificial intelligence symptoms checkers have gained traction in primary care.

The failure of adoption at scale relates to:

* lack of good evidence
* failure to integrate with existing GP workflows and the creation of burdensome parallel workflows
* failure to integrate technically with GP practice IT systems, and
* failure to identify the real customer

Whilst there are existing trials of artificial intelligence against clinical vignettes, there are none which have been set in the context of usual healthcare practice.

# 2 Proposal

Patient-facing symptom checkers are systems in which a patient inputs a series of symptoms pertaining to their current condition.

The system outputs useful clinical information, which may include a diagnosis and/or triage advice. These systems have gained traction in the last two years, and are being actively marketed towards a large potential user base.

However, there is no meaningful evidence to support the use of patient-facing symptom checkers. By ‘no meaningful evidence’, we mean that there is very little evidence to assess the quality of symptom checker algorithms and no evidence that assesses the impact of such systems in real clinical situations. This includes the short-term impact (i.e. is a patient treated correctly/advised to take the right next steps/given the right assessment of personal risk) and the wider systemic impact (i.e. are too many people being referred to primary care or does the tool reduce referrals to primary care).

The dearth of meaningful evaluation, when considered alongside the potential user base and impact of symptom checker technology, means that this topic is highly important and topical.

In December 2018, the UK’s National Institute of Clinical Evidence (NICE) published an Evidence Standards Framework for Digital Health Technologies. Under their framework, symptom checkers are classified as class 3b. This recommends a minimum evidence standard of ‘High quality experimental or quasi-experimental studies showing improvements in relevant outcomes.’

Fraser et al. provide more detailed guidance on how this can be achieved in the case of symptom checkers. They suggest five sequential stages:

1. Problem definition
2. Pre-deployment lab testing – algorithm performance, user interaction
3. Early observational studies in a clinical setting
4. Formal field trials
5. System efficacy

Your.MD is currently seeking support to develop a series of tests that address the first three stages of Fraser’s model. The intention is to develop a protocol in which tests can be replicated and applied to other symptom checkers. This will allow fair comparison between symptom checkers. To ensure fairness, the test must be conducted independently, and the symptom checker should not have sight of the test beforehand.

# 3 Relevance

If Your.MD can demonstrably prove its ability to accurately assess individual risk and deliver best and safe next steps advice and practical self-care information, it and systems like it can have a major impact on global health systems.

Symptom checking AI is relevant for Low and Middle Income Countries (LMIC) with limited primary care where it can be delivered to citizens and community health workers with basic levels of clinical training. In LMICs symptom checking AI can improve the quality, consistency and timeliness of care.

In developed countries these systems can spearhead behaviour change towards more self-care and less reliance on over stretched primary care providers. In the UK appointments for general practitioners rose by 500,000 in November 2018 [2] (up by 1.7% year on year) and the number of GPs working has reduced by more than 1,000 since 2015.

This trend is continuing and leading clinical practices to examine why people attend and if they can be appropriately referred at a pre-primary stage to other sources of help, information and services.

# 4 Impact

Your.MD’s algorithms can provide high quality triage and self-care advice to reduce costs and transform quality of care. Benchmarking is essential to ensure confidence in the service and to facilitate behaviour change.

A proven and trusted symptom checking AI algorithm has the potential to increase self-care while also encourage early consultation for serious conditions. This will lead to both better care, lower costs and greater efficiency in primary care.

Long-term the algorithms can significantly contribute to the creation of a new health care sector - **pre-primary care** which is AI supported, community linked and dovetailed with primary care. In the developed world the aim would be to create a trusted AI algorithm which led, overtime, to 30% of people currently attending primary care either self-caring or self-referring to more appropriate services i.e. pharmacies, physiotherapists or community mental health workers.

This shift will only occur if both professionals and citizens trust the advice AI algorithms are providing. AI algorithms can only attain the same ‘kudos’ as a face-to-face consultation if they are benchmarked and proven. This does not mean they are proven to be as diagnostically good as a doctor. It requires them to understand an individual’s personal risk as well as a doctor and to describe next steps that either lead to easement of the symptoms or further investigation.

# 5 Existing work

Your.MD was first to launch an AI chat-based triage service and operates globally performing over 500,000 consultations a month in English. We have a proprietary algorithm that has been developed from scratch over the last four years by our in-house data science and medical team.

The system uses Natural Language Processing, a Chatbot, Neural Networks, Deep Learning, K Means Clustering, Nearest Neighbours and Bayesian Networks to create an experience aligned to a medical consultation.

Your.MD’s mixed approach examines free text and determines if a question is being asked about a condition or if symptoms are being inputted. Questions are answered by referring the user to medical content provided by the UK’s National Health Service or by reference to clinically validated questionnaires.

Symptoms are analysed using a chatbot AI. A series of questions are generated by the artificial intelligence to determine the most probable condition and appropriate triage response. Typically the AI takes three to seven questions to make a determination.

Your.MD then generates a consultation report with triage advice, most probable condition (if appropriate) and relevant information. Where appropriate the user can also be signposted to products and services that may assist them.

This differs from current linear decision tree approaches and triage using forms based on existing conditions.

To date we have replicated Semigran et al vignettes study to test our algorithms and we use continuous Monte Carlo testing but are only now starting work on a definitive benchmarking approach.

# 6 Feasibility

The technology exists to deliver this project. The main challenge is the design of the study and the need to **emphasise risk assessment rather than diagnostic accuracy**. Clinical trials will also need to be designed to account for patient bias and clinical feedback. Patient bias will occur depending on where and when patients are asked to interact with Your.MD. Our objective is to design a trial which captures users before they have made an appointment to see a doctor/before they have made a decision to see a doctor.

# 7 Data availability

Your.MD has well annotated data covering millions of consultations. Your.MD has also developed a proprietary medical model linking hundreds of conditions to thousands of symptoms and influencing factors. All the data is pseudonymised and covers location, gender, age, positive symptoms, negative symptoms, influencing factors, probable condition. We see no legal obstacles with sharing this data but it would not be relevant for the benchmarking exercise.

The proprietary data-set that powers Your.MD’s medical brain contains symptoms, global prevalence, personal factors and their relationship to conditions mapped to numerical probabilities. Your.MD would not make this available to other AI developers and we would not be able to develop a training module.

# 8 Data quality

The data collected from users and the Your.MD’s medical brain are both high-quality well structured sources. All input is tagged against UMLS codes whenever a medical term is found.

# 9 Annotation/label quality

Automatic annotation based on our own platform (Mediterm™) maps terms and colloquialism to medical codes (UMLS). Mediterm is one of the most advanced medical terminology tagging systems in the world.

# 10 Data provenance

The data for Your.MD’s medical brain - wherever possible - is sourced from peer reviewed medical papers. Where a recognised paper does not exist we take soundings from a network of 30 clinicians. Your.MD is also now instituting an independent audit of its data. The British Medical Journal’s Best Practice team has carried out a pilot to determine if its team can verify Your.MD’s data sources and alert Your.MD to new sources.

# 11 Benchmarking

Your.MD, in connection with Leeds University’s Department of eHealth, is working on the creation of a synthetic clinical scenario testing which is the second step of Fraser et al.’s suggested approach for staged evaluation - mentioned in section 2 above.

Its use is important for two reasons: (i) it allows fair comparison between different systems (ii) it allows testing in a pre-clinical setting in which patients are not at any risk.

Scenario tests have been conducted multiple times previously (Semigran, Babylon, etc). This ‘pre-deployment’ is intended to be the definitive version of this type of study. It draws on information from previous studies and addresses publicly-known concerns.

The aim is to compare the symptom checker output to human clinical consensus, taking human consensus as the gold standard. The weakness of this approach is that it does not easily allow for above-human levels of diagnosis/triage. The approach can be extended, by replacing the ‘Symptom Checker’ with a human clinician (as has been done previously). However, the value of this is questionable, given that a human triage process does not solely rely on questions about symptoms (i.e. this process is a hamstrung version of a real human triage process, which may take into account patient history, how they look, etc.)

The secondary aim is to provide a comprehensive methodology to describe how this sort of testing may be done in-house. This would allow future comparison between Symptom Checker providers.

The benchmarking as a minimum would measure the following outcomes:

* Diagnostic accuracy
* Triage accuracy (what should I do)
* Advice accuracy
* Attrition rate

# 12 Organizer

Your.MD has 4 years’ experience in the field, a team of 40 people, won the UNESCO/NetExplo Award, Seal Excellent Horizon 2020. We have no previous experience of running clinical trials but we are finalising a partnership with Leeds University’s eHealth Department.

Reference

[1] Self Care: N. Pillay: The economic burden of minor ailments on the national health service in the UK (2010)

[2] GP Online <https://www.gponline.com/gp-appointments-data-reveal-year-on-year-rise-workload/article/1523213>

\_\_\_\_\_\_\_\_\_\_\_\_\_