A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis

Salman Razzaki*, Adam Baker*, Yura Perov*, Katherine Middleton*, Janie Baxter*, Daniel Mullarkey*, Davinder Sangar*, Michael Taliercio*, Mobasher Butt*, Arnold DoRosario‡, Megan Mahoney§ and Saurabh Johri*

* Babylon Health ‡ Northeast Medical Group, Yale New Haven Health § Division of Primary Care and Population Health, School of Medicine, Stanford University

Introduction & Background. AI virtual assistants have significant potential to alleviate the pressure on overly burdened healthcare systems by enabling patients to self-assess their symptoms and to seek further care when appropriate. For these systems to make a meaningful contribution to healthcare globally, they must be trusted by patients and healthcare professionals alike, and service the needs of patients in diverse regions and segments of the population. We developed an AI virtual assistant based on a probabilistic graphical model (PGM) and demonstrate that it is able to provide patients with triage and diagnostic information with a level of clinical accuracy and safety comparable to that of human doctors. Importantly, this evaluation assessed the accuracy and safety of both the AI and human doctors and, unlike previous studies, also takes in account the information gathering process of both agents [1, 2]. Through this approach, we hope to build trust in AI-powered systems by directly comparing their performance to human doctors, who do not always agree with each other on the cause of patients' symptoms or the most appropriate triage recommendation. Crucially, the system is based on a generative model, which allows for relatively straight-forward reparameterization to reflect local disease burden in diverse regions and population segments. This is an appealing property, particularly when considering the potential of AI virtual assistants to improve the provision of healthcare on a global scale.

Methods. The core of our AI system is a PGM [3], designed to provide users with triage advice and to suggest likely conditions. The structure of the graphical model is defined by medical experts and is parameterized through a combination of epidemiological data and expert elicitation. Given a set of user-entered presenting symptoms and risk-factors, the model infers the most likely conditions and generates follow-up questions [4, 5, 6, 7]. The decision-making functionality of the system is provided by extending the underlying generative model with a utility model that operates as a function of disease posteriors, and is designed to provide triage recommendations that minimize the expected harm to the patient, whilst also penalizing over-triaging.

Experiments. We compared the accuracy and safety of the AI against that of human doctors. We assessed the relevance of the suggested conditions, and the appropriateness and safety of the recommended triage action. The evaluation was performed using a semi-naturalistic role-play scenario that involved mock consultations between a patient and either a human doctor or the AI, based on realistic clinical vignettes. The roles of both doctors and patients were played by practicing primary care physicians.

One hundred clinical vignettes were created by independent medical practitioners who were not involved in the role-play experiment. Each vignette was designed to simulate a medical condition from the list of all conditions currently modeled by the AI[1]. The vignettes contained information about

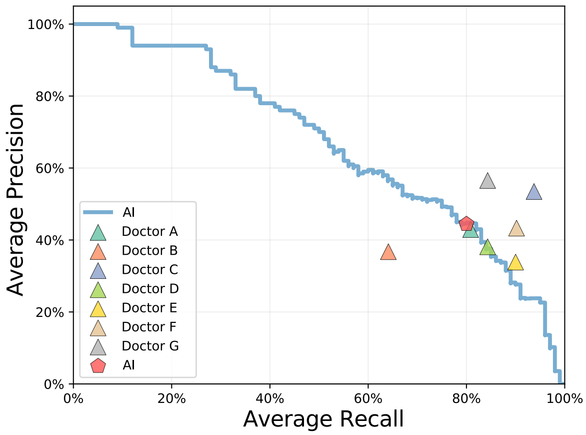

Figure 1. Average recall and precision for doctors and for the AI for different threshold parameters. Varying the internal thresholds allows the model to behave more similarly to different individual doctors, while maintaining a high level of performance, suggesting that it is not overly tuned to a particular operating point.

[1] The list of conditions modeled by the AI includes the majority of those encountered in General Practice in the United Kingdom, but does not include skin conditions, pregnancy-related conditions or paediatric conditions.

the patient, their initial complaint(s), information about their symptoms and past medical history that should be offered on open questioning, and information that should only be reported on direct questioning.

The study was conducted in multiple rounds. In each round, a “patient” was assigned a vignette and had independent consultations with up to four doctors and the AI. After each consultation the differential diagnosis (list of possible conditions) and recommended triage produced by the doctor or the AI was recorded.

Results. We assessed the precision and recall of the AI and doctors against the condition modeled by the vignette. Recall is the proportion of relevant diseases that are included in the differential. When considering only the single disease modeled by the vignette, this corresponds to the proportion of differentials that contained the modeled disease, over all vignettes. Precision is the proportion of the diseases in the differential that are relevant, and penalizes long differentials (that would result in a higher recall). A precision of 100% would be achieved if the differential diagnosis contained only the disease modeled by the vignette. In general this level of certainty is unlikely and even undesirable, given only the information provided on the vignette (i.e. in the absence of diagnostic tests).

In this study, the AI was able to predicted the modeled disease, with precision and recall comparable to human doctors and in some cases exceeding human-level performance (see Table 1 and Figure 1). Average doctor recall was found to be 83.9% (64.1–93.8%), meaning that doctors failed to include the disease modeled by the vignette in their differentials in 16% of cases.

To assess triage accuracy, an independent expert judge was asked to specify a range of safe and appropriate triage outcomes for each vignette. Providing a range of acceptable triage recommendations was motivated by the fact that doctors often disagree on the most appropriate triage recommendation, however it is not necessarily the case that any of these different opinions are inappropriate or unsafe [8]. By providing the minimum and maximum appropriate triage, the judge indicates the range of recommendations that are neither unsafe nor overly cautious. We compared the triage recommendations of doctors and the AI against the judge’s “gold standard” range. We define a “safe” triage as any recommendation which was of equal or greater urgency than the judge’s minimum triage, and an “appropriate” triage as any recommendation that fell within the judge’s range of acceptable recommendations. In this study, we found that the AI provided a safer triage recommendation than doctors on average (97.0% versus 93.1%), at the expense of a marginally lower appropriateness (90.0% versus 90.5%; see Table 1).

Table 1 Average diagnostic and triage performance for the AI and 7 doctors (with ranges). A total of 100 vignettes was assessed by the AI. Each doctor assessed a subset of vignettes (from 47 to 78).

	Recall (%)	Precision (%)	Safety (%)	Appropriateness (%)
Doctors	**83.9** [64.1−93.8]	43.6 [33.9−56.5]	93.1 [88.2−100.0]	**90.5** [85.7−94.1]
AI	80.0	**44.4**	**97.0**	90.0

Conclusion. AI virtual assistants have the potential to reduce the burden on healthcare systems worldwide, particularly when underpinned by flexible models such as Bayesian generative models that are relatively straightforward to adapt to different regions or population segments. Such systems may hold the promise of reduced costs and improved access to healthcare worldwide, but realising this requires greater levels of confidence from the medical community and the wider public. Key to this confidence is a better understanding of the relative strengths and weaknesses of both AI virtual assistants and human doctors. We have shown that a generative model based AI virtual assistant is able to diagnose and triage with a degree of accuracy and safety comparable to that of human doctors, when evaluated using simulated consultations. Further studies using larger, real-world cohorts will be required to demonstrate the relative performance of these systems to human doctors.

References

[1] Hannah L Semigran, Jeffrey A Linder, Courtney Gidengil, and Ateev Mehrotra. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ*, 351:h3480, 2015.

[2] Hannah L Semigran, David M Levine, Shantanu Nundy, and Ateev Mehrotra. Comparison of physician and computer diagnostic accuracy. *JAMA internal medicine*, 176(12):1860–1861, 2016.

[3] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[4] Laura Douglas, Iliyan Zarov, Konstantinos Gourgoulias, Chris Lucas, Chris Hart, Adam Baker, Maneesh Sahani, Yura Perov, and Saurabh Johri. A universal marginalizer for amortized inference in generative models. *arXiv preprint arXiv:1711.00695*, 2017.

[5] Jian Cheng and Marek J. Druzdzel. AIS-BN: An adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. *Journal of Artificial Intelligence Research*, 2000.

[6] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[7] Shixiang Gu, Zoubin Ghahramani, and Richard E Turner. Neural adaptive sequential Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 2629–2637, 2015.

[8] Alicia O' Cathain, Elizabeth Webber, Jon Nicholl, James Munro, and Emma Knowles. NHS Direct: consistency of triage outcomes. *Emergency Medicine Journal*, 20(3):289–292, 2003.