Estudo piloto de validação de um chatbot de rastreamento, implementado para direcionar a teleassistência em COVID-19

(*Pilot validation of a frontline chatbot to face COVID-19 using telehealth assistance*)

Gabriel F. Cateb[1], Samuel Amaral[1], Samuel C. L. Gonçalves[1], Isaias J. R Oliveira[2], Raquel O. Prates[3], Bruno A. Chagas,[3] Milena S. Marcolino[1], Zilma S. N. Reis[1,2]

[1]Faculdade de Medicina e Centro de Telessaúde do Hospital das Clínicas da UFMG – Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

[2]Centro de Informática em Saúde da Faculdade de Medicina – UFMG – Belo Horizonte, MG – Brasil

[3] Instituto de Ciências Exatas – UFMG – Belo Horizonte, MG – Brasil

gabrielfuly@ufmg.br,samuel.rsamaral@gmail.com,samuel98@ufmg.br,isaias@medicina.ufmg.br,milenamarc@gmail.com,zilma@ufmg.br, rprates@dcc.ufmg.br, bruno.azevedo.chagas@gmail.com

Abstract. *The novel coronavirus pandemic has overloaded healthcare systems to the limit. Our aim was to assess the effectiveness of a chatbot to identify symptoms of COVID-19. The chatbot was developed to screen patients before teleconsultation. The symptoms informed in the dialogue were compared with those reported to the doctors in an emergency service. Among 96 patients assessed, dyspnea was the most frequent symptom (16,6%), and the only one that showed moderate agreement with the medical history recorded in electronic medical records (Kappa=0.605). In conclusion, the technology was useful in detecting one of the major symptoms of COVID-19. However, it was not possible to evidence its effectiveness to assess minor symptoms.*

Resumo. *A pandemia do novo coronavírus tem sobrecarregado os sistemas de saúde ao limite da capacidade de atendimento. Nosso objetivo foi avaliar a eficácia de um chatbot desenvolvido para triagem de pacientes, antes de teleconsulta, para identificar sintomas de COVID-19. Sintomas informados no diálogo foram comparados com os relatados aos médicos, em um serviço de urgência. Em 96 pacientes, dispneia foi o sintoma mais frequente (16,6%) e o único que mostrou concordância moderada com a história registrada em prontuário eletrônico (Kappa=0,605). Concluindo, a tecnologia mostrou-se útil para detectar um dos sintomas graves da COVID-19, mas não foi possível evidenciar sua eficácia em relação aos sintomas menores.*

1. Introdução

A pandemia da COVID-19 afetou 125 milhões de pessoas até março de 2021, sendo responsável por 2,7 milhões de óbitos e uma enorme sobrecarga para os sistemas de saúde em todo o mundo [WHO 2016]. A implementação de tecnologias digitais, entre elas os chatbots, foram introduzidas em sistemas de triagem preparados para selecionar casos

mais graves através de diálogos pré-estabelecidos com base em evidência científica [Judson et al. 2020]. O uso dessas tecnologias, ao permitir avaliar pessoas à distância, foi impulsionado, uma vez que o distanciamento social é uma das medidas mais efetivas na prevenção contra o vírus [WHO 2016]. Uma das potenciais utilidades dos bots é conseguir filtrar os pacientes que realmente precisam de atendimento presencial [Judson et al. 2020]. Estudos de implementações de chatbots neste contexto na população brasileira são escassos. Este estudo piloto buscou avaliar a eficácia e a viabilidade de um chatbot para identificar sintomas relacionados à infecção pelo COVID-19, comparando as respostas com as queixas relatadas ao médico, em consulta presencial.

2. Metodologia

Trata-se de estudo piloto de validação de uma nova tecnologia de saúde digital [WHO 2016]. O delineamento epidemiológico foi do tipo coorte. O cenário do cuidado foi o Hospital das Clínicas da UFMG/Ebserh, um hospital universitário terciário de referência para o tratamento da COVID-19, no período de 01/10/2020 a 15/01/2021. Os pacientes foram abordados no serviço de pronto atendimento, enquanto aguardavam a consulta médica e concordaram com a participação voluntária no estudo, assinado um termo de consentimento livre e esclarecido. O projeto foi aprovado pelo Comitê de Ética da UFMG sob o número CAAE 35953620.9.0000.5149, base legal para uso de dados identificados.

2.1. O desenvolvimento da tecnologia

Um grupo multidisciplinar de especialistas das ciências da saúde e da computação da UFMG desenvolveu um aplicativo chatbot especialista da empresa TAKE BLIP 2021. O acesso foi disponibilizado via aplicativo de mensagem instantânea (WhatsApp) ou pelo site do Centro de Telessaúde da Hospital das Clínicas da UFMG/Ebserh (https://telessaude.hc.ufmg.br/).

O chatbot foi programado para executar um algoritmo que aponta dois caminhos de cuidado: (i) avaliação do estado de saúde; ou (ii) tirar dúvidas sobre a COVID-19. Uma equipe de médicos pesquisadores e especialistas em infectologia fundamentou o conteúdo científico utilizado no *software*, mantendo-o sob contínua atualização [WHO 2021]. Na opção de avaliação do estado de saúde, o sistema automatizou uma árvore de decisão baseada em evidências e inteligência artificial para triar pacientes com sintomas de COVID-19, segundo manifestações clínicas e comorbidades. A estratégia foi sugerir aos usuários com quadros leves a permanecerem em seus domicílios com tratamento sintomático, separando-os daqueles com sinais de alerta e/ou comorbidades que se enquadrem nos grupos de risco. As perguntas foram escalonadas por prioridade, de forma que o sintoma mais grave foi o primeiro a ser perguntado, seguido de outros sintomas, em ordem decrescente de gravidade. No caso de resposta positiva às três primeiras perguntas, o chatbot direciona o caso para o teleatendimento e os sintomas de menor gravidade não são perguntados. Na opção para tirar dúvidas, o sistema realiza orientações sobre a COVID-19, agrupadas por temas e no formato de perguntas-respostas. Este caminho não foi alvo do presente estudo.

2.2 Procedimentos do estudo de validação

Uma amostra por conveniência foi selecionada por critérios de elegibilidade para abordagem presencial: idade ≥18 anos; estar no serviço de pronto-atendimento clínico, incluindo a maternidade, com alguma demanda de saúde; estar aguardando consulta

médica de urgência ou exame complementar, sem ciência de seu diagnóstico definitivo para a demanda de urgência. O cálculo amostral teve como base a elevada ocorrência esperada de sintomas em serviços de urgência. Por questões de segurança do paciente e dos pesquisadores de iniciação científica, casos suspeitos de COVID-19 não foram avaliados.

Acadêmicos de Medicina foram treinados para realizar a abordagem aos voluntários da pesquisa presencialmente, mesmo que o chatbot em sua versão de produção seja totalmente em ambiente virtual. Os pacientes foram abordados, pelos acadêmicos treinados, em um momento que já tinham o risco de gravidade geral da urgência classificado pela escala de Manchester [Storm-Versloot 2011], mas ainda sem avaliação médica. Os pacientes responderam questionário sobre experiência prévia com tecnologias digitais. Em seguida, utilizaram o chatbot em um tablet com uma breve explicação dos pesquisadores, sem qualquer interferência destes. Problemas de conexão da rede internet ou dificuldades encontradas na utilização do tablet ou navegação nas interfaces do chatbot foram registradas.

Posteriormente, de forma cega ao conhecimento das respostas do chatbot, o registro clínico realizado na consulta médica presencial foi analisado pelos acadêmicos. Dados da consulta foram extraídos nas entradas padronizadas do prontuário eletrônico do paciente (PEP): "queixa principal" e "história clínica". Os mesmos sintomas perguntados no chatbot foram pesquisados: dificuldade de respirar (dispneia), hipotensão (síncope ou pré-síncope), febre, sintomas menores (tosse, expectoração, odinofagia, disfagia, cefaleia, congestão nasal, anosmia ou ageusia). A ausência de menção a um determinado sintoma foi considerada ausência da queixa. Além disto, dados como idade, sexo e classificação de risco de Manchester atribuídos no processo de atendimento da urgência foram recuperados.

As queixas diretamente informadas pelos pacientes no chatbot foram confrontadas com aquelas extraídas pelos pesquisadores no PEP. A validação do diálogo do chatbot na detecção das queixas de interesse foi obtida através de testes de confiabilidade. O índice de concordância Kappa comparou as queixas obtidas pelas duas técnicas. A acurácia do diálogo do chatbot em reconhecer os sintomas foi calculada em termos de sensibilidade e especificidade. O nível de significância considerado nos testes de hipóteses foi $p < 0,05$.

3. Resultados

No período do estudo, 130 pacientes foram entrevistados (Figura 1).

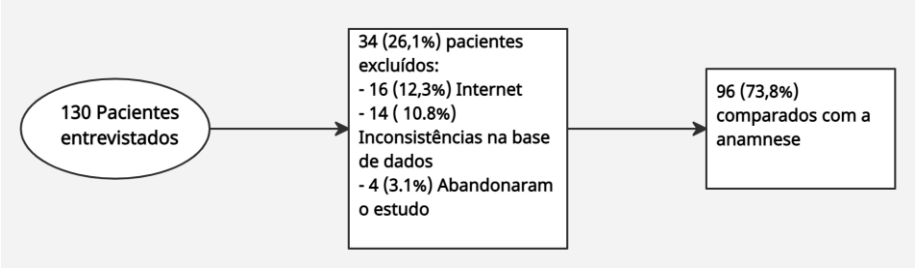

Figura 1. Fluxograma da coorte

Entre os 96 pacientes com registros válidos, a idade mediana foi 35.0 ± 24, variando de 18 a 77 anos, 77,6% eram mulheres, 23 (23,9%) marcaram algum dos sintomas no chatbot. Segundo a escala de Manchester, a maioria dos pacientes foi

classificada como pouco urgentes ou não urgentes 44 (45.8%), 39 (40.6%) urgentes, 13 (13.5%) muito urgentes e nenhum estava em situação de emergência. A experiência prévia com tecnologia móvel foi diversa, mas 97,8% dos participantes do estudo já utilizavam pelo menos um aplicativo (Tabela 1).

Tabela 1. Características dos participantes do estudo

Experiência em tecnologia móvel	Estatística
Possui um smartphone, n (%)	89 (90,8%)
Possui internet no smartphone, n (%)	83/89 (93,3%)
Aplicativo mais usado: WhatsApp, n (%)	87/89 (97,8%)
Usa o smartphone ≥ 4 hours ao dia, n (%)	52/96 (54,2%)
Checa o smartphone a cada 30 min, n (%)	20 (20,4%)

Dispneia foi o sintoma mais frequente no diálogo com o chatbot 16 (16,7%), e apresentou concordância significativa e moderada quando comparados os métodos de registro (Tabela 2). Em relação à acurácia do chatbot na detecção dos sintomas em relação à anamnese clínica, a dispneia foi detectada em 71,4% dos pacientes. A ausência deste sintoma foi detectada em 92,7% no chatbot, em relação à consulta médica. Hipotensão, febre e sintomas menores se apresentaram em baixa frequência no chatbot, 4 (4,2%), 1 (1%) e 2 (2,1%) respectivamente. Mas 13 pacientes relataram sintomas menores para o médico. Não houve concordância significativa nas demais comparações.

Tabela 2. Confiabilidade do diálogo do chatbot em obter as queixas dos usuários em comparação com os registros médicos

	Chatbot	Registro médico	Kappa	Sensibilidade	Especificidade
Dispneia	16	14	0.605 (p<0.001)	71,4%	92,7%
Hipotensão	4	3	-0.039 (p=0.727)	0%	96,1%
Febre	1	1	-0.013 (p=0.908)	0%	98,4%
Sintomas menores*	2	13	0.091 (p=0.216)	7,7%	98,4%

* Tosse, catarro, dor de garganta, dificuldade para engolir, dor de cabeça, nariz entupido ou escorrendo, parou de sentir cheiro ou gosto

4. Discussão

O estudo mostrou uma boa concordância, sensibilidade e alta especificidade do chatbot em relação à queixa de dispnéia, um dos principais critérios de gravidade na infecção pelo SARS-CoV2. No entanto, não foi útil para mostrar a acurácia em relação aos outros sintomas. Há que se considerar que o diálogo é interrompido quando um sintoma grave, como a dispneia, hipotensão ou febre são relatados, nesta sequência. Com isso, não foi possível verificar a coexistência dos sintomas menores e comorbidades seguintes a um deles.

Em relação à familiaridade com tecnologia digital, mais de 90% dos participantes tinham um telefone celular próprio, acesso à internet e usavam aplicativos. Sendo assim, é possível supor que o acesso a este tipo de tecnologia não tenha sido um impedimento

para os usuários informarem seus sintomas. No entanto, não se descarta que o baixo letramento em saúde da população brasileira possa ter influenciado os resultados, assim como a própria linguagem empregada nos diálogos. Este projeto possui apoio de especialistas em linguagem humana, o que desencadeará uma revisão nas perguntas com baixa acurácia, na expectativa de prover melhorias na interação com os usuários. Apesar de iniciativas semelhantes terem sido desenvolvidas por outros grupos no país, este estudo piloto foi pioneiro no cenário nacional.

Para comparação, iniciativas internacionais de maior sucesso já foram descritas, porém com metodologia mais complexa. O chamado "*Symptoma*", chatbot utilizado na Alemanha, Grécia e Reino Unido, alcançou mais de 20.000 usuários e mais de 90% de acurácia no diagnóstico de COVID-19 [Martin,Alistrair et al, 2020]. Um chatbot desenvolvido pela Universidade da Califórnia (EUA), mostrou também grande eficiência na triagem de COVID-19 em trabalhadores da saúde, reduzindo o tempo do procedimento de entrada dos profissionais nos hospitais [Judson et al. 2020].

Quanto ao processo de desenvolvimento do chatbot, um estudo em fase piloto busca iniciar uma solução tecnológica para um problema de interesse. Com a análise dos resultados, as vantagens e limitações são levantadas. Por se tratar de uma abordagem recente em saúde, o seu processo de validação também o é. Estudo prévio destaca a importância de se testar os chatbots confrontando-os com a consulta presencial ou à distância, em relação às técnicas de simulação. Os principais problemas relatados quando validados fora do contexto são a pouca profundidade das conversas não reais [Milnes-Ives et al,2020]. No presente estudo, a escolha de uma validação presencial e prospectiva, trouxe como vantagem a coleta de dados qualificados, respeitou as prioridades dos pacientes na consulta médica de urgência, direcionada para a sua demanda individual. Talvez por este motivo, 13 pessoas tenham relatado sintomas menores para o médico e apenas 2 no *chatbot*.

A maior limitação deste estudo tem relação com o pequeno número de pacientes com sintomas respiratórios e gripais na amostra de usuários selecionada, dificultando a validação de outros sintomas relacionados à COVID-19, além da dispneia. Mesmo assim, em pacientes assintomáticos, o chatbot é útil na educação em saúde, abordagem também muito relevante, em relação aos cuidados durante a pandemia e no combate às *fake news,* não abordados nesta validação. No contexto da urgência, a participação voluntária pode ser prejudicada por sintomas álgicos, desconforto do paciente ou necessidade de exames complementares, o que também foi um obstáculo na coleta dos dados. Diante do crescimento exponencial da saúde digital e da inteligência artificial dentro da prática médica, mais estudos sobre o tema são necessários. Em análise subsequente, pretende-se incluir pacientes com suspeita de COVID-19, com amostragem maior para explorar possibilidades e aprimoramentos na aplicação.

5. Conclusão

Este estudo piloto mostrou que é viável a aplicação e a difusão do chatbot para reconhecimento de sintomas graves de COVID-19.

Agradecimentos

Agradecemos ao Hospital das Clínicas e Reitoria da UFMG, pela viabilização do estudo e recursos: 4/2020/CGPO/DIFES/SESU-MEC, CAPES (88887.507149/2020-00).

Referências

World Health Organization. Coronavirus disease (COVID-19) pandemic. Acessível em: https://www.who.int/emergencies/diseases/novel-coronavirus-2019?adgroupsurvey={adgroupsurvey}&gclid=Cj0KCQjwjPaCBhDkARIsAISZN7Re8Ln2XfDSjoAiN3jnNzGVSHCApPNsY8EDe8f6Pw7uA7t0OaNzHRYaAsBkEALw_wcB. Data do acesso: 26/03/2021.

JUDSON, Timothy J. et al. Implementation of a digital chatbot to screen health system employees during the COVID-19 pandemic. Journal of the American Medical Informatics Association, v. 27, n. 9, p. 1450-1455, 2020.

Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. Geneva: World Health Organization; 2016. Licence: CC BY-NC-SA 3.0 IGO.

Storm-Versloot MN, Ubbink DT, Kappelhof J, et al. Comparison of an informally structured triage system, the emergency severity index, and the Manchester triage system to distinguish patient priority in the emergency department. Acad. Emerg. Med. 2011; 18(8):822-829.

WILLIAMS, Noelle L. et al. Clinical Integration of Digital Solutions in Health Care: An Overview of the Current Landscape of Digital Technologies in Cancer Care. American Society of Clinical Oncology, 2018.

MARTIN,Alistrair et al. An artificial intelligence-based first-line defence against COVID-19: digitally screening citizens for risks via a chatbot, 2020**.**

GABARRON, Elia et al. What do we know about the use of chatbots for public health? European Federation for Medical Informatics, 2020.

MILNES-IVES et al. The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review .Journal of Medical Internet Research,vol. 22, e20346,p.1, 2020.