When Penn Medicine decided to launch a chatbot to help patients with questions about Covid-19, leaders recognized that the kind of information patients were seeking (based on call center interactions) often requires an institution-specific response. Even standardized responses to symptom checker questions need to be crafted in a manner consistent with the local capabilities and pathways of the health system in which the initiative is based. The need for unique response mapping, complex contextualization, and dynamic, human-guided validation of content led leaders to develop a specialized chatbot through collaboration with companies experienced in machine learning and natural language processing. Two weeks after the initial planning meeting with the companies, the chatbot went live.

Throughout the U.S., people are struggling with questions, lots of questions, about Covid-19. Seeking answers, they turn to whatever resources are available: the Internet, friends and family, or they try to reach their clinicians for guidance. Medical call centers and clinicians bear the brunt of this volume burden as patients seek reliable answers to questions about Covid-19, the risks for them and their families, and the disruptions it introduces to their usual health maintenance.

Many of the questions patients have about Covid-19 are questions that other patients have as well, like What are symptoms of Covid-19? or What should I do if I fear I have been exposed? or Where can I get tested? While each patient has a unique medical history, many of these questions lend themselves to standardized answers that should reflect the most up-to-date knowledge and guidelines about the virus and be readily accessible for patients.
Across a wide range of industries, chatbots have been an efficiency-enhancing way for business teams to interact with their customers. Chatbots are conversational agents that leverage machine learning and natural language processing to understand intent in order to reply with appropriate answers, and they have advantages relevant to the present crisis.\textsuperscript{1-3} First, they are accessible any time, allowing patients to obtain answers 24/7 and to avoid wait times on hold before reaching a human. Second, chatbots have a higher handling capacity than any human.\textsuperscript{4} A single chat bot can simultaneously have conversations with thousands of people no matter what time of day. Regardless of the volume of calls, every question may be answered immediately.

Less obvious benefits may be the fact that chatbots offer the potential to standardize answers, consistency that can be difficult to achieve in the context of significant numbers of people rotating through phone line staffing assignments. Users seek quick, accurate responses when searching for information or assistance, and this productivity has been demonstrated to be one of the most important reasons organizations deploy chatbots.\textsuperscript{5} While chatbots require ongoing training to achieve accuracy, the ability to learn as they go, incorporating shifts in conversation structure and flow, can help address the fluidity and variability of language used in describing health related needs, in both normal and dynamic circumstances.\textsuperscript{6} Bots can also interact in many different languages. Of course, bots are still machines where the \textit{I in AI} is not yet equal to human cognitive abilities, and thus work in this arena requires deliberate and transparent validation methods, finesse with disclaimers, and institutional tolerance for imperfection and post-release improvement.

Chatbot diversity is derived from frameworks that vary based on type, direction, guidance, interaction style, and communication channel.\textsuperscript{7} Despite this diversity of function, most chatbots perform mundane tasks in a predictable fashion, with less evidence on how they might handle the level of variation and complexity in medical questions and responses arising in the time of a pandemic.\textsuperscript{8} In this study, we outline the creation of a Penn Medicine chatbot collaboratively created with Verily, Google Cloud, and Quantiphi, a Google Cloud strategic partner. (Penn Medicine includes the Perelman School of Medicine and six hospitals and hundreds of outpatient centers throughout Southeastern and Southern New Jersey.)

When we chose this direction, a number of organizations had already deployed off-the-shelf chatbots. We recognized the potential value of these tools, as many questions — like Should I wear a mask? — should produce consistent answers. Our analysis of incoming patient questions, though, revealed the need to address institution-specific queries, such as Where should I go to get tested? or How do I get my test results? Even interactions that might logically be standardized, such as symptom checkers, must be crafted in a manner consistent with the capacity, capabilities, and pathways of the health system using it to translate information into desired patient actions while responsibly managing constrained resources. This need for unique response mapping, complex contextualization, and dynamic, clinician-guided validation of content required us to develop a specialized chatbot (Figure 1).
FIGURE 1

Chatbot Answers Questions About Covid-19

In this exchange, the patient has agreed to use the automated system and is asking some general questions about the coronavirus.

Answers to Your COVID-19 Questions

Great! To ask a question, type it into the box below.

What is coronavirus?

The World Health Organization states that Coronaviruses (CoV) are a large family of viruses that cause illness ranging from the common cold to a more severe lung infection. COVID-19 is the disease caused by a novel coronavirus.

Will it go away in the summer?

At this time, it is not known whether the spread of COVID-19 will decrease when weather becomes warmer. There is much more to learn about the transmissibility, severity, and other features associated with COVID-19 and investigations are ongoing.

Ask something... Feedback

Source: Penn Medicine
NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

Process of Creating the FAQ Bot

The process of creating the FAQ bot was initiated by first determining the most frequently asked questions to assess which content would best offload volume from care team members. Using a database of Covid-19-related telephone encounters with the Penn Medicine phone lines and patient-submitted secure messages to our patient portal, a team of medical student volunteers analyzed a sample of 800 telephone encounters and 450 secure messages from the early stage of the epidemic in Philadelphia region (3/19/20–3/22/20) and categorized them by topic, sub-topic, and main question. These questions were then re-ordered by frequency and further sorted into general categories of Covid-19-related/Infectious Diseases–specific questions (e.g., How long am I infectious for? What do I do if I have a fever?), Oncology/Immunosuppressed–specific questions (e.g., Am I at higher risk on immunosuppressants? Should I discontinue my immunosuppressants?),
Occupational Health–specific questions (e.g., “I had an exposure; should I go into work?”), and Logistics/Operational–specific questions (e.g., “How do I get tested?”). By this process, we generated an initial list of 97 high-frequency questions. Beyond the frequency of questions, it was important to leverage patient-submitted messages to understand the phrasing patients used in making inquiries. This language became a critical element in training the bot, providing a foundation for the technology to begin learning how to link various phrasings for a question to the underlying intent and, therefore, an appropriate answer.

The next step in the content process was formulating answers to these questions. With a team of medical students, information from PubMed-searched studies, recommendations from the Centers for Disease Control and Prevention and the World Health Organization, reputable news outlets, and Penn clinicians, we crafted answers to the extracted questions.

Two of the key challenges to overcome were variation in information across a large, complex health system and rapidly changing guidance due to the reality of both continually evolving evidence and changing conditions, such as where tests or services were available.

This content then required validation and the creation of a system for ongoing answer verification. Two of the key challenges to overcome were variation in information across a large, complex health system and rapidly changing guidance due to the reality of both continually evolving evidence and changing conditions, such as where tests or services were available. As new information became available and was assessed and contextualized for our system, our team had to insert ourselves into existing information distribution channels to avoid creating failure points where content could fall out of sync. Translating the information into clear answers for a consumer audience — and applying what was known to a broad range of contexts of interest to patients, such as whether aerosols staying suspended in the air mean it’s not safe to go to the grocery store or to stand where an infected person might have coughed hours ago — was also important. We created a content team with fellows and attendings from infectious disease, oncology, occupational health, primary care, and operations leaders along with a deliberate process for drafting and approving answers.

The attempt to ensure consistency in providing rules-based, algorithmic answers revealed opportunities for policy alignment across our hospitals, from whether a spouse could go in with a patient having surgery to occupational health rules for our employees. Close engagement by our operations and clinical leadership teams enabled rapid resolution of discrepancies. Where differences were appropriate, we had to identify questions in the bot requiring divergent answers, triggering follow-up questions for clarity and branching once the person provided more information, such as whether they were in Lancaster or Philadelphia or were a Penn Medicine employee or not.

We identified two critical contexts and audiences who needed help securing fast, accurate answers to coronavirus related questions: nurses on our phone lines seeking support when addressing patient questions and patients interested in self-serve answers via the chatbot. The needs of these
groups clearly differ. Nurses preferred to quickly access all the information on a topic, consider alternative answers, and select the response they found most appropriate. Thus, we created an internal search tool that would search the FAQ database and display all relevant question-answer pairs ranked by relevance. This search interface eliminated the time and effort previously required to scroll through many pages hunting for relevant content. We incorporated input from the Penn Medicine Patient and Family Advisory Council in the bot testing and feedback process pre-launch and strove to provide a single best response to reduce ambiguity for patients.

A central challenge in enabling the desired patient experience required addressing the reality that people phrase and frame questions in seemingly endless ways. Myriad word choices can represent the same intent. For instance, Should I keep taking my Humira?; Are immunosuppressant medications safe?; and Does my lupus drug increase my risk of Covid-19? all express patients’ need to know whether one should alter their regimen of immunomodulating medications during the pandemic. Partnering with Google presented the opportunity to leverage natural language processing and machine learning in pursuit of mapping diverse inquiries to common underlying intents. The Google team offered technical acumen, resources, and their Dialogflow technology, where we could establish validated answers for high-volume questions and map varying observed language representing the intended meaning to initiate training the bot, a process that could then be amplified and accelerated by Google’s machine learning capability. For each question-answer pair, our teams collaborated to provide 10–15 training phrases consisting of alternate ways a patient might phrase a question driving at the same intent. After training the bot on these phrases, it was then able to recognize many more phrasings, driving selection of the appropriate response.

“For each question-answer pair, our teams collaborated to provide 10–15 training phrases consisting of alternate ways a patient might phrase a question driving at the same intent. After training the bot on these phrases, it was then able to recognize many more phrasings, driving selection of the appropriate response.”

Testing and Going Live: The urgency and importance of the crisis attracted support from across the Penn Medicine community, enabling a crowdsourced testing effort. In several rounds of internal testing leveraging medical students, faculty, non-clinical staff, and the patient and family advisory council, we collected more than 500 feedback submissions in three days, ranging from missing answers to clarifying edits and new ways people worded questions that the bot would need to learn. This input allowed us to quickly make improvements by altering and adding responses and teaching the bot by adding training phrases based on how testers structured inquiries. Prior to launch, we established clear and measurable criteria for going live, based on industry standards for accuracy of answers, and then tested the bot on 150 patient-submitted questions from our patient portal. In the first round of testing the bot did not perform well enough, reaching an accuracy of 61%. On review, we discovered that this was due to a small subset of questions that were not yet included in the bot’s training data. Therefore we added the relevant content and training phrases. In the subsequent testing phase, the bot produced a sufficient response 75% of the time (“sufficient” defined as
correctly answering questions to which it knew the answer and recognizing questions to which it did not and providing an appropriate fallback response), meeting our launch criteria.

While we still had to set clear expectations with health system leaders that people wouldn’t receive their desired outcome in one-quarter of interactions, the combination of ongoing manual training based on patients’ and clinicians’ incoming questions and machine learning algorithms supporting the bot would enable it to improve with use. Based on this, we decided to proceed with launching on our institution’s Covid webpage on April 9, 2020, two weeks after our initial planning meeting with our industry partners.

**Process of Creating a Triage Tool**

In addition to building a chatbot to provide accurate information to patients, we also equipped it with the ability to triage symptomatic patients to the appropriate level of care. Staying focused on the goal of offloading volume from phone lines and reducing clinician burden, analyses of incoming questions revealed that people reporting symptoms and seeking direction on what to do was quite common. Patients would be offered the triage tool when first opening the bot and whenever they submitted a question that suggests they are experiencing symptoms. Similar symptom checkers already existed on several platforms, provided by startups and larger technology vendors. However, we found we needed to both customize our approach to appropriately triage patients to the avenues for receiving care at our institution, mirroring system protocols, and have back-end control of the content and algorithm for rapid adjustments as the pandemic evolved (Figure 2).
Attaining buy-in on the topic of automated triaging of patients required extensive dialogue with clinical leaders. Several key stakeholders initially expressed reluctance to triage algorithmically without an individualized clinical assessment performed by a trained provider. Valid concerns were raised, most notably about how a bot could both understand the nuances of conversation and communicate as well as a seasoned clinician, and regarding liability in the event of inappropriately triaging patients. The concern was raised that many patients would likely prefer to speak directly to a clinician, so that option was always preserved. Still, most patients, about two-thirds of a sample of 2,400, agreed to proceed with the bot after the disclaimer message, and about 90% of those finished to an endpoint.
Rather than train a bot to understand all the vagaries of a patient describing symptoms, we constructed an algorithm that mirrors what seasoned clinicians do: distill complicated situations down to their most crucial elements and make decisions based on those factors. While any number of additional data points can be useful to a clinical provider, we chose to focus on only those questions that would influence one’s disposition decision and gave users binary answer options.

Our work started with a helpful meeting with Tim Judson, MD, MPH, from UCSF’s Clinical Innovation Center. He and Ralph Gonzales, MD, MSPH (Associate Dean for Clinical Innovation and Chief Innovation Officer for UCSF Health), and their team had implemented an algorithm for triaging Covid and influenza-like illness patients that they deployed through their EHR-linked patient portal. Using these questions and disposition endpoints as a foundation, we convened a team of clinicians at our institution to edit the questions asked, triage algorithm, and four disposition endpoints to match our patient population, priorities, and capabilities. Additionally, we restructured the order in which questions were asked to prioritize appropriately triaging patients in as few questions as possible (algorithm shown in Appendix). For instance, take a 50-year-old male with unrelenting chest pain. Eventually a provider team will want to know whether this patient has a history of hypertension, but that information will not change the fact that he needs an emergent evaluation and can be triaged straight into the emergent evaluation category, no further questions asked.

A bot cannot yet replicate the art of medicine, so we positioned the triage tool’s role as expediting the process for patients to access the level of care they needed and prioritizing our human resources for where they could have the greatest impact. The context and magnitude of the Covid-19 outbreak provided a catalyst for exploring alternate methods to triage patients, particularly those who are lower risk or the worried well. As noted above, bots aren’t limited in how many patients they can serve simultaneously, nor are they susceptible to fatigue in the case of high volume.

On the risk of inappropriately triaging patients, involving clinical leaders in the algorithm’s refinement was critical to assuage concerns. The nurse leaders from our Covid hotline also provided valuable input on clarifying and simplifying language, having a wealth of experience from live phone conversations. We intentionally made the algorithm conservative in erring toward recommending either remote or in-person evaluation as opposed to self-care at home. We also included anticipatory guidance at each disposition endpoint, encouraging patients to seek care if they remained concerned and to repeat the tool or contact their providers if their symptoms change, worsen, or fail to improve. While all stakeholders were comfortable with the triage algorithm for those placed into the highest and lowest severity groups (emergent and low risk for complications, respectively), achieving convergence about how best to handle the two middle
groups (those requiring an urgent or non-urgent evaluation) was more challenging. For these groups, initial deployment included validation of the bot’s triage decision with an intentionally redundant clinical evaluation by a triage nurse to enable comparison of the decisions made by the bot and the triage staff. We agree with our colleagues who have called for methodical validation of facilitated self-service tools to assess safety and effectiveness. We’ve structured the hand-off from the bot to our clinicians to enable a follow-up analysis comparing the bot’s determination to that of clinical judgment.

**Ongoing Process for Updating**

Providing continued accurate, consistent information regarding Covid-19 requires frequent updating, given the steady evolution of information and policies. We developed two processes for managing updating: one for new information and one for ongoing review of previous content.

For the first process, ongoing content generation occurred similarly to the initial content generation. New frequently asked questions were extracted from ongoing communications through our patient portal, news articles, CDC/WHO recommendations, and daily reports of frequently asked questions coming into Penn Medicine phone lines. Post-launch, we also used user-submitted questions to guide topics for new content generation. Medical students drafted answers, which were then reviewed by fellows and attendings. This system was organized into a twice-weekly review. Fellows and attendings would update the project manager after review and the project managers migrated verified FAQs to the production version of the Google bot. In addition to this process, there was a more immediate process for adding new operational updates. CDC and Penn Medicine–specific updates continue to be tracked in real time. The project manager then drafts and forwards a response for immediate review by the relevant reviewer, especially for expedited execution of logistics or operations content change, such as when testing or safety protocols change.

The chatbot automatically detects and flags for analysis *intents* (phrasings with similar meanings) from users that go unanswered. High-frequency unanswered questions can be analyzed in a rapidly evolving clinical situation to produce more question/answer pairs on a near real-time basis. Phrases with similar meanings can be added to an intent if local, colloquial phrases are newly identified after deployment, allowing for better answers next time.

**Success Metrics to Track**

The key questions to answer in assessing newly deployed technology typically ask whether people use it, whether they like or recommend it, and whether it moves the targeted needle to change the outcome. In this case, we added metrics related to accuracy to ensure consistently high-quality responses. We also chart accuracy weekly to enable us to measure the degree of ongoing improvement.

Usage can be assessed based on a number of within-bot metrics, including: number of users, average number of questions asked per conversation, most frequent responses given by the bot,
and the percentage of fallback responses where the bot did not recognize or know the answer to a question.

Patients’ experience with the bot can be measured by directly asking for feedback. We integrated a permanent feedback link in the user interface visible at all times to a user. We’ve also contemplated ending sessions by directly asking for feedback, which we have not yet deployed due to design considerations including not wanting to bother users at a potentially stressful time.

Call volume hitting our clinicians and wait times affecting our patients were key outcome metrics for initiating this project. As the pandemic unfolded, the health system was able to re-deploy clinicians to phone lines in a manner that handled incoming volume effectively, such that wait times never became an issue of concern. That being said, as we return staff to their original clinical roles, and in preparation for any forthcoming surge or crisis, we will continue to examine the bot’s potential to offload volume.

Accuracy remains of utmost importance in establishing credibility and earning leadership confidence for broader deployment of bot technology. We’re able to automatically track the percent of questions that do not result in an answer, and this improved steadily from 33.1% during initial testing pre-launch to 12.7% in the most recent week for which we have data (5/24–5/30/20), but that’s only part of the equation. We also need to assess how often the answer produced does not address the question intended, which requires a qualitative assessment of logs recording questions asked and answers given.

To round out evaluation, the final critical piece will be the degree of concordance between the triage decisions made by the bot and those made by clinical personnel. Other important measures include the number and percentage of patients triaged to each disposition (emergent, urgent, non-urgent, self-care at home) and tracking the outcomes for patients who call into our Covid hotline from the triage tool and are then followable in our EHR. These are important not only for research and monitoring, but also for altering the triage algorithm as needed.

**Broader Implications**

The images of overwhelmed health care systems in Italy, Spain, and New York City have highlighted the importance of leveraging novel tools to reduce the demands on front-line clinical personnel. Bots have made substantial inroads in many industries due to the efficiencies that can be achieved but have not been as widely adopted within health care settings. Much of this relates to the culture of medicine — the notion that an in-person evaluation with a clinician is the best way to diagnose and treat a patient. However, as we are seeing with as much as 90% of outpatient visits becoming virtual in some systems, there are substantial shifts underway in how we collectively view what is ideal in terms of diagnosing and treating patients. Tools such as bots can also be adopted with low incremental effort.
As we are seeing with as much as 90% of outpatient visits becoming virtual in some systems, there are substantial shifts underway in how we collectively view what is ideal in terms of diagnosing and treating patients. Tools such as bots can also be adopted with low incremental effort.

To facilitate wider-scale adoption, Verily and Google Cloud have collaborated to create an open-source Pathfinder virtual agent template, which is freely available to health systems and hospitals. This facilitates creation of chat or voice bots that answer questions about Covid-19 symptoms and provide the latest guidance from public health authorities like the Centers for Disease Control and Prevention and World Health Organization. Contact Center AI’s Rapid Response Virtual Agent program is available around the world in any of the 23 languages supported by Dialogflow.

Part of the appeal of a bot is that it is instantly available 24/7 to meet patient demands for support and poses a compelling, convenient option if in fact it gives clear, accurate answers. We have already seen the potential to standardize rigorously researched answers to eliminate concerns about variability in responses. Some of the questions bots can answer are mundane logistical questions but many of these are highly important to patients, such as the location of parking or how to get test results. Others are clinical questions where self-service customization is possible due to available technology capabilities. Deployment of bots during this crisis may be done to prevent clinical and clerical staff from being overwhelmed by inbound call volume; in time, however, health systems may also find it helpful to either serve larger populations of patients without adding staff or to reallocate staff to roles of providing clinical care best done by humans and letting bots handle the algorithmic tasks.

Bots can make information available at a scale well beyond telemedicine approaches due to automation and this can provide support for people who cannot afford care, who can’t communicate well in English, or who prefer anonymity.

Some of this may be uncomfortable. Medicine has always prided itself on being a people business where direct contact between humans was paramount. Perhaps because of this, the technology-enabled savings in costs that have characterized most other industries have largely evaded medicine, and health care costs in the U.S. have risen well above the level of affordability for most American families. Bots can make information available at a scale well beyond telemedicine approaches due to automation and this can provide support for people who cannot afford care, who can’t communicate well in English, or who prefer anonymity. These approaches can be developed to tie to institution-specific resources and that will enhance customer service by providing the best answer to patients whenever they want it 24/7 without having to wait and without variability based on who answers the phone.
Maguire Herriman,
MD/MBA student, Perelman School of Medicine, Wharton School, University of Pennsylvania

Elana Meer,
MD/MBA student, Perelman School of Medicine, Wharton School, University of Pennsylvania

Roy Rosin, MBA
Chief Innovation Officer, Penn Medicine Center for Health Care Innovation, University of Pennsylvania Health System Member, Penn Center for Health Incentives and Behavioral Economics

Vivian Lee, MD, PhD, MBA
President Health Solutions, Verily Life Sciences LLC

Vindell Washington, MD
Chief Clinical Officer, Verily Life Sciences LLC

Kevin G. Volpp, MD, PhD
Director, Penn Center for Health Incentives and Behavioral Economics; Professor, Perelman School of Medicine and the Wharton School of Business, University of Pennsylvania Chief, Division of Health Policy, Penn Medical Ethics and Health Policy Editorial Board, NEJM Catalyst Innovations in Care Delivery

Penn Medicine Triage Tool Algorithm, Current Version 5/30/20

Acknowledgments

The authors wish to acknowledge the following for their assistance: Tim Judson, MD, MPH, from UCSF’s Clinical Innovation Center, for early guidance. Team members Krisda Chaiyachati, MD; Mohan Balachandran; Glenn Fala; Tim Jones; Kevin Van Horn; David Asch; Susan Day; Phillynn Hepschmidt; Susan Regli; Ann Huffenberger; Danielle Werner; Christine Vanzandbergen; Colleen Mallozzi; Kathleen Lee; Aaron Johnson; John McGreevey; and the Google/Verily/Quantiphi team for helping us to stand this up so quickly.

Disclosures: Maguire Herriman, Elana Meer, and Roy Rosin have nothing to disclose. Vivian Lee and Vindell Washington are employees of Verily Life Sciences. Kevin Volpp is a partner and part-owner of VAL Health and has received speaking fees from the Center for Corporate Innovation and the Lehigh Valley Health System and research support from CVS, WW, Vitality/Discovery, Oscar Health Insurance, and Hawaii Medical Services Association.

References

1. Dale R. The Return of the Chatbots. Nat Lang Eng. 2016;22(6):811-7
2. Lebeuf C, Storey MA, Zagalsky A. How Software Developers Mitigate Collaboration Friction with Chatbots. GroundAI. 2017;1. https://www.groundai.com/project/how-software-developers-mitigate-collaboration-friction-with-chatbots/

3. Abdul-Kader SA, Woods J. Survey on Chatbot Design Techniques in Speech Conversation Systems. International Journal of Advanced Computer Science and Applications. 2015;6(6):72-80

4. Chandel S, Yuying Y, Yujie G, Razaque A, Yang G. Chatbot: efficient and utility-based platform. In Science and Information Conference (pp. 109-122). Springer, Cham. July 2018.

5. Brandtzaeg PB, Følstad A. Why people use chatbots. In International Conference on Internet Science (pp. 377-392). Springer, Cham. November 2017.

6. Radziwill NM, Benton MC. Evaluating quality of chatbots and intelligent conversational agents. Available at: arXiv:1704.04579

7. Paikari E, Van Der Hoek A. (2018, May). A framework for understanding chatbots and their future. In 2018 IEEE/ACM 11th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE) (pp. 13-16). IEEE. https://dl.acm.org/doi/10.1145/3195836.3195859

8. Hill J, Ford WR, Farreras IG. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. Comput Human Behav. 2015;49(6):245-50

9. Asch DA, Nicholson S, Berger ML. Toward Facilitated Self-Service in Health Care. N Engl J Med. 2019;380(6):1891-3