NEJM Catalyst | Innovations in Care Delivery

ARTICLE

Designing Accountable Health Care Algorithms: Lessons from Covid-19 Contact Tracing

Lisa Lu, Alexis D'Agostino, MPP, Sarah L. Rudman, MD, MPH, Derek Ouyang, MS, Daniel E. Ho, JD, PhD

Vol. 3 No. 4 | April 2022

DOI: 10.1056/CAT.21.0382

An academic team at Stanford University worked with the County of Santa Clara Public Health Department to develop a machine-learning system for language matching in Covid-19 contact tracing through a collaborative design process. Although developed for a specific public health activity, the setting is representative of a wide range of health care delivery contexts, and the experience lays out how health care organizations can develop accountable algorithms that improve care, mitigate risk, and enable evaluation by stakeholders. Key elements of the design process involved: (1) a partnership and stakeholder consultation to develop a common understanding of and iteration around algorithmic design; (2) the use of a model understandable to all stakeholders, which exhibited only modest performance degradation relative to more complex models; and (3) a randomized controlled trial and qualitative survey of how the algorithm impacted real-world contact tracing, providing an evaluation that goes beyond narrow technical measures of algorithmic performance.

While rapid advances in artificial intelligence (AI) have the potential to revolutionize health care, there are also serious concerns around its adoption. AI systems can exacerbate bias,[1] fail to produce output that concurs with best health care judgment,[2] or even degrade human decision-making when relying on artifacts in data used to train the algorithm.[3] The question of how to build out health care algorithms reflects the broader debate around the accountability of AI systems.[4] The core question is whether and how we should build algorithms that reproduce what is good about human clinical judgment — at scale or with reduced resources — without exacerbating what is flawed in the same manual processes. Accountability means that stakeholders must be able to understand, interrogate, and evaluate the risks and benefits of the algorithm.[5]

Much of this debate has remained abstract.[6] In this article, we use a concrete example of how an AI system in an important health care and public health setting improved patient outcomes and, specifically, overcame an existing disparity in care. Through an academic–public collaboration, we developed a machine-learning approach to match the language needs of incoming patients with contact tracers, which reduced case completion times by 14 hours and improved patient and contact tracer engagement.

We document the technical components and results elsewhere,[7] focusing in this article on our design process and extracting broader lessons in how to scope out, develop, interrogate, and assess an algorithmic solution. We attempt to address the following question: What is the process by which multiple actors — computer scientists, social scientists, doctors, public health officials, managers, and contact tracers — should exercise "the obligation to explain and justify" the algorithm's use, design, decisions, and impact?[8] In addition, public organizations — and government entities in particular — may struggle in academic and technical collaborations to deploy strategies at the forefront of scientific ability while ensuring the level of responsibility and minimal risk required of the public sector. The approach we used has broad applicability for language access across health care systems[9] and serves as a potential model with lessons for how to build accountable algorithms in important settings.

Challenges of Contact Tracing

As part of its Covid-19 response, Santa Clara County, which has 1.9 million residents, established a shelter-in-place order on March 16, 2020, along with five other Bay Area counties.[10,11] Santa Clara County created a contact tracing operation and scaled up to more than 900 tracers within a matter of weeks; the tracers called patients upon diagnosis, provided information about isolation, and elicited close contacts to notify of a potential exposure. The ultimate goal was to reach and encourage contacts to quarantine before they became contagious, which required optimization for speed.

> "*The core question is whether and how we should build algorithms that reproduce what is good about human clinical judgment — at scale or with reduced resources — without exacerbating what is flawed in the same manual processes.*"

Across all Covid-19 response efforts, contact tracing has proven challenging. First, many individuals are reluctant to engage, trust, and offer information about contacts to the government. Second, the infrastructure for testing and rapid dissemination of results has been strained locally and nationally by supply and personnel limitations as well as major data-transmission backlogs and errors. Third, the pandemic's disparate impact made contact tracing particularly prone to language barriers. While Latinx individuals represent roughly 25% of the Santa Clara County population, they constituted more than 50% of Covid-19 cases. Laboratory reports provided incomplete information about language needs. In the software used for case investigation and contact tracing (CICT) efforts across the state of California, there was no way to match incoming cases to tracers on the basis of language spoken, and such needs were hence met by a secondary step of dialing in state interpreter services.

The Design Process for a Language-Matching Algorithm

In July 2020, the County of Santa Clara Public Health Department (PHD) and Stanford's Regulation, Evaluation, and Governance Lab (RegLab) began a series of conversations to scope out potential collaborations. The RegLab has a particular expertise in data science, machine-learning, and public sector applications. PHD had created its own Covid-19 Situational Awareness Branch, so key to collaboration was a series of conversations to identify complementary efforts. Early conversations surfaced language-matching capability as a priority because of both the perceived acute need and the feasibility of meeting the need.

Data Use

One of the major barriers was the contracting process for the data use agreement. Because information from laboratory reports was sparse, there was a strong need to be able to use all available information (e.g., name, date of birth, and address) to develop accurate predictions of language need. Extensive negotiation was required to satisfy the security and privacy requirements. While this process took 3 months — a long time in pandemic response terms, but short relative to ordinary contracting — the Stanford team was able to develop the core algorithm using a non–Covid-19 administrative data set. All data were ultimately placed on a secure server, suitable for protected health information.

Collaborative Process

Our teams met weekly for the Stanford team to share project updates and for PHD to share how the pandemic and contact tracing response was evolving. These meetings were critical to developing a common understanding of data sources and the potential limits of an algorithmic solution (Table 1).

For instance, the county used the state's contact tracing data-management platform, the California CONfidential NEtwork for Contact Tracing (CalCONNECT), which had a field for language, but the county had little confidence the field was used in a reliable way, which led the Stanford team to seek out different administrative data sets to train the model. In addition, these meetings enabled the Stanford team to learn as much as possible about the substance and process of contact tracing. A key question, for instance, lies in the human–computer interaction:[12,13] How should human contact tracers use *risk scores* for language need? Our solution centered on the designation of a language specialty team to enable team leads to route cases without breaking the existing organizational structure. These meetings were also critical

Table 1. Translating Key Technical and Public Health Questions Related to Algorithm Development

A. Technical question	B. Public health question	C. Impact on intervention
What *threshold* should be used to predict "positive cases" in a classification model? Should the threshold vary over time?	What is the capacity of the language specialty team? How does the team adjust with fluctuating case counts?	Building in a flexible threshold that allowed for trade-offs between model performance and operational capacity in response to case count and team size fluctuations
What *feature set* should be included to train the model?	Do we know what kinds of patients are most likely to benefit from bilingual contact tracing?	Selecting features associated with increased likelihoods of having a non-English preferred language (i.e., age, area, and name)
Should the model be regularly *retrained* on the basis of CalCONNECT fields that indicate language?	How consistent are our guidance and usage of the language field in the workforce?	More standardized guidance on usage of the language field in the workforce and early monitoring of the model's performance on CalCONNECT language data
If we retrain, should we be worried about *runaway feedback loops*, if language specialty team will pay more attention to cases flagged by the algorithm than if they had been coincidentally assigned a Spanish speaker without the assignment algorithm?	How different should the instructions be to the language specialty team, and is it possible to improve how all contact tracers engage patients around language issues?	Trained all staff on the importance of correctly identifying and documenting language, improving data in all parts of the program (such as case reporting by subgroups), not just the case assignment process
What kind of a *model* should be employed (e.g., logistic regression, random forest, heuristic model)?	How can we explain the approach to key decision makers and make discretionary choices (e.g., about the cutoff and team size) transparent?	Deployed an explainable heuristic approach based on a small, understandable set of variables rather than a more accurate but more complex random forest model
How should the model results be *deployed* into the contact tracing process?	Should the process be automated? Or should team leads or individual tracers choose cases?	A language specialty team composed entirely of bilingual contact tracers was created from among the existing pool of contact tracers to facilitate case assignment. Model results funneled cases to the team, after which team leads had discretion in case assignment. Risk scores were also included as part of the case record.
What *performance measures* should be used?	How would we know when language matching is not working?	Close monitoring of algorithmic performance through data collection and feedback from contact tracers early on in the pilot

This table provides an illustration of technical-domain translation through a collaborative design process. Each row provides a question stated in technical machine-learning terms (italicized) in column A; the resulting public health conversation around the goals, constraints, and operation of contact tracing in column B; and the ultimate real-world impact on the intervention in column C. CalCONNECT = California CONfidential NEtwork for Contact Tracing (the state's contact tracing data management platform). Source: The authors

to: (1) developing user comfort around the process and accuracy of the algorithm, which meant we used more explainable approaches, and (2) building in an evaluation to assess impact.

Ongoing Monitoring

Upon deployment, we provided channels for language specialty team members to report any issues, and we tailored the algorithm to fluctuating caseloads. We developed a dashboard of results — including the actual preferred language of the case and the language the interview was conducted in — to enable continuous monitoring, which ensured fidelity with the assignment process and enabled adaptation when the third surge hit Santa Clara County in December 2020. After the pilot period, which ran from December 2020 to February 2021, we also fielded a survey to all contact tracers to understand their assessment of existing state interpreter services and, where applicable, the language-matching program.

Results

The core results of the pilot show that the intervention resulted in more efficient contact tracing through significant time savings, despite the addition of a process for identifying and reassigning the case records in the software system. Time from a case being opened to completion of the initial interview was estimated to be reduced by nearly 14 hours for patients who would have been matched to a bilingual contact tracer if flagged as a likely Spanish speaker. The likelihood of completing a case interview on the same day it was opened increased by 12 percentage points. Additionally, there was a 4-percentage-point reduction in patient refusal to interview, indicating improved patient engagement under the program. To put this in context, for all patients in the month prior to the pilot, the average time from a case being opened to completion of the initial interview was nearly 30 hours; the average refusal rate was nearly 2%; and the average rate of interviewed cases that were completed on the same day they were opened was 37%.

> “*Time from a case being opened to completion of the initial interview was estimated to be reduced by nearly 14 hours for patients who would have been matched to a bilingual contact tracer if flagged as a likely Spanish speaker.*”

The survey results from the contact tracers corroborated the trial results, with 45% of respondents indicating that using a third-party interpreter increased the time for contact tracing *considerably* or by *a great deal*, showing the time-saving benefit of the program. In addition, 67% of bilingual respondents reported that being able to conduct the interview in a non-English language themselves was easier than dialing in a third-party interpreter. Respondents reported that the interpretation service, although valuable, could sometimes make it more difficult to build rapport with their clients because of technical issues and increased interview length. Reported benefits of language matching included increased client satisfaction and engagement, an improved ability to seek and provide important information, and an improved ability to build trust.

Lessons

Overall, the results showed that the machine-learning system for Covid-19 contact tracing was a strong success, and we believe it has several important general lessons on how to develop accountable health care algorithms.

Bias and Equity

One of the dominant concerns with AI systems is built-in bias against certain demographic groups.[14] Usually, algorithms are built to optimize for accurately predicting an outcome and only attempt to correct for underlying bias by adding constraints on how the algorithm works.[15] For instance, if an algorithm had predicted interview time in order to optimize case processing time, developers might take supplementary steps to attempt to mitigate potential disparities across demographic groups.

But here, health equity (not case processing time) was itself the animating *objective*. The serious concern was a status quo system — quasi-random assignment of patients to tracers — in which patients with a preference to speak Spanish (Spanish speakers) assigned to a non–Spanish-speaking contact tracer would have experienced unique downsides: difficulty in communication, the potential use of a time-intensive state interpreter, and, thus, significant barriers to their ability to isolate and quarantine. In contrast, under the system we adopted, some number of patients were algorithmically identified as potential Spanish speakers and routed to Spanish-speaking tracers; by design, there was virtually no downside to a false-positive identification, because all contact tracers also spoke English. The identification of this potential “do no harm” solution was critical to a first-order focus on equity. The interdisciplinary team also surfaced a range of discretionary choices that had second-order equity impacts: we rejected an approach based solely on voter registration data, for instance, because that could disproportionately leave out demographic groups affected most acutely by the pandemic.

Stakeholder Consultation

Algorithms go astray when there is no genuine exchange between domain and technical experts.[16] Development of the algorithm involved extensive iteration among technical developers, PHD leadership, and end users on the contact tracing team to develop a common understanding of the goals, data, and main discretionary decisions in building out the AI system (Table 1).[17] We were helped here by the fact that the county team was already moving in a more algorithmic direction by experimenting with the use of ZIP Codes to match bilingual contact tracers with patients from specific geographic communities.

The AI system was, in that sense, a logical extension of prior efforts, and our extensive exchanges guided development and integration. We note that the design process did not scope as far as some participatory frameworks might advocate,[18] namely to include broad segments of other community members. We consulted with the county’s Racial and Health Equity Senior Manager, who had extensive ties to the community, but a more extensive process would have significantly delayed the intervention in an already time-sensitive setting. In other contexts, deeper community engagement may be warranted.

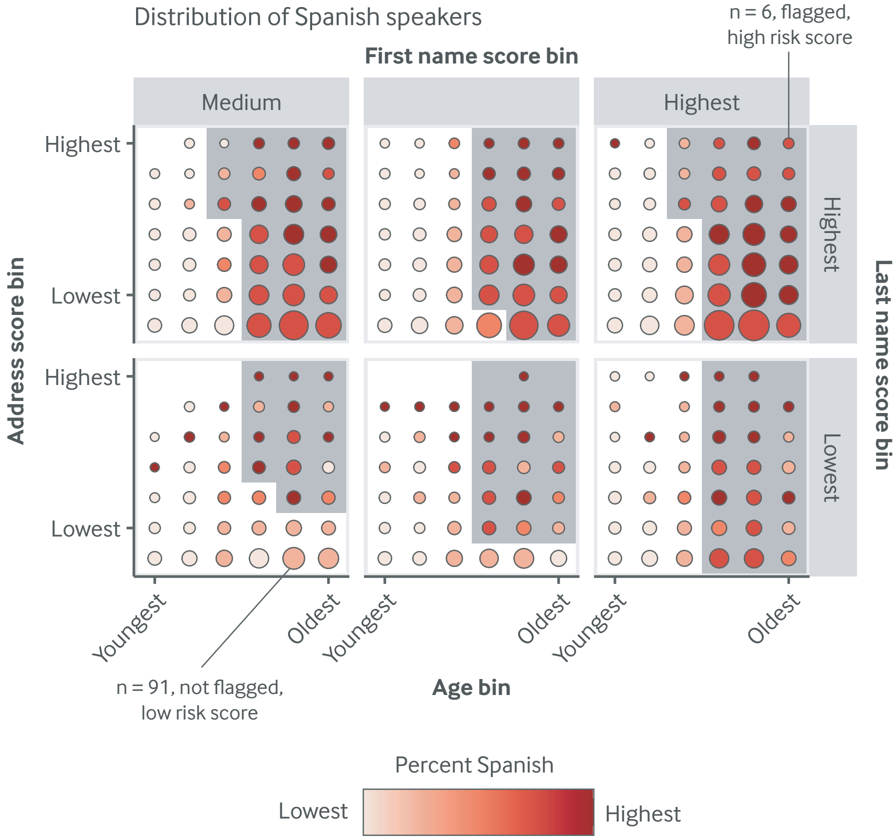
Transparency and Interpretability

One of the key elements in stakeholder consultation was a clear articulation of the algorithm. There is natural resistance to *black box* methods (which involve testing or deployment without a clear understanding of how an algorithm arrives at its predictions), and one of the roles for the technical team was to explain the logic behind the algorithm. What sources of data can be relied upon? How should we determine cutoffs from risk scores? A key decision was to favor a simpler, more explainable *heuristic approach* instead of a more complex, fully machine-learned *random* *forest model.*[19] A random forest model leverages multiple randomly created decision trees, each of which works on a random subset of features, which makes for a much less interpretable decision process.[19] The approach we designed is illustrated in Figure 1 and shows how the key laboratory report information — name, address, and age — are combined with census information to predict the likelihood of a Spanish speaker.

FIGURE 1

Explainable Algorithm Design

This figure shows the distribution of a subset of Spanish speakers in Santa Clara County from an administrative data set that was used to train the algorithm across the heuristic approach's inputs: age, address score, first name score, and last name score. The deployed algorithm binned first name score into six total bins, and this figure displays only the highest three first name score bins. Each bin was associated with a *risk score* representing the proportion of individuals in that bin who were Spanish speakers. We then calculated a cutoff boundary to identify cases who were more likely to have Spanish language needs. In the pilot, cases who fell into one of the flagged bins would be routed to a language specialty team. Size: circle size corresponds to the number of individuals in each bin. Hue: darker red corresponds to a higher risk score for Spanish speakers. Hue: lighter red corresponds to a lower risk score for Spanish speakers. Shading: gray shading denotes populations predicted to have Spanish language needs.



Source: The authors

NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

Individuals were classified into bins on the basis of those attributes, with each bin being associated with a *risk score* representing the likelihood of an individual being a Spanish speaker, on the basis of administrative data, making it transparent to understand why an individual may or may not have been flagged. This kind of figure became a key way to articulate how the algorithm extended earlier ideas of ZIP Code targeting to take advantage of all information contained in laboratory reports.

We compared the performance between these two approaches by analyzing the trade-off between sensitivity and specificity (calculated as the area under the receiver operating characteristic curve [AUC]) and the relationship between precision and sensitivity (calculated as the area under the precision-recall curve [AUCPR]). The greater the AUCs, the better the model is in terms of accurately distinguishing between Spanish and non-Spanish speakers. The heuristic approach had an AUC of 0.94 and an AUCPR of 0.85, while the random forest had an AUC of 0.95 and AUCPR of 0.98. While there was a boost in performance from the less interpretable random forest model, its performance was not deemed a practically significant improvement over the heuristic approach, which was easier to explain to stakeholders. Collaborative design enables decision-makers to make informed decisions about potential trade-offs between accuracy and interpretability.[20]

> “*A key decision was to favor a simpler, more explainable* heuristic approach *instead of a more complex, fully machine-learned* random forest model.”

Complementarity Between AI and Human System

AI systems do not sit alone. The algorithm we developed ultimately had to integrate with a complex case assignment and contact tracing process carried out manually by staff, illustrated in Figure 2.

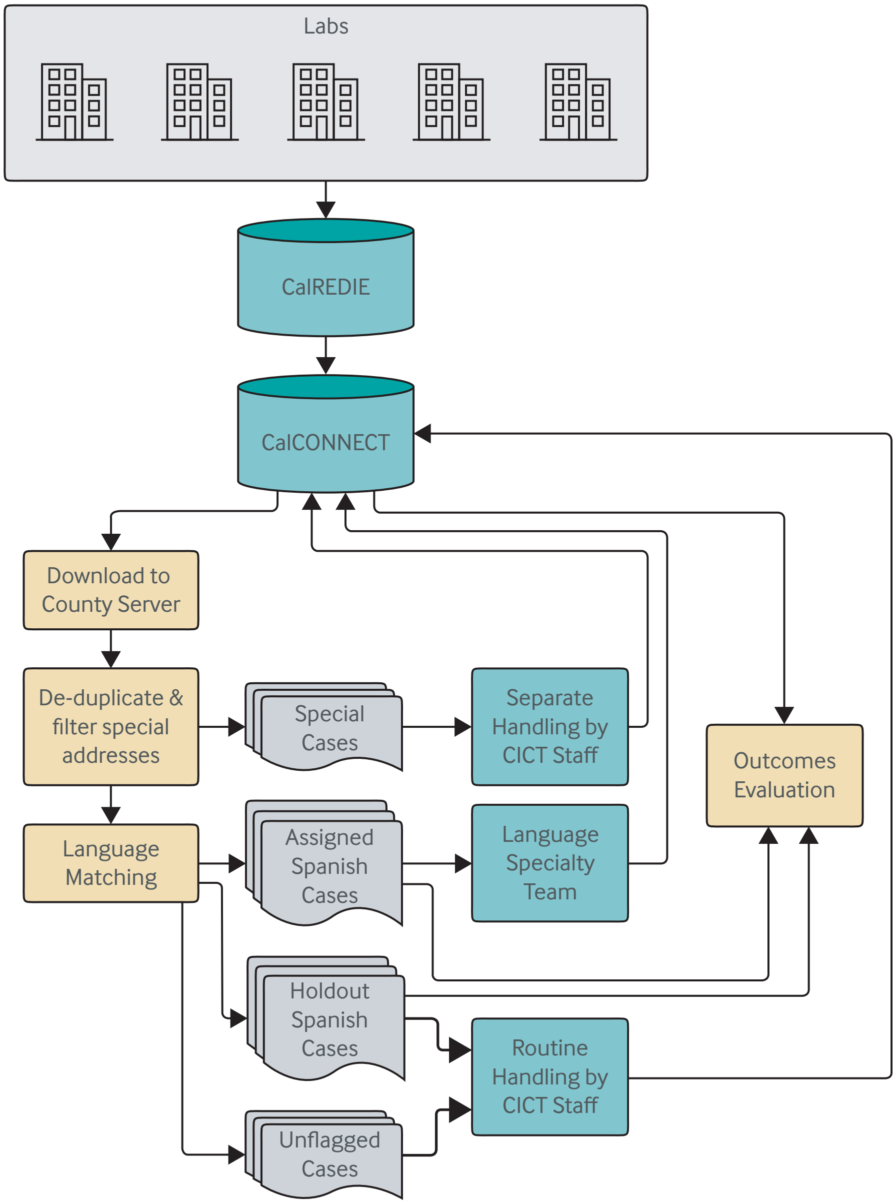
This multistep process, which involved human and AI systems, had to happen in real time, given the urgency of contacting patients as soon as possible after a diagnosis.

Laboratories all across the state upload digital laboratory reports containing Covid-19 testing data to the statewide testing system, California Reportable Disease Information Exchange (CalREDIE). The case data are transferred in batches to CalCONNECT, a state-managed contact tracing system built on a Salesforce cloud platform. From there, data are downloaded to the county’s server, and cases are deduplicated. Duplicates are handled by the data team. Cases from congregate settings, such as jails or long-term care facilities, are filtered out and routed to specialized CICT teams. The rest go through an additional language-matching step we embedded. During the pilot, flagged cases were randomly assigned to a Spanish-speaking language specialty team. Holdout flagged cases and unflagged cases then went through the normal contact-tracing process and were assigned by team leads to contact tracers on the basis of scheduling and

FIGURE 2

Overview of the County's Contact Tracing Data Pipeline

Laboratories transferred Covid-19 test information to the statewide testing system (CalREDIE); test data were transferred in batches to the contact tracing system (CalCONNECT); the county ran all new cases through a processing code to identify duplicates for handling by the data team and cases from congregate settings for assignment to special teams; and team leads assigned remaining cases on the basis of capacity and scheduling constraints. CalCONNECT = California CONfidential NEtwork for Contact Tracing (the state's contact tracing data management platform), CalREDIE = California Reportable Disease Information Exchange, CICT = case investigation and contact tracing.



Source: The authors

capacity constraints. Outcomes related to the interview, such as the patient's preferred language or engagement, were tracked and recorded in CalCONNECT to evaluate the pilot.

The technical team worked extensively to understand and streamline these existing systems before embedding the pilot code in existing operations. The manual, time-intensive batch runs were converted into dynamic, one-click batch runs that made the existing process more efficient and enabled the embedding of the model and subsequent routing of likely Spanish-speaking patients to the appropriate tracer. To embed the model, the technical team added an additional step to the processing code for duplicates and special cases that applied the algorithm to each new case and outputted a spreadsheet of predicted Spanish-speaking patients and their risk scores. These data were routed to the language specialty team leads as part of the case-assignment process. The team leads would review the spreadsheet and assign cases to their Spanish-speaking team members. These case records appeared in each tracer's queue in the system with a note included by their team lead that the case was identified as possibly needing a Spanish-speaking tracer. Technical staff had to understand the operational demands, and managers had to identify at which point the language risk scores would be most useful.

A More Expansive Notion of Evaluation

AI researchers often have narrow technical evaluation criteria for model performance, such as accuracy at the time of training on a limited data set. Yet what these criteria miss is that we should also care about how an AI system performs in the human context in which it is actually deployed.[12,13] There are many ways in which these technical measures may fail.

First, there might be "domain shift" from the training data to the patient population: the population of interest might be different from the sample on which the algorithm was trained, just as AI medical devices may perform differently across different hospital settings.[21] For this reason, we monitored performance of the algorithm throughout the pilot period, adjusting the algorithm to better fit the language data the team were collecting in their interviews.

Second, the AI system interacts with humans, so the algorithm is only useful if contact tracers can make efficient use of the information. Much of the time spent preparing for implementation focused on how to incorporate the algorithm into existing processes efficiently, to ensure that team leads could actually assign cases on the basis of the findings before the cases were assigned to someone else.

> “*The compelling results here actually formulate the business case for hiring more bilingual contact tracers. In that sense, language matching became much less necessary — the AI system, in a sense, actually automated itself away.”*

Lastly, the AI system could, at least theoretically, worsen outcomes relative to the baseline. Just as FDA drug approval hinges on evidence of efficacy and safety, we developed a randomized

Table 2. Language Matching: Sample of Contact Tracer Survey Responses

Enhances collection of data, supports compliance	"[M]y experience has been overwhelmingly positive! I think the 1:1 communication in the client's language has helped me to establish rapport and elicit more contacts… I think it also has helped in promoting isolation/quarantine compliance and relaying information on health monitoring, keeping household contacts as safe as possible. Finally, I hear a lot of appreciation from clients and that adds significantly to my satisfaction doing this work."
Improves efficiency,reduces frustration	"[Using the interpreter service resulted in] too much to repeat… [With language matching,] I was able to take more time to get specific questions answered. With the [interpreter, clients] became more impatient and frustrated and focused on ending the call."
Increases value of information	"I gather more valuable information, clients open up to me and are not afraid of asking for any additional questions or if they need any referrals. A rapport is made when you speak their language."
Establishes connection, increases rate of success	"As a Team Lead, I noticed that the rate of successful interviews (calls answered and clients willing to interview and answer questions) is much higher when the call is made by a native speaker of the same language as the client. [Tracers] who need to use the [interpreter] line had a higher percentage of incomplete interviews or no answers. When making calls myself in my native language, there was an immediate connection with the client that helped the interviews as it made the client feel more comfortable asking questions and sharing information."

Drawing on a survey of 411 responses of contact tracers, here focusing on members of the language specialty team, we see comments that express the perceived benefits of language matching.

Source: The authors

controlled trial to assess the effect of the language-matching intervention in actual operation that evaluated both the potential benefit to efficient use of human resources and the risk of interrupting functioning systems or exacerbating problems. In addition, we fielded a survey to understand the dimensions that could not be captured by data fields in the contact-tracing system (Table 2).

These results suggest that contact tracers agreed that the language-matching component improved outcomes for themselves and their patients. This aligns with a general lesson from this experience that a more expansive notion of evaluation will be necessary to understand performance when the human–computer interaction is taken seriously.[22]

Looking Ahead

This case study shows how we can design AI systems that are transparent and accountable and do not exacerbate existing harms and biases. The use of iterative stakeholder consultation and partnership across a range of perspectives — involving social scientists, machine learners, and public health officials — was critical to developing an equitable and effective system. This partnership approach can serve as a model to address one of the core questions in health care systems and the public sector: how to tame AI for good.

Another current concern is the anxiety that exists around whether AI systems may automate away the labor force.[23] But the compelling results here actually formulate the business case for hiring more bilingual contact tracers. In that sense, language matching became much less necessary — the AI system, in a sense, actually automated itself away. That is one of the lesser-known promises of evaluating AI systems for health care: they may teach us what the currently irreducibly human elements to quality of care are and how to maximize their utility.

Lisa Lu
Research Fellow, Regulation, Evaluation, and Governance Lab, Stanford University, Stanford, California, USA

Alexis D’Agostino, MPP
Senior Research and Evaluation Specialist, County of Santa Clara Public Health Department, San Jose, California, USA

Sarah L. Rudman, MD, MPH
Assistant Health Officer, County of Santa Clara Public Health Department, San Jose, California, USA

Assistant Clinical Professor (Affiliated), Department of Medicine, Division of Infectious Diseases, Stanford University School of Medicine, Stanford, California, USA

Derek Ouyang, MS
Program Lead, Regulation, Evaluation, and Governance Lab, Stanford University, Stanford, California, USA

Daniel E. Ho, JD, PhD
William Benjamin Scott and Luna M. Scott Professor of Law, Stanford University, Stanford, California, USA

Senior Fellow, Stanford Institute for Economic Policy Research, Stanford, California, USA

Associate Director, Stanford Institute for Human-Centered Artificial Intelligence, Stanford, California, USA

Director, Regulation, Evaluation, and Governance Lab, Stanford University, Stanford, California, USA

Disclosures: Lisa Lu, Alexis D’Agostino, Sarah L. Rudman, Derek Ouyang, and Daniel E. Ho have nothing to disclose.

References

1. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447–53 https://www.science.org/doi/10.1126/science.aax2342.

2. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 10, 2015. Accessed October 15, 2021. https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/KDD2015FinalDraftIntelligibleModels4HealthCare_igt143e-caruanaA.pdf.

3. Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated classification of skin lesions: from pixels to practice. J Invest Dermatol 2018;138:2108–10 https://www.jidonline.org/article/S0022-202X(18)32293-0/fulltext.

4. Engstrom DF, Ho DE. Algorithmic accountability in the regulatory state. Yale J Regul 2020;37:800–54 https://openyls.law.yale.edu/handle/20.500.13051/8311.

5. Institute of Medicine (US) Committee on Public Health Strategies to Improve Health. For the public's health. The role of measurement in action and accountability. Washington (DC): National Academies Press, 2011.

6. Rességuier A, Rodrigues R. AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data Soc 2020;7:1–5 https://journals.sagepub.com/doi/10.1177/2053951720942541.

7. Lu L, Anderson B, Ha R, et al. A language-matching model to improve equity and efficiency of COVID-19 contact tracing. Proc Natl Acad Sci USA 2021;118:e2109443118 https://www.pnas.org/content/118/43/e2109443118.

8. Wieringa M. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. January 27, 2020. Accessed October 18, 2021. https://dl.acm.org/doi/10.1145/3351095.3372833.

9. Flores G. Language barriers to health care in the United States. N Engl J Med 2006;355:229–31 https://www.nejm.org/doi/10.1056/NEJMp058316.

10. ABC 7 News. First to Order COVID-19 Lockdown, Santa Clara's Dr. Sara Cody Reflects on Pandemic. June 16, 2021. Accessed January 6, 2022. https://abc7news.com/dr-sara-cody-santa-clara-county-coronavirus-shelter-in-place/10795754/.

11. Eby K. Coronavirus Timeline: Tracking Major Moments of COVID-19 Pandemic in San Francisco Bay Area. ABC News 7. March 16, 2020. Accessed January 4, 2022. https://abc7news.com/timeline-of-coronavirus-us-covid-19-bay-area-sf/6047519/.

12. Wilder B, Horvitz E, Kamar E. Learning to complement humans. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence;* 2021:1526-33. https://dl.acm.org/doi/abs/10.5555/3491440.3491652.

13. Green B, Chen Y. The principles and limits of algorithm-in-the-loop decision making. Proc ACM Hum Comput Interact 2019;3:1–24 https://dl.acm.org/doi/abs/10.1145/3359152.

14. Barocas S, Hardt M, Narayanan A. Fairness and machine learning: limitations and opportunities. FairMLBook, 2019. https://fairmlbook.org/.

15. Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A. Algorithmic decision making and the cost of fairness. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 4, 2017. Accessed October 20, 2021. https://doi.org/10.1145/3097983.3098095.

16. Engstrom DF, Ho DE, Sharkey CM, Cuéllar M-F. Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. Administrative Conference of the United States. February 2020. Accessed October 19, 2021. https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf.

17. Veale M, Van Kleek M, Binns R. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. April 21, 2018. Accessed October 20, 2021. https://dl.acm.org/doi/10.1145/3173574.3174014.

18. Organisation for Economic Co-operation and Development. Focus on Citizens: Public Engagement for Better Policy and Services. OECD Studies on Public Engagement. OECD iLibrary. June 8, 2009. Accessed December 14, 2021. https://www.oecd.org/gov/focusoncitizenspublicengagementforbetterpolicyandservices.htm.

19. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15 https://www.nature.com/articles/s42256-019-0048-x.

20. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI-Explainable artificial intelligence. Sci Robot 2019;4:eaay7120 https://www.science.org/doi/10.1126/scirobotics.aay7120.

21. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021;27:582–4 https://www.nature.com/articles/s41591-021-01312-x.

22. Ho DE, Black E, Agrawala M, Fei-Fei L. Evaluating facial recognition technology: a protocol for performance assessment in new domains. Denver Law Rev 2020;98:753–73 https://static1.squarespace.com/static/5cb79f7efd6793296c0eb738/t/611c4adf9c3c7801cfddc0db/1629244128058/Vol98_Issue4_Ho_PRINT_FINAL.pdf.

23. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6:94–8 https://www.rcpjournals.org/content/futurehosp/6/2/94.