Author Affiliations: Center on Aging, Department of Medicine, University of Connecticut School of Medicine, Farmington.

Corresponding Author: Karina M. Berg, MD, MS, Center on Aging, Department of Medicine, University of Connecticut School of Medicine, UConn Health, 263 Farmington Ave, Farmington, CT 06030 (kberg@uchc.edu).

Published Online: April 8, 2019. doi:10.1001/jamainternmed.2018.8689

Conflict of Interest Disclosures: None reported.

1. Rowe JW, Berkman L, Fried L, et al. *Preparing for Better Health and Health Care for an Aging Population: A Vital Direction for Health and Health Care*. Washington, DC: National Academy of Medicine; 2016. https://nam.edu/preparing-for-better-health-and-health-care-for-an-aging-population-a-vital-direction-for-health-and-health-care/.

2. National Academies of Sciences, Engineering, and Medicine. *Families Caring For an Aging America*. Washington, DC: The National Academies Press; 2016.

3. Burgdorf J, Roth DL, Riffin C, Wolff JL. Factors associated with receipt of training among caregivers of older adults [published online April 8, 2019]. *JAMA Intern Med*. doi:10.1001/jamainternmed.2018.8694

4. Reinhard SC, Feinberg LF. The escalating complexity of family caregiving: meeting the challenge. *Fam Caregiving N Norm*. 2015:291-303. doi:10.1016/B978-0-12-417046-9.00016-7

5. AARP. New state law to help family caregivers. https://www.aarp.org/politics-society/advocacy/caregiving-advocacy/info-2014/aarp-creates-model-state-bill.html. Accessed January 23, 2019.

6. Shugrue N, Kellett K, Gruman C, et al. Progress and policy opportunities in family caregiver assessment: results from a national survey. *J Appl Gerontol*. 2017:733464817733104. doi:10.177/0733464817733104

7. RAISE Family Caregivers Act, S 1028, 115th Cong, 1st Sess (2017).

Automatically Charting Symptoms From Patient-Physician Conversations Using Machine Learning

Automating clerical aspects of medical record keeping through speech recognition during a patient's visit[1] could allow physicians to dedicate more time directly with patients. We considered the feasibility of using machine learning to automatically populate a review of systems (ROS) of all symptoms discussed in an encounter.

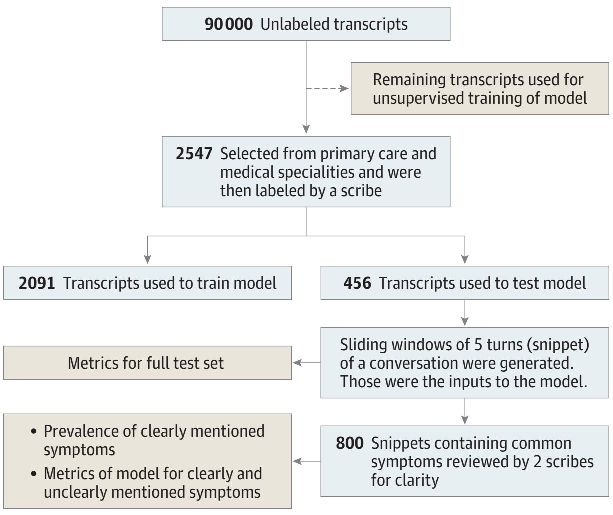
Methods | We used 90 000 human-transcribed, deidentified medical encounters described previously.[2] We randomly selected 2547 from primary care and selected medical subspecialties to undergo labeling of 185 symptoms by scribes. The rest were used for unsupervised training of our model, a recurrent neural network[3,4] that has been commonly used for language understanding. We reported model details previously.[5]

Because some mentions of symptoms were irrelevant to the ROS (eg, a physician mentioning "nausea" as a possible adverse effect), scribes assigned each symptom mention a relevance to the ROS, defined as being directly related to a patient's experience. Scribes also indicated if the symptom was experienced or not. A total of 2547 labeled transcripts were randomly split into training (2091 [80%]) and test (456 [20%]) sets.

From the test set, we selected 800 snippets containing at least 1 of 16 common symptoms that would be included in the ROS, and asked 2 scribes to independently assess how likely they would include the initially labeled symptom in the ROS. When both said "extremely likely" we defined this as a "clearly mentioned" symptom. All other symptom mentions were considered "unclear."

The input to the machine learning model was a sliding window of 5 conversation turns (snippets), and its output was each symptom mentioned, its relevance, and if the patient experienced it. We assessed the sensitivity and positive-predictive value, across the entire test set. We additionally calculated the sensitivity of identifying the symptom and the accuracy of correct documentation, in clearly vs unclearly mentioned symptoms. The **Figure** outlines the study design. The study was exempt from institutional review board approval because of the retrospective deidentified nature of the data set and the snippets presented in this manuscript are synthetic snippets modeled after real spoken language patterns, but are not from the original dataset and contain no data derived from actual patients.

Figure. Study Design


90 000 Unlabeled transcripts
Remaining transcripts used for unsupervised training of model
2547 Selected from primary care and medical specialities and were then labeled by a scribe
2091 Transcripts used to train model
456 Transcripts used to test model
Metrics for full test set
Sliding windows of 5 turns (snippet) of a conversation were generated. Those were the inputs to the model.
• Prevalence of clearly mentioned symptoms
• Metrics of model for clearly and unclearly mentioned symptoms
800 Snippets containing common symptoms reviewed by 2 scribes for clarity

Description of how data were used to construct the model, how subsets were labeled, and where metrics were calculated.

Results | In the test set, there were 5970 symptom mentions. Of these 5970, 4730 (79.3%) were relevant to the ROS and 3510 (74.2%) were experienced.

Across the full test set, the sensitivity of the model to identify symptoms was 67.7% (5172/7637) and the positive predictive value of a predicted symptom was 80.6% (5172/6417). We show examples of snippets and model predictions in the **Table**.

From human review of the 800 snippets, slightly less than half of symptom mentions were clear (387/800 [48.4%]), with fair agreement between raters on the likelihood to include a symptom as initially labeled in the ROS ($\kappa = 0.32$, $P < .001$). For clearly mentioned symptoms the sensitivity of the model was 92.2% (357/387). For unclear ones, it was 67.8% (280/413).

The model would accurately document—meaning correct identification of a symptom, correct classification of relevance to the note, and assignment of experienced or not—in 87.9% (340/387) of symptoms mentioned clearly and 60.0% (248/413) in ones mentioned unclearly.

Discussion | Previous discussions of autocharting take for granted that the same technologies that work on our smartphones will work in clinical practice. By going through the process of adapting such technology to a simple ROS autocharting task, we report a key challenge not previously considered: a substantial proportion of symptoms are mentioned vaguely, such that even

Table. Examples of Predictions on Various Snippets

Example	Snippet Conversation	Label	Prediction
Colloquial references to symptoms were correctly handled by the model.	PT: Yeah. DR: Anything else? PT: [I have pain in my belly] or [I have stomach-aches] or [My stomach has been hurting]. DR: When? PT: After I eat.	Abdominal pain (experienced)	Abdominal pain (experienced)
The model can identify when symptoms are not about the patient's experience (ie, irrelevant).	DR: That must have been really scary for you and your son. PT: Yeah, what are the normal signs of an allergic reaction? DR: Some people have a hard time breathing and get hives all over. PT: What should I do if it happens again to my son? DR: Does he have an injector?	Shortness of breath (not about patient); hives (not about patient)	Shortness of breath (not about patient); hives (not about patient)
The model can detect descriptions of symptoms that are clearly explained but not explicitly mentioned. This is a complex natural-language understanding task.	DR: Any problems with your urination? PT: I feel like I need to go all the time. DR: Any accidents? PT: No, I always make it on time. DR: Oh, okay.	Frequent urination (experienced); urinary incontinence (not experienced)	Frequent urination (experienced); urinary incontinence (not experienced)
Some normal physiological experiences can sound like symptoms, but it is unclear if a clinician would even document this as abnormal, although the scribe and model both identified it.	DR: What happens after you wake up? PT: I get up to turn off the alarm and my heart rate jumps up. DR: You feel your heart racing? PT: Yeah, then it goes back to normal in a few seconds. DR: Okay.	Palpitations (experienced)	Palpitations (experienced)
The model identified fever and cough correctly. Although clear to a human, "decreased appetite" is not identified by the model. We note that it is mentioned only implicitly (the patient could mean anorexia or discomfort with swallowing).	PT: It has been a hard few days. DR: Tell me what has been going on. PT: Two days ago I noticed I was running a fever and I also started having this bad cough. My throat also started hurting and I didn't feel like eating anything. I was worried I was getting the flu, so I didn't go to work and came here instead. DR: Sorry to hear that. PT: What should I do?	Fever (experienced); cough (experienced); sore throat (experienced); decreased appetite (experienced)	Fever (experienced); cough (experienced); sore throat (experienced)
The model incorrectly identified the patient as reporting depression, which is implicitly negated.	DR: It is not uncommon to feel different after starting steroids. PT: Oh, I didn't know that. DR: So you think you are getting depressed after starting it? PT: I think I am feeling more anxious than feeling depressed. DR: Go on.	Anxiety (experienced); depression (not experienced)	Anxiety (experienced); depression (experienced)

Abbreviations: DR, physician; PT, patient.

human scribes do not agree how to document them. Encouragingly, the model performed well on clearly mentioned symptoms, but its performance dropped significantly on unclearly mentioned ones. Solving this problem will require precise, though not necessarily jargon heavy, communication. Further research will be needed to assist clinicians with more meaningful tasks such as documenting the history of present illness.

Alvin Rajkomar, MD
Anjuli Kannan, AB
Kai Chen, PhD
Laura Vardoulakis, PhD
Katherine Chou, MSc
Claire Cui, PhD
Jeffrey Dean, PhD

Author Affiliations: Google LLC, Mountain View, California.

Corresponding Author: Alvin Rajkomar, MD, Google LLC, 1600 Amphitheatre Pkwy, Mountain View, CA 94043 (alvinrajkomar@google.com).

Accepted for Publication: December 8, 2018.

Published Online: March 25, 2019. doi:10.1001/jamainternmed.2018.8558

Author Contributions: Dr Rajkomar had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: All authors.
Acquisition, analysis, or interpretation of data: Rajkomar, Kannan, Chen, Vardoulakis, Chou.
Drafting of the manuscript: Kannan, Chen, Rajkomar, Vardoulakis.
Critical revision of the manuscript for important intellectual content: Rajkomar, Kannan, Chen, Chou, Cui, Dean.
Statistical analysis: Rajkomar, Kannan, Chen, Vardoulakis.

Obtained funding: Chou.
Administrative, technical, or material support: Chou.
Study supervision: Rajkomar, Chou, Cui, Dean.

Conflict of Interest Disclosures: All authors are employed by and own stock in Google. In addition, as part of a broad-based equity portfolio intending to mirror the US and International equities markets (eg, MSCI All Country World, Russell 3000), Jeff Dean holds individual stock positions in many public companies in the health care and pharmacological sectors, and also has investments in managed funds that also invest in such companies, as well as limited partner and direct venture investments in private companies operating in these sectors. All other health care–related investments are managed by independent third parties (institutional managers) with whom Jeff Dean has no direct contact and over whom Jeff Dean has no control. The authors have a patent pending for the machine learning tool described in this study. No other conflicts are reported.

Additional Contributions: We thank Kathryn Rough, PhD, and Mila Hardt, PhD, for helpful discussions on the manuscript; Mike Pearson, MBA, Ken Su, MBA, MBH, and Kasumi Widner, MS, for data collection; Diana Jaunzeikare, BA, Chris Co, PhD, Daniel Tse, MD, and Nina Gonzalez, MD, for labeling; Linh Tran, PhD, Nan Du, PhD, Yu-hui Chen, PhD, Yonghui Wu, PhD, Kyle Scholz, BS, Izhak Shafran, PhD, Patrick Nguyen, PhD, Chung-cheng Chiu, PhD, Zhifeng Chen, PhD, for helpful discussions on modeling; and Rebecca Rolfe, MSc, for illustrations. All individuals work at Google. They were not compensated outside of their normal duties for their contributions.

1. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA*. 2018;319(1):19-20. doi:10.1001/jama.2017.19198

2. Chiu C-C, Tripathi A, Chou K, et al. Speech Recognition for Medical Conversations. In: Interspeech 2018. ISCA: ISCA; 2018. https://www.isca-speech.org/archive/Interspeech_2018/abstracts/0040.html. Accessed December 8, 2018.

3. Sutskever I, Vinyals O, Le QV. Sequence to Sequence Learning with Neural Networks. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. Advances in Neural Information Processing Systems. vol 27. 2014. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf. Accessed December 8, 2018.

4. Cho K, van Merriënboer B, Gülçehre Ç, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014:1724-1734.

5. Kannan A, Chen K, Jaunzeikare D, Rajkomar A. Semi-supervised Learning for Information Extraction from Dialogue. In: Interspeech 2018. ISCA: ISCA; 2018. https://www.isca-speech.org/archive/Interspeech_2018/abstracts/1318.html. Accessed December 8, 2018.

Characteristics of Digital Health Studies Registered in ClinicalTrials.gov

Digital health is the application of software or hardware, often using mobile smartphone or sensor technologies to improve patient or population health and health care delivery.[1] In contrast to drugs and traditional medical devices, which have strict regulatory guidelines on safety and efficacy, the clinical evidence generation for digital health tools may be motivated by other factors, including adoption, utilization, and value, that may influence study design and quality. The landscape of clinical evidence underlying digital health interventions has not been well characterized.[2,3] We sought to evaluate the characteristics of digital health studies registered in ClinicalTrials.gov.

Invited Commentary
page 840

Methods | We performed a cross-sectional analysis of digital health studies in ClinicalTrials.gov.[4,5] To identify studies evaluating mobile-, web-, and electronic-based tools as well as digital medical devices, we searched ClinicalTrials.gov on January 22, 2017, using the Medical Subject Heading concepts (*mobile health*, *mHealth*, *ehealth*, *telehealth*, and *telemedicine*) and commonly used lay terms (*digital health*, *consumer health*, *mobile application*, and *wireless technology*). Variables were exported as structured fields when downloaded from ClinicalTrials.gov.[6] A single reviewer (C.E.C.) verified studies for inclusion, removed duplicates, and assigned each study to 1 of 13 clinical areas determined by iterative qualitative clustering against commonly accepted medicine domains. Descriptive statistics were calculated for key study characteristics, with additional stratification by study type (interventional vs observational, randomization status). We used the χ^2 test to compare proportions, and $P < .05$ was considered to be statistically significant.

Table 1. Digital Health Studies Registered in ClinicalTrials.gov

Study Type	No. (%) of Studies
All (N = 1783)	
Interventional	1570 (88.1)
Observational	213 (11.9)
Study allocation (n = 1776)	
Randomized	1257 (70.8)
Nonrandomized	519 (29.2)
Recruitment status[a]	
Not yet recruiting	218 (12.2)
Recruiting or enrolling	535 (30.0)
Active, not recruiting	176 (9.9)
Completed	692 (38.9)
Withdrawn or terminated	56 (3.1)
Unknown	103 (5.8)
Interventional (n = 1570)	
Intervention model (n = 1563)	
Parallel assignment	1147 (73.4)
Crossover assignment	88 (5.6)
Factorial assignment	43 (2.8)
Single group assignment	282 (18.0)
Masking (n = 1561)	
Double-blind	107 (6.9)
Single-blind[a]	417 (26.7)
Open-label or no masking	1031 (66.0)
Observational (n = 213)	
Observational model (n = 180)	
Case-control	26 (14.4)
Case-only	39 (21.7)
Cohort	100 (55.6)
Other	15 (8.3)
Time perspective (n = 199)	
Prospective	157 (78.9)
Retrospective	19 (9.5)
Cross-sectional	21 (10.5)
Completed, Suspended, Withdrawn, Terminated (n = 751)	
Study results	
Available	85 (11.3)
Not available	666 (88.7)

[a] Randomization status unknown for 7 studies.

[b] Patient, principal investigator, or assessor.