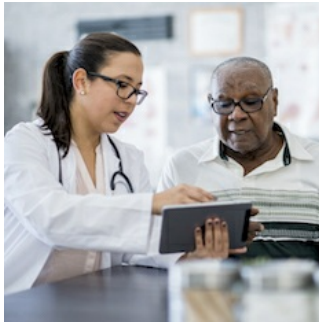# Machine Learning can automate charting using patient-doctor conversations

Automating symptoms recording 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, according to a new report published in [JAMA Intern Med](#).
Researchers considered the feasibility of using **machine learning to automatically populate a review of systems** (ROS) **of all symptoms discussed** in an encounter.

Methods

For the report, researchers used **90,000** human-transcribed, **de-identified medical encounters** described previously[2]. The 2547 subjects were then randomly collected from primary care and selected medical subspecialties to undergo labelling of **185 symptoms** by scribes. The rest were used for unsupervised training of the research model, a recurrent neural network[3,4] that has been commonly used for language understanding. There were previously reported model details[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, researchers then 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** used in the report, was a sliding window of 5 conversation turns (**snippets**), and its output was each symptom mentioned, its relevance, and if the patient experienced it. Then the team assessed the sensitivity and **positive-predictive value**, across the entire test set. They additionally calculated the sensitivity of identifying the symptom and the accuracy of correct documentation, in clearly vs unclearly mentioned symptoms.

The study was exempt from institutional review board approval because of the retrospective de-identified nature of the data set and the snippets presented in the manuscript are **synthetic snippets modelled after real spoken language patterns**, but are not from the original **dataset** and contain no data derived from actual patients.

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). Researchers presented examples of snippets and model predictions in the report.

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 **auto-charting** 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** auto-charting task, researchers reported a key challenge not previously considered: a substantial proportion of symptoms were mentioned vaguely, such that even human scribes do not agree on how to adequately 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, reported the researchers. Further research will be needed to assist clinicians with more meaningful tasks such as documenting the history of present illness.

Source: JAMA Intern Med.
Image Credit: iStock

Published on : Tue, 26 Mar 2019