# Is Auto-generated Transcript of Patient-Nurse Communication Ready to Use for Identifying the Risk for Hospitalizations or Emergency Department Visits in Home Health Care? A Natural Language Processing Pilot Study

**Jiyoun Song*, PhD, AGACNP-BC, APRN[1], Maryam Zolnoori*, PhD[1], Danielle Scharp, MSN, FNP-BC, APRN[1], Sasha Vergez, BS[2], Margaret V. McDonald, MSW[2], Sridevi Sridharan, MSc[2], Zoran Kostic, PhD[3], Maxim Topaz, PhD, RN[1,2]**

**[1]Columbia University School of Nursing, New York, NY; [2]Center for Home Care Policy & Research, Visiting Nurse Service of New York, New York, NY; [3]Columbia University, Fu Foundation School of Engineering and Applied Science, Department of Electrical Engineering, New York, NY**

## Introduction

Home healthcare (HHC), which is defined as skilled healthcare that is provided to an individual in their own home environment, is one of the fastest-growing segments of the healthcare industry in the United States (U.S)[1]. Approximately 5 million U.S. adults are currently receiving HHC, and the demand is expected to grow over the next few years[2]. With the shift in healthcare delivery from hospitals to the community due to increasing patient ages, shorter hospital stays, and improved health technology[3], an increasing number of patients with complex medical conditions rely on HHC after being discharged from an acute care setting[4, 5]. These changes in healthcare delivery have led to discussions about patient safety in HHC, including efforts to reduce acute care utilization (i.e., unplanned hospitalizations or emergency department (ED) visits)[2, 6]. In spite of several national and local quality improvement efforts focused on acute and chronic ambulatory care sensitive conditions (i.e. health conditions for which adequate management, treatment, and interventions delivered in the ambulatory care setting could potentially prevent hospitalizations)[7], approximately one in five patients still experience hospitalizations or ED visits during HHC services. There has not been a significant reduction in these occurrences over the past decade[8, 9]. With up to 30% of hospitalizations or ED visits deemed as potentially preventable[10], further efforts (such as identifying patients at risk and implementing risk mitigation strategies in a timely manner) may prevent these negative outcomes[11].

As data science methods are evolving, the use of cutting-edge data science technologies is gaining attention for identifying patients at risk and enhancing the predictive ability of risk prediction models. In addition, verbal communication between healthcare providers and patients are increasingly recognized as valuable informational resources because they involve information-seeking and sharing behaviors and often include extensive information[12]. Thus, as part of the trend, using artificial intelligence techniques developed for extracting insights from patient-nurse verbal communication (e.g., speech recognition) may be used to enhance risk identification[13]. Several recent studies used audio-recorded encounters between patients and health providers to predict further health outcomes. For example, one study applied machine learning to successfully predict the presence of post-traumatic stress disorder based on recorded communications between psychologists and army veterans[14]. Another study used the content of military couples' communications to predict risk for suicide[15]. However, despite that active verbal communications, including medical history taking, nursing assessments, and problem-focused discussions, take place between patients and HHC nurses during home visits, this has not been studied previously, nor have natural language processing (NLP) methods been applied to these verbal communications.

Converting speech to text is the first step in exploring the contents of verbal communication. The process of automatic speech recognition is a way for voice technology to recognize spoken sounds as words[16]. In this technology, computers can interact with humans in the most natural way using spoken interactions that combine voice and text techniques as a form of NLP. Over the past few years, speech recognition accuracy has increased as a result of advances in deep learning[17]. Even though automatic speech-to-text systems are becoming more prevalent, and they can even be found in toys, smartphones, and other devices with voice assistance, they have not yet been widely adopted in healthcare, especially in HHC to enable an analysis of verbal communication. Due to its significant potential to assist healthcare providers in analyzing verbal data by saving time and effort from manually generating transcripts and enabling real-

* Contributed equally

time transcription, auto-generated transcripts, including their efficiency and accuracy, should be evaluated to determine if they can replace human-generated transcripts.

To summarize, there have been no studies that have applied NLP to verbal communication between nurses and patients, nor have auto-generated transcripts been evaluated for the feasibility of replacing human-generated transcripts to improve patient risk identification. To address these knowledge gaps, the purpose of this study was (1) to refine the NLP algorithm that was developed by our team to identify terms associated with risk for hospitalizations or ED visits from clinical narrative notes in order for the algorithm to be applied to patient-nurse verbal communication in HHC; and (2) to compare the performances of NLP algorithms on auto-generated transcripts with human-generated transcripts in patient-nurse verbal communication.

## Methods

This retrospective cohort pilot study utilized information derived from patient-nurse verbal communications. In addition, sociodemographic and structured clinical assessment data (i.e., routine clinical assessment data called the Outcome and Assessment Information Set [OASIS] that is a federally mandated assessment in HHC, and administrative records) were extracted from electronic health records (EHRs) of a large HHC agency. **Figure 1** provides a general overview of the study methods. The study was approved by the Institutional Review Boards of the participating institutions.

### *Study Population*

We collected data from the largest non-profit HHC organization in the Northeastern U.S. between 2/16/2021 and 9/2/2021. Informed consent was obtained from all patients and nurses before their participation in the study. Nurses were recruited through email outreach to the participating HHC organization. After explaining the purpose, risks, and benefits of the study, our research assistant (employed by the study agency) obtained a written consent from those who agree to participate. Afterwards, the nurses were instructed on how to use the audio recording devices to record patient-nurse verbal communications. To recruit participating patients, the research assistant explained the purpose, risks, and benefits of the study over the phone prior to the nurse visit Once the patient verbally consented, a consent form was mailed for the patient's record. As a token of appreciation for their participation, patients and nurses received gift cards. More information on how the participants were recruited can be found elsewhere[18,19].

### *Verbal Communications*

In the study, five nurses recorded 127 patient-nurse visits with 44 patients using two audio recording devices (Saramonic Blink500 Pro B2 Pro and Sony ICD-TX6). We selected these two devices based on their high rating for functionality and usability compared to other audio recorders. Further details about the evaluation of the devices' functionality and usability were presented in our previous study[19].

For this analysis, we aimed to generate a maximum variation sample of the recordings. To accomplish that, we assessed each recording for information richness (i.e., verbal communication in which at least four health problems were discussed) and recording quality (i.e., recordings that had minimal background noise, echo, or interruptions). Consequently, a total of 22 recordings representing 15 patients were selected. A detailed description has been provided in our previous study[18].

### *Converting Speech to Text: Human-generated Transcripts*

The selected audio recordings were transcribed verbatim by a human expert transcriber using TranscribeMe, a HIPAA-compliant transcription services company (https://www.transcribeme.com). The transcripts included speaker information (e.g., speaker 1, speaker 2), timestamps, and content at the utterance level (i.e., the smallest unit of speech ending and beginning with a clear pause). Based on listening to audio recordings and reviewing transcript content, the authors identified and marked the speaker's role in transcripts (e.g., patient, nurse). To evaluate the accuracy of the human-generated transcripts of patient-nurse verbal communication, 500 utterances of communication were randomly selected and the author (JS) reviewed each utterance based on the time stamps attached to every utterance to ensure the quality of human-transcripts.

### *Converting Speech to Text: Auto-generated Transcripts*

We created auto-generated transcripts for selected audio recordings using the Amazon Transcribe Medical service (https://aws.amazon.com/transcribe/medical/), which is tailored for medical professionals who need to transcribe medically-related speech such as physician dictation, drug safety monitoring, and clinician-patient communications. Transcripts included speaker information (e.g., speaker 1, speaker 2), timestamps (i.e., start and end time), and content

(transcription) of verbal communication for each speaker[20]. To evaluate the accuracy, the word error rate (WER), a metric commonly used to evaluate speech recognition systems, was calculated against the human-generated transcripts by co-authors (MZ, SV)[19]. In the WER, transcription errors are calculated by substitution errors (such that the identification system incorrectly recognizes one word for another), deletion errors (such that the recognition system missed words), and insertion errors (such that words are introduced into the generated text by the recognition system) with refence to the total word in the transription[21].

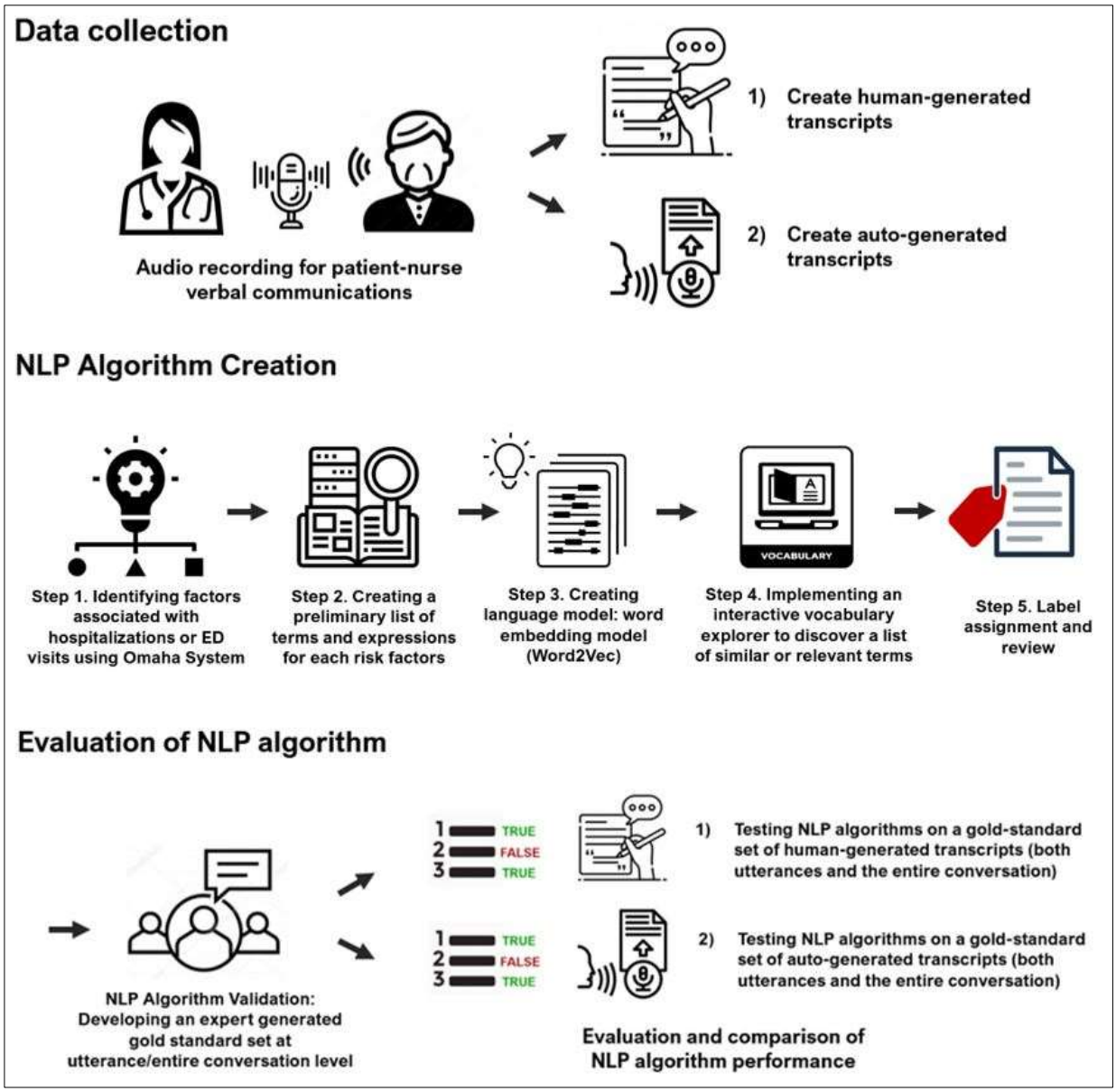


**Figure 1.** Study methods including natural language prcoessing (NLP) algorithm creation and validation

*Natural Language Processing (NLP) algorithm*

Following sections describe our previously developed rule-based NLP algorithm[22] and its refinement to process verbal communications. We used both nurses' and patients' utterances to identify the risk factors using the NLP algorithm since nurses' utterances can indicate signs and symptoms by describing their observations or educating them on the signs and symptoms as an intervention.

*A. NLP Algorithm Development*

*(1) Identifying factors associated with hospitalizations or ED visits using Omaha System*

Our research team previously identified factors associated with hospitalizations or ED visits among HHC patients by assembling a team of experts in nursing, HHC and informatics[22]. To generate a list of risk factors, we used the Omaha System problems, which is - a commonly used standard terminology for documenting clinical information among community-based care (https://www.omahasystem.org/)[23]. The Omaha System comprehensively covers environmental, psychosocial, physiological and health related behavior problem domains[23]. The experts determined that a subset of 31 out of 42 Omaha System problems (e.g., "Circulation", "Pain", "Medication regimen") with 160 sign/symptoms (e.g., sign/symptoms of 'edema' or 'discoloration of skin or cyanosis' under "Circulation" problem, or sign/symptoms of 'expresses discomfort or pain' or 'compensated movement or guarding' under "Pain" problem) could be considered as risk factors for hospitalizations or ED visits in HHC. See our prior publication for a complete list of risk factors[22].

*(2) Creating preliminary list of terms for risk factors using a vocabulary of standardized terminology*

The common approach is to first map narrative clinical notes to concepts from biomedical knowledge sources of standardized terminology like Unified Medical Language System [UMLS][24], which provides a richer set of concurrent concepts to integrate documents that contain related concepts[25, 26]. Thus, our team developed a comprehensive lexicon on each risk factor with synonyms from a standardized medical terminology database (i.e., UMLS). For example, for the problem of "pain", we identified UMLS synonyms such as "express discomfort," "express pain," or "aches."

*(3) Creating language model: word embedding model (Word2Vec)*

In this study, we used publicly available NimbleMiner NLP software (http://github.com/mtopaz/NimbleMiner)[27]. As a part of the tools used in the software, Word embedding (Word2Vec) was used. It is a language model that generates statistical representations of texts that learns to identify synonyms for terms of interest relevant to a specific domain by analyzing word associations[28]. To prepare clinical notes for the word embedding model training, the notes were pre-processed to remove punctuation and to lowercase all letters[29]. Next, it created an embedding vector for each clinical note with a maximum word length of four (4-grams)[29] before applying a skip-gram model that is a generalization of n-grams that doesn't require that the components of the text under consideration be consecutive but may skip over gaps[30]. Using this approach, we identified synonyms for risk factors in a large collection of HHC clinical notes (n = approximately 2.3 million) and terms used in patient-nurse verbal communication (n = 5,118 utterances).

*(4) Implementing an interactive rapid vocabulary explorer*

As part of the NimbleMiner software, two authors who are experts in HHC and informatics (JS, MT) implemented an interactive rapid vocabulary explorer to discover large vocabularies of relevant terms and expressions. The interactive rapid vocabulary explorer allowed us to develop an expanded list of synonymous expressions based on pre-populated lists of synonyms extracted from the UMLS. When the user enters a query term of interest (e.g., the problem of "Skin"), the system returns a list of similar terms it identified as relevant (e.g., "wound," "wound noted," "wd," "ulcer") based on cosine similarity metric. The user selects and saves relevant terms by clicking on them until no further relevant terms appear on the user interface. Negated terms (e.g., "denies," "no," "not," "ruled out") or other irrelevant terms not selected by the user were used for further processing, such as negation detection to distinguish and exclude negated terms.

*(5) Label assignment and review*

An NLP algorithm assigned labels to transcripts of patient-nurse verbal communications using terms that have been selected and saved (while excluding notes with negations and other irrelevant terms). Assigning a "positive label" means that a risk factor of interest is present in the transcripts. The users reviewed the assigned labels for accuracy.

*B. NLP Algorithm Validation*

To test whether the NLP algorithm was able to identify risk factors as accurately as a clinical expert, we employed "gold standard" testing methods. A total of 5,118 utterances from 22 transcripts of audio-recordings were manually annotated by co-authors (JS, DS) using Microsoft Excel to check for the presence of any of the Omaha System problem categories (risk factors) at the utterance level. Interrater agreement between annotators was good (Fleiss' Kappa = 0.77)[31] and all discrepancies were resolved through several consensus group meetings with a

senior author (MT)[18]. In addition, we created gold standard sets that would test for the presence of any categories at the level of the entire communication.

*C. Evaluation and Comparison of NLP algorithm Performance*

We applied our NLP system to the manually annotated 22 transcripts (i.e., gold standard testing set) and for each category calculated precision (i.e., positive predictive value defined as the number of true positives out of the total number of predicted positives), recall (i.e., sensitivity defined as the number of true positives out of actual number of positives), and F-score (i.e., the weighted harmonic mean of the precision and recall). Since F-measure measures both precision (the number of instances correctly classified) and robustness (not missing many instances), it is used widely to evaluate the algorithm performance. To avoid the inevitable error arising when evaluating algorithms' performance based on the low frequency of risk factors identified during verbal communication, only risk factors found more than 20 times were used to evaluate the performance.

The performance of NLP algorithms in the auto-generated and human-generated transcripts of patient-nurse verbal communication was compared at both the utterance level and the level of the entire communication within each visit. Because of the time and effort of manually annotating both human-generated and auto-generated transcripts, we measured only precision when measuring performance on auto-generated transcripts at the utterance level. In other words, we applied the NLP algorithm on the utterances of auto-generated transcripts, then reviewed the positively labeled utterances for confirmation.

**Results**

*Patient's Characteristics*

Among the sample of 15 patients, the average age was 67.3 (range 40.5-93.8) and 53% were male. The majority of patients were White (60%), followed by Black (33%), and Asian (7%) patients. About one-third of all patients lived alone. All patients had five or more prescription medications. The most common comorbid conditions were hypertension, chronic pulmonary disease, and diabetes (73%, 40%, and 33%, respectively). Furthermore, 33% of the visits involved postoperative care.

*Verbal Communications and Transcript Accuracy*

In the audio-recorded verbal communication, the average communication lasted 22 minutes (range 9 – 46.2), and each communication contained an average of 245 utterances (range 76 – 562). Out of 5,118 total utterances, 46.4% were produced by patients, and 53.6% by nurses.

The accuracy of human-generated transcripts was approximately 89% based on a random sample review of utterances from audio recordings. Many errors were caused by the transcriber's inability to identify the specific name of the medication as part of the sentence (e.g., "The only thing here is antibiotics [inaudible]") or by inaudible due to cross-talking or environmental noise. On the other hand, an average of the error rates across the auto-generated transcripts that were calculated against the human-generated transcripts yields a WER of 0.26 (SE: 0.004).

*Identify Risk Factors for Hospitalizations or ED Visits within Patient-Nurse Verbal Communication*

During the HHC visit, an average of 10 different risk factors (range 4 – 23) were manually identified. One or more risk factors were verbally discussed in approximately 7.1% (365/5118) of utterances. The problems of "Circulation," "Pain," and "Skin" were the most commonly mentioned at the utterance level of verbal communication. The following risk factors were mentioned over 20 times, therefore, they were included for further analysis to evaluate NLP performance: "Circulation," "Medication regimen," "Neuromusculoskeletal function," "Pain," "Respiration," "Skin."

*Evaluation and Comparison of NLP Algorithm Performance*

**Table 1** provides a comparison of the risk factor identification performance on human-generated and auto-generated transcripts at the utterance level. Overall, NLP algorithms showed good risk factor identification performance at identifying risk factors for hospitalizations or ED visits within utterances on both human-generated and auto-generated transcripts, with a precision of 0.81 and 0.79, respectively. According to evaluation metrics, NLP algorithms on auto-generated transcripts performed less accurately than human-generated transcripts from the perspective of risk factor identification performance. However, a closer look at the frequency of risk factors found within utterances detected by NLP algorithms indicated large differences between the human-generated and automatic transcripts. These examples illustrated the reason auto-generated transcripts detected a significant difference in the frequency of terms compared to human-generated transcripts.

Example 1) The sentence within human-generated transcripts "did leg swelling [risk factor: "Circulation"] happen again?" was incorrectly transcribed in auto-generated transcripts as "is captain tommy reaction at home leah leah". Therefore, the NLP algorithm was not able to identify the risk factor in the utterance.

Example 2) The sentence within human-generated transcripts "the problem is not really the salt but you don't want to feel breathless [risk factor: "Respiration"]" was not captured in auto-generated transcripts. Thus, as the part of transcripts was read as "the problem is not really the salt but you dont want to yeah alright," this utterance could not be identified as including the risk factors by the NLP algorithm.

Example 3) The sentence of human-generated transcripts "maybe they will say that insurance won't pay" was incorrectly transcribed in auto-generated transcripts as "maybe they will say that insurance wont pain." Thus, this was wrongly detected as having the risk factor of "pain."

**Table 1.** Evaluation and comparison of NLP algorithms at the utterance level

| **Risk factors (Omaha System Problem)** | **Human-generated transcripts** | | | **Auto-generated transcripts** |
|---|---|---|---|---|
| | **Precision** | **Recall** | **F-score** | **Precision** |
| Circulation | 0.94 | 0.81 | 0.87 | 0.89 |
| Medication regimen | 0.48 | 0.84 | 0.62 | 0.54 |
| Neuromusculo-skeletal function | 0.96 | 0.84 | 0.9 | 1 |
| Pain | 0.72 | 0.88 | 0.79 | 0.72 |
| Respiration | 1 | 0.56 | 0.72 | 0.66 |
| Skin | 0.75 | 0.88 | 0.81 | 0.92 |
| **Overall** | **0.81** | **0.8** | **0.79** | **0.79** |

* Note: Full annotation was not done on auto-generated transcripts, hence only precision was calculated

**Table 2** shows a comparison of the risk prediction ability for human-generated and auto-generated transcripts on the basis of the entire verbal communication of each HHC visit. On both human-generated and auto-generated transcripts, the NLP algorithms demonstrated excellent accuracy with a low number of false positives and false negatives. F-score was 0.91 for both types of transcripts. Throughout the entire verbal communication of each HHC visit, NLP algorithms detected exactly the same risk factors, in either human-generated or auto-generated transcripts.

**Table 2.** Evaluation and comparison of NLP algorithms at the entire verbal communication of each HHC visit

| **Risk factors (Omaha System Problem)** | **Human-generated transcripts** | | | **Auto-generated transcripts** | | |
|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F-score** | **Precision** | **Recall** | **F-score** |
| Circulation | 1 | 0.95 | 1 | 1 | 0.95 | 1 |
| Medication regimen | 0.67 | 0.75 | 0.71 | 0.67 | 0.75 | 0.71 |
| Neuromusculo-skeletal function | 1 | 1 | 1 | 1 | 1 | 1 |
| Pain | 0.95 | 1 | 0.97 | 0.95 | 1 | 0.97 |
| Respiration | 0.77 | 1 | 0.87 | 0.77 | 1 | 0.87 |
| Skin | 1 | 0.82 | 0.9 | 1 | 0.82 | 0.9 |
| **Overall** | **0.9** | **0.92** | **0.91** | **0.9** | **0.92** | **0.91** |

**Discussion**

This was the first study to investigate speech-to-text converted transcripts of patient-HHC nurse verbal communication to identify terms related to the risk for hospitalizations or ED visits during HHC. According to our recent study, verbal communication contained 50% more health problems compared to health problems captured by EHR, including structured data and narrative clinical notes[18]. Thus, using rich data sources, such as verbal communication, can improve recognition of problems verbally discussed during the HHC visit. This can ultimately lead to a more accurate identification of patients at risk for hospitalizations and better understanding of the risk factors during HHC. This study complements and extends our team's recent study that developed NLP algorithms utilizing clinical notes to identify risk factors associated with HHC patient hospitalizations and ED visits[32, 33].

Our findings suggested that information about risk of hospitalizations or ED visits was actively discussed verbally during HHC visits. One or more risk factors were mentioned in about 7.1% (365/5118) of utterances. The most frequently discussed risk factor was the "Circulation" problem, followed by "Pain" and "Skin". These findings were reflected in the characteristics of patients who participated in this study; a majority of them had hypertension (73%) and 33% of the visits included postoperative care. Thus, verbal interaction between patients and nurses can provide clues as to the type of care needed. On the other hand, the remaining utterances contained (1) health-related problems that mapped to Omaha System problems categories but were not considered risk factors for hospitalizations or ED visits (e.g., sign/symptoms of 'smokes or uses tobacco products' under "Substance use" problem or sign/symptoms of 'rhinorrhea or nasal congestion' under "Respiration" problem), (2) the intervention provided (e.g., "Treatment," "Surveillance," "Teaching, Guidance, and Counseling" or "Case management"), (3) "small talk," (4) answers to questions (e.g., "all right," "yes," etc.) and (5) clinical assessments that were within a normal range (thus they were not annotated as problems). Although these utterances constituted the majority of verbal communications, they are still important since they reinforce behaviors, confirm information, and allow for physical and emotional interaction between nurses and patients[34].

Furthermore, we confirmed the potential usefulness of the NLP approaches to extract information from verbal communication for understanding risk factors in HHC. Although NLP performance was evaluated only on the six risk factors because of the nature of this pilot study that had a low frequency of risk factors annotated in transcripts, the NLP algorithm performed well for identifying risk factors found in verbal communication: human-generated transcripts have a precision of 0.81 while auto-generated transcripts have a precision of 0.79 at the utterance level; the F-score on both human-generated and auto-generated transcripts is 0.91 on entire communication. In addition, we found that NLP performance was superior when it was measured on the entire verbal communication rather than on the level of utterances. This can indicate that risk factors were mentioned many times throughout the entire verbal communication, so NLP algorithms had a higher chance of capturing them. Since this study utilized rule-based NLP approaches, it offered potential for utilizing domain knowledge directly within the information extraction process, which results in algorithms that are clinically meaningful and transparent[35]. However, due to the declarative nature of rule-based approaches, they are inherently inflexible to generalize to minor linguistic variations, noise, or in the case of when the input data differ only in minor nuances[35]. Consequently, certain risk factors, such as the "Medication regimen" problem, showed poor performance. For example, the patient explicitly stated "but I have to buy some more" in the verbal communication. Even though it implied that the medication was not in stock, it was necessary to read utterances before or after the specific utterance to understand the problem in the context. Thus, NLP algorithms were unable to identify this risk factor which was annotated by human reviewers as "fails to obtain refills appropriately" under the "Medication regimen" problem. Alternatively, machine learning-based approaches have the advantage that statistical clues from documents and clusters of related words in a document can be gleaned from the keywords, while the major challenge with machine learning is that the system has to be trained so that it can recognize and respond to queries[36]. To resolve some of these issues, the use of hybrid NLP approaches that combine the rule-based and machine learning-based NLP approaches should be considered in future studies.

This study also aimed to compare human-generated transcripts and automatic transcripts in terms of NLP performance. This aim set out to explore, in the sense of a proof-of-concept, whether auto-generated transcripts could potentially replace the need for human-generated transcripts for the purpose of identifying risk factors for hospitalizations or ED visits in HHC through the application of NLP algorithms. In terms of prediction ability at the utterance level, NLP algorithms on auto-generated transcripts performed slightly worse than human-generated transcripts. Nevertheless, the fundamental issue was that some key terms or expressions were not captured due to the inaccuracy of the auto-generated transcripts, therefore, comparing the risk factor identification performance of two types of transcripts solely on the basis of evaluation metrics needs to be reconsidered. The automatic speech recognition (i.e., Amazon Transcribe Medical service for this study) uses statistical probabilities to deduce whole words from phonemes, and then derives complete sentences from these full words by analyzing them sequentially, starting from the first phoneme[16]. For this reason, the presence of grammatical errors, disfluencies, and other imperfections in recorded verbal communication can generate errors in auto-generated transcripts, then errors introduced by either the speaker or the automatic speech recognition system will be propagated to the next task, such as the development of NLP algorithms. Thus, we suggest sophisticated post-processing from an automatic speech recognition output should be applied in order to turn unintelligible text into readable text while maintaining the semantic meaning of the speaker[37].

The previous studies which used the auto-generated transcript for voice recordings (i.e., qualitative interview, physician note dictating, or reporting pathology results) were mostly recorded in a quiet environment or the voice recording devices were directly connected to a desktop computer[38-40]. Therefore, the quality of transcripts in such

controlled environments might be higher. For this study, despite using the noise removal feature to create the auto-transcript, due to the nature of HHC services offered in the home environment, the quality of our voice recordings between patients and nurses could be inferior to those recorded in a more controlled environment. Indeed, some of the recordings used in this study had constant background sounds (such as television noise, cross-talking, or distant caregiver voices), or the recorded voice was too far away while nurses provided certain interventions. Hence, there is a need to identify how to integrate speech- recognition into HHC clinical workflows and collect high-quality recordings, which will likely improve transcript accuracy.

*Clinical implications and future research*

Our findings showed that NLP algorithms risk factor identification performance was relatively high. Future studies need to expand the sample size to determine whether the algorithm is effective with all the risk factors for hospitalizations or ED visits in HHC. In addition, given the diversity of the population in HHC organizations, further research is needed on how these NLP algorithms work for patients or nurses whom English is not the primary language. Future NLP algorithms should be able to be flexible in their use to deal with grammatical errors or unusual expressions from non-native speakers.

Further efforts are also needed to improve the accuracy of the auto-generated transcripts. Nonetheless, this study's results suggest that automated voice recognition technologies could potentially be used to capture patient-nurse verbal communication and create clinical records in the EHR to (1) reduce the time and costs associated with human-generated documentation, and (2) reduce clinician burnout and cognitive load as a result of a heavy documentation load[41, 42]. With these insights, predictive analytics may also improve the identification of patients at risk for negative outcomes (e.g., hospitalizations or ED visits).

*Limitations*

Since this pilot study was exploratory, it has a small sample size, which can affect the generalizability of the results. However, at the utterance level, we analyzed over 5,000 utterances that were empirically acceptable for the classification task. In addition, the analysis of audio recordings of patient-nurse verbal communication was only performed at one HHC organization in the same geographical area, thus limiting generalizability due to the possibility of organizational-specific practice patterns (for instance, certain care protocols during HHC visits). The participants were not limited to native speakers. Since participants whose first language is not English may influence the quality of content of audio recordings. Although the NLP algorithm has been refined to include verbal communication, a small percentage of utterances may not provide sufficient data to detect statistical associations in a text containing verbal communication. Lastly, as verbal communication was directly reflected an individual's knowledge, perception, and attitude, the communication could be influenced by a range of factors, including the nurse's individual communication style.

## Conclusion

This study highlights the potential application of NLP approaches for detecting care needs in HHC by extracting information from verbal communication. The NLP algorithm performed well for identifying risk factors in verbal communication, with precision of over 0.79 at an utterance level and F-score of 0.91 on the entire communication. The risk factor identification performance of the NLP algorithm on auto-generated transcripts performed slightly worse at the utterance level than those on human-generated transcripts, but it was equally good over the entire communication. However, it is not yet possible to determine the usability of auto-generated patient-nurse communication just based on these pilot results. Future research should prioritize improving the accuracy of auto-generated transcripts to determine whether they can potentially replace human-generated transcripts for identifying risk factors for hospitalizations or ED visits in HHC by applying NLP algorithms.

## Acknowledgements

This study is funded by the pilot grant from Columbia University School of Nursing (Clinical Meaningful Concept in Communication "MC-ICON") and the Pilot Grant from Columbia University Center of Artificial Intelligence Technology ("Using artificial intelligence to identify homecare patient's risk of hospitalization and emergency department visits: speech-recognition feasibility study").

**References**


1. The Medicare Payment Advisory Commission. Report to the congress- medicare payment policy: Home health care services 2019 [Available from: http://www.medpac.gov/docs/default-source/reports/mar19_medpac_entirereport_sec.pdf.
2. Landers S, Madigan E, Leff B, Rosati RJ, McCann BA, Hornbake R, et al. The future of home health care: A strategic framework for optimizing value. Home health care management & practice. 2016;28(4):262-78. doi: 10.1177/1084822316666368.
3. Mitzner TL, Beer JM, McBride SE, Rogers WA, Fisk AD. Older adults' needs for home health care and the potential for human factors interventions. Proc Hum Factors Ergon Soc Annu Meet. 2009;53(1):718-22. doi: 10.1177/154193120905301118.
4. Jarvis WR. Infection control and changing health-care delivery systems. Emerging infectious diseases. 2001;7(2):170-3. doi: 10.3201/eid0702.010202.
5. Hardin L, Mason DJ. Bringing it home: The shift in where health care is delivered. The Journal of the American Medical Association. 2019;322(6):493-4. doi: 10.1001/jama.2019.11302.
6. The Medicare Payment Advisory Commission. Home health care services 2020 [Available from: http://www.medpac.gov/docs/default-source/reports/mar20_medpac_ch9_sec.pdf?sfvrsn=0.
7. Agency for Healthcare Research & Quality. Potentially avoidable hospitalizations: Ambulatory care-sensitive conditions 2015 [Available from: https://www.ahrq.gov/research/findings/nhqrdr/chartbooks/carecoordination/measure3.html.
8. O'Connor M, Hanlon A, Bowles KH. Impact of frontloading of skilled nursing visits on the incidence of 30-day hospital readmission. Geriatric nursing (New York, NY). 2014;35(2 Suppl):S37-44. doi: 10.1016/j.gerinurse.2014.02.018.
9. Centers for Medicare and Medicaid Services. Home health compare 2019 [Available from: https://www.medicare.gov/homehealthcompare/search.html.
10. Solberg LI, Ohnsorg KA, Parker ED, Ferguson R, Magnan S, Whitebird RR, et al. Potentially preventable hospital and emergency department events: Lessons from a large innovation project. Perm J. 2018;22:17-102. doi: 10.7812/TPP/17-102.
11. Zolnoori M, McDonald MV, Barrón Y, Cato K, Sockolow P, Sridharan S, et al. Improving patient prioritization during hospital-homecare transition: Protocol for a mixed methods study of a clinical decision support tool implementation. JMIR research protocols. 2021;10(1):e20184. doi: 10.2196/20184.
12. Duffy FD, Gordon GH, Whelan G, Cole-Kelly K, Frankel R, Buffone N, et al. Assessing competence in communication and interpersonal skills: The kalamazoo ii report. Academic medicine : journal of the Association of American Medical Colleges. 2004;79(6):495-507. doi: 10.1097/00001888-200406000-00002.
13. Narayanan S, Georgiou PG. Behavioral signal processing: Deriving human behavioral informatics from speech and language: Computational techniques are presented to analyze and model expressed and perceived human behavior-variedly characterized as typical, atypical, distressed, and disordered-from speech and language cues and their applications in health, commerce, education, and beyond. Proceedings of the IEEE Institute of Electrical and Electronics Engineers. 2013;101(5):1203-33. doi: 10.1109/jproc.2012.2236291.
14. Marmar CR, Brown AD, Qian M, Laska E, Siegel C, Li M, et al. Speech-based markers for posttraumatic stress disorder in us veterans. Depression and anxiety. 2019;36(7):607-16. doi: 10.1002/da.22890.
15. Chakravarthula SN, Nasir M, Tseng S, Li H, Park TJ, Baucom B, et al., editors. Automatic prediction of suicidal risk in military couples using multimodal interaction cues from couples conversations. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020 4-8 May 2020.
16. Yu D, Deng L. Automatic speech recognition: Springer; 2016.
17. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine. 2012;29(6):82-97. doi: 10.1109/MSP.2012.2205597.
18. Song J, Zolnoori M, Scharp D, Vergez S, McDonald MV, Sridharan S, et al. Do nurses document all discussions of patient problems and nursing interventions in the electronic health record? A pilot study in home healthcare. JAMIA OPEN. 2022;5(2). doi: 10.1093/jamiaopen/ooac034.
19. Zolnoori M, Vergez S, Kostic Z, Sid R, Topaz M. Audio recording patient-nurse verbal communications in home health care settings: Pilot feasibility and usability study. JMIR Hum Factors. 2021;9(2). doi: 10.2196/35325.
20. Amazon Web Services. Amazon transcribe developer guide 2022 [Available from: https://docs.aws.amazon.com/transcribe/latest/dg/API_streaming_Item.html.

21. Ali A, Renals S, editors. Word error rate estimation for speech recognition: E-wer. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018: Association for Computational Linguistics.
22. Song J, Ojo M, Bowles KH, McDonald MV, Cato K, Rossetti S, et al. Detecting language associated with home health care patient's risk for hospitalization and emergency department visit. Nursing research. 2022. doi: 10.1097/NNR.0000000000000586.
23. Martin KS. The omaha system: A key to practice, documentation, and information management: Elsevier Saunders; 2005.
24. Pradeep KS, Gaur S, Prashant B, Manisha M, Atreya D. Unified medical language system. Electronic health record: Standards, coding systems, frameworks, and infrastructures: IEEE; 2013. p. 145-52.
25. Garla VN, Brandt C. Knowledge-based biomedical word sense disambiguation: An evaluation and application to clinical document classification. Journal of the American Medical Informatics Association. 2013;20(5):882-6. doi: 10.1136/amiajnl-2012-001350.
26. Wilcox AB, Hripcsak G. The role of domain knowledge in automating medical text report classification. Journal of the American Medical Informatics Association. 2003;10(4):330-8. doi: 10.1197/jamia.M1157.
27. Topaz M, Murga L, Bar-Bachar O, McDonald M, Bowles K. Nimbleminer: An open-source nursing-sensitive natural language processing system based on word embedding. Computers, informatics, nursing : CIN. 2019;37(11):583-90. doi: 10.1097/cin.0000000000000557.
28. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality 2013 [3111-9]. Available from: https://papers.nips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.
29. Topaz M, Murga L, Gaddis KM, McDonald MV, Bar-Bachar O, Goldberg Y, et al. Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. Journal of biomedical informatics. 2019;90:103103. doi: 10.1016/j.jbi.2019.103103.
30. Ma L, Zhang Y, editors. Using word2vec to process big text data. 2015 IEEE International Conference on Big Data (Big Data); 2015 29 Oct.-1 Nov. 2015.
31. McHugh ML. Interrater reliability: The kappa statistic. Biochemia medica. 2012;22(3):276-82. doi:
32. Topaz M, Woo K, Ryvicker M, Zolnoori M, Cato K. Home healthcare clinical notes predict patient hospitalization and emergency department visits. Nursing research. 2020. doi: 10.1097/nnr.0000000000000470.
33. Song J, Woo K, Shang J, Ojo M, Topaz M. Predictive risk models for wound infection-related hospitalization or ed visits in home health care using machine-learning algorithms. Advances in skin & wound care. 2021;34(8):1-12. doi: 10.1097/01.Asw.0000755928.30524.22.
34. Street RL, Jr., Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient Educ Couns. 2009;74(3):295-301. doi: 10.1016/j.pec.2008.11.015.
35. Waltl B, Bonczek G, Matthes F. Rule-based information extraction: Advantages, limitations, and perspectives 2018 [Available from: https://wwwmatthes.in.tum.de/file/47fs4e04rvtp/Sebis-Public-Website/-/Rule-based-Information-Extraction-Advantages-Limitations-and-Perspectives/Wa18b.pdf.
36. Li H. Deep learning for natural language processing: Advantages and challenges. National Science Review. 2017;5(1):24-6. doi: 10.1093/nsr/nwx110.
37. Liao J, Eskimez SE, Lu L, Shi Y, Gong M, Shou L, et al. Improving readability for automatic speech recognition transcription. J arXiv preprint arXiv:04438. 2020. doi:
38. Badal VD, Nebeker C, Shinkawa K, Yamada Y, Rentscher KE, Kim HC, et al. Do words matter? Detecting social isolation and loneliness in older adults using natural language processing. Frontiers in psychiatry. 2021;12:728732. doi: 10.3389/fpsyt.2021.728732.
39. Sanz C, Carrillo F, Slachevsky A, Forno G, Gorno Tempini ML, Villagra R, et al. Automated text-level semantic markers of alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring. 2022;14(1):e12276. doi: 10.1002/dad2.12276.
40. Al-Aynati MM, Chorneyko KA. Comparison of voice-automated transcription and human transcription in generating pathology reports. Archives of pathology & laboratory medicine. 2003;127(6):721-5. doi: 10.5858/2003-127-721-covtah.
41. Flanagan ME, Militello LG, Rattray NA, Cottingham AH, Frankel RM. The thrill is gone: Burdensome electronic documentation takes its toll on physicians' time and attention. Journal of general internal medicine. 2019;34(7):1096-7. doi: 10.1007/s11606-019-04898-8.
42. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, et al., editors. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. Mayo Clinic proceedings; 2016: Elsevier.