A Randomized Trial of Voice-Generated Inpatient Progress Notes: Effects on Professional Fee Billing

Andrew A. White[1] Tyler Lee[1] Michelle M. Garrison[2] Thomas H. Payne[1]

[1] Department of Medicine, University of Washington School of Medicine, Seattle, Washington, United States
[2] Department of Health Services, University of Washington School of Public Health, and Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, Seattle, Washington, United States

Address for correspondence Andrew A. White, MD, Department of Medicine, University of Washington School of Medicine, Seattle, WA 98195, United States (e-mail: andwhite@uw.edu).

Appl Clin Inform 2020;11:427–432.

Abstract

Background Prior evaluations of automated speech recognition (ASR) to create hospital progress notes have not analyzed its effect on professional revenue billing codes. As ASR becomes a more common method of entering clinical notes, clinicians, hospital administrators, and payers should understand whether this technology alters charges associated with inpatient physician services.

Objectives This study aimed to measure the difference in professional fee charges between using voice and keyboard to create inpatient progress notes.

Methods In a randomized trial of a novel voice with ASR system, called voice-generated enhanced electronic note system (VGEENS), to generate physician notes, we compared 1,613 notes created using intervention (VGEENS) or control (keyboard with template) created by 31 physicians. We measured three outcomes, as follows: (1) professional fee billing levels assigned by blinded coders, (2) number of elements within each note domain, and (3) frequency of organ system evaluations documented in review of systems (ROS) and physical exam.

Results Participants using VGEENS generated a greater portion of high-level (99233) notes than control users (31.8 vs. 24.3%, $p < 0.01$). After adjustment for clustering by author, the finding persisted; intervention notes were 1.43 times more likely (95% confidence interval [CI]: 1.14–1.79) to receive a high-level code. Notes created using voice contained an average of 1.34 more history of present illness components (95% CI: 0.14–2.54) and 1.62 more review of systems components (95% CI: 0.48–2.76). The number of physical exam components was unchanged.

Conclusion Using this voice with ASR system as tested slightly increases documentation of patient symptom details without reliance on copy and paste and may raise physician charges. Increased provider reimbursement may encourage hospital and provider group to offer use of voice and ASR to create hospital progress notes as an alternative to usual methods.

Keywords
- encounter notes
- billing
- speech recognition
- inpatient care

received
November 18, 2019
accepted
May 1, 2020

DOI https://doi.org/10.1055/s-0040-1713134.
ISSN 1869-0327.

Background and Significance

U.S. hospitals have broadly adopted electronic health records (EHRs) for clinician documentation.[1] Although electronic documentation benefits patients, providers, and health systems,[2,3] some clinicians find manual note entry with keyboard and mouse dissatisfying and time consuming.[4,5] In response, many clinicians use templates or copy and paste to save time.[6] Unfortunately, copy and paste may propagate outdated, unnecessary information, creating safety problems, and hindering auditors' and professional fee coders' ability to assess the correct evaluation and management (E/M)[a] code.[7,8] Using voice dictation with automatic speech recognition (ASR) offers an alternative mechanism for rapid document creation[9,10] and its use is growing.[11] Although ASR may reduce misuse of copy and paste, little is known about the effects of voice with ASR on inpatient E/M coding used for professional fee billing. Instead, prior evaluations of ASR have analyzed document turnaround time, accuracy, and physician task efficiency.[11] A deeper understanding of how voice with ASR affects professional fee coding could influence decisions about ASR system implementation, particularly if physicians who adopt ASR would generate lower billing charges.

We developed a new approach to write general medical inpatient progress notes with voice and ASR called voice-generated enhanced electronic note system (VGEENS).[12] VGEENS used a smartphone application to record note dictation and create a draft note in the EHR for physician review. The system included links to EHR data to import patient data in response to verbalized commands. It differed from commercial note writing systems available at the time in its portability, suitability to hospitalist workflow, and interaction with the EHR to insert patient data. A randomized controlled trial comparing VGEENS to manual note entry did not find superior outcomes in the domains of note timeliness, quality (as measured by PQRI-9), and physician satisfaction.[13] When notes were dictated soon after rounds,[b] notes were available in the EHR within 10 minutes.

Objectives

In this secondary analysis of a randomized trial, we measured the effect of VGEENS on E/M coding associated with inpatient notes. We also sought to understand the mechanism for any observed difference in E/M codes by measuring the frequency of note components used in calculating the E/M code. Although professional fee coders attempt to apply codes consistently, there is some subjectivity in E/M assignment and the assigned code can be influenced by small differences in how physicians describe elements of the patient visit. Because copy and paste may facilitate inflating the E/M coding level,[14] we hypothesized that note dictation with VGEENS would reduce the portion of notes coded as 99233–the highest of three levels for hospital progress notes–relative to a control group using standard manual note entry techniques.

Methods

Setting, Participants, and Intervention

This study was conducted on the inpatient general medicine services at University of Washington (UW) Medical Center and Harborview Medical Center, teaching hospitals of the UW using Cerner Millennium EHR (Cerner Corp., Kansas City, Missouri, United States).[15] The population of patients cared for on the medicine services includes those with advanced heart, renal, and hepatic failure, solid organ transplant candidates and recipients, and underserved populations including homeless and those with complications of HIV and substance use disorders. For all patients on the study services during the year of this study, the average length of stay was 5.93 days and the case mix index was 1.56. Prior to the VGEENS trial, progress notes were typed into templates that automatically import patient-specific data such as laboratory results. All progress notes are assigned E/M codes by professional coders employed by the faculty practice group and blinded to whether notes were created by physicians in the intervention or control group.

We conducted a randomized controlled trial over 8 months in 2016. Participants were randomized to either the intervention (VGEENS) or control (keyboard and template) group. The primary outcomes of interest related to the time required to complete notes, physician satisfaction with the note writing process, and measurements of note quality. Detailed descriptions of the VGEENS randomized trial and selected system outcomes have been published elsewhere.[12,13,16] In brief, physicians used VGEENS during or after rounds to record a dictation on a cell phone with an Android application. On a server, the digitally recorded dictation was converted to text using ASR (Dragon Medical Practice Edition, Nuance) and automatically edited. Scripts were used to insert requested patient data, format the note, and to send the note to the EHR inbox. From the EHR inbox, the physician could further edit and sign the document. Participants could not make their own commands in the software.

The intent of the trial was to enroll 30 participants who would generate 140 notes each, which was estimated to detect an 8% reduction in note completion time with an interclass correlation of 0.01 and a power of 0.80. Power calculations were not performed for comparison of the E/M code comparison. For the primary endpoints, the enrollment target was met, but fewer notes were generated by participants than sought because of scheduling of patient care duties and because some participants stopped using the system before completion of the trial.

All internal medicine residents and attending hospitalist physicians at the study sites were contacted through

[a] In the United States, for a given encounter, the selection of the appropriate level of evaluation and management (E/M) services is determined according to the code of definitions in the American Medical Association's Current Procedural Terminology (CPT) book and any applicable documentation guidelines. For hospital physician professional fees, 99231 99232, 99233 are used for subsequent care progress notes. Professional fees are lowest for 99231, intermediate for 99232, and highest for 99233. See reference 8 for more detail.

[b] Hospital physicians make at least daily visits to their hospitalized patients; these visits are referred to as "rounds."

meetings and e-mail messages and invited to participate in a trial of VGEENS. Forty-nine physicians volunteered to participate in the study. After a description of the study, physicians who consented to participate were randomly assigned to either the intervention or control group in a 1:1 ratio.

Because some randomized participants were not on a medical service rotation in which their responsibilities included writing daily progress notes during the study period they did not contribute notes. This situation was more common in the intervention group than in the control group. There were 13 intervention participants and 18 control participants who contributed at least one note to the study. Participants in the intervention arm contributed a median of 26 notes (interquartile range: 42.5). No changes were made to the trial outcomes after the trial commenced. No interim analysis was performed, and the trial was ended due to the completion of funding. Regarding the outcomes reported here, the participants and coders assigning E/M codes were not aware the notes were part of a trial and were blinded to the assignment; the codes were assigned in the course of usual medical care billing. The participants were not blinded to their arm of the trial.

Outcomes

In this secondary analysis, we compared intervention and control notes using the following three outcomes: (1) professional fee billing levels (99231 [level 1], 99232 [level 2], 99233 [level 3]) assigned by blinded coders, (2) number of note components provided within each note domain necessary for E/M billing (e.g., quality and severity are among elements counted in the "history of present illness"),[8] and (3) number of organ systems documented within the review of systems and physical exam. Outcomes 2 and 3 provide greater details to show why notes were assigned the E/M level (level 1, 2, or 3).

Data Extraction and Analysis

Notes were included based on the primary E/M codes assigned on that date of service. Notes were excluded if they lacked a code or were coded as admission notes, discharge summaries, or critical care progress notes. Notes without an attributed author were excluded from multivariate analysis, as this information was required for regression analysis. The complete, author-attributed text of control notes was available for 100 randomly selected notes for this secondary analysis; as per institutional review board requirements in the primary analysis, note text was deidentified if not selected for primary analysis. However, E/M codes were available for all notes.

We used a commercially available computer-assisted coding tool (nCode, Cerner Corp, Kansas City, Missouri, United States) to assess the presence of note components that contribute to E/M code determination.[17] nCode analyzes progress notes for attributes of history, physical, and medical decision making. For example, the software would analyze "regular rate and rhythm, no edema" as evidence of a cardiac exam. Evaluation of nCode has shown that nCode and the human coding experts agreed 90% of the time.[18] Outcomes calculated using nCode include count of components in the history of present illness (HPI score, eight possible components), review of systems (ROS score, 13 possible components, excluding "allergies" that were uniformly documented due to listing drug allergies in the template), exam (exam score, 31 possible components), and whether the exam was categorized as "comprehensive," "detailed," "expanded problem focused," "problem focused," or "none." nCode was not used to quantify subcomponents of medical decision making because it is not optimized for inpatient templates that may inflate E/M coding; human coders were regarded as the gold standard for this subjective element.

Analysis included descriptive statistics and regression models. Here, the individual chart note is the level of analysis, as we have multiple chart notes per note author and note characteristics are likely to be more similar within versus between authors, clustering effects need to be taken into account in the statistical method. As a result, each regression model used cluster-robust standard errors, which use the robust variance estimator to take into account this clustering in the calculation of standard errors (and hence 95% confidence intervals [CI] and p-values) and the variance–covariance matrix.[19–23] Bivariate linear regression models examined the impact of intervention group status on three nCode-derived outcomes as follows: HPI components, ROS components, and exam components; bivariate logistic regression examined the intervention impact on whether the physical exam was coded as "comprehensive" or "detailed" rather than "problem focused," "expanded problem focused," or "none"; and multinomial logistic regression examined intervention impact on the professional fee current procedural terminology (CPT) code assigned and billed by the hospital, with the most common code (99232) chosen as the referent group.

Results

From the original corpus of 1,852 notes from the VGEENS trial, 239 were excluded per protocol (►Fig. 1). Of the 1,613 analyzed notes, 690 were eligible for multinomial logistic analysis, including 603 in the intervention group, and 87 in the control group. An additional 901 notes lacking author attribution were included in descriptive statistics (►Table 1). In crude analyses, VGEENS users generated a greater portion of high-level (99233) notes than control users (31.8 and 24.3%, respectively, $p < 0.01$). After adjustment for clustering by author, the finding persisted in the multinomial logistic regression (►Table 2). Intervention group notes were 1.43 times more likely (95% CI: 1.14–1.79) than control notes to be assigned a high-level (99233) code, and 0.74 times less likely (95% CI: 0.46–1.20) to be assigned a lower-level code (99231), with 99232 notes as the referent group.

Intervention notes were given credit for an average of 1.34 more HPI components (95% CI for the difference 0.14–2.54) and 1.62 more ROS components (95% CI: 0.48–2.76), with no significant difference in the physical exam components. We also did not find statistically significant changes in the portion of exams coded as "comprehensive" or "detailed," rather than "problem focused" or "expanded problem focused" ($p = 0.22$). In a post hoc exploration to generate

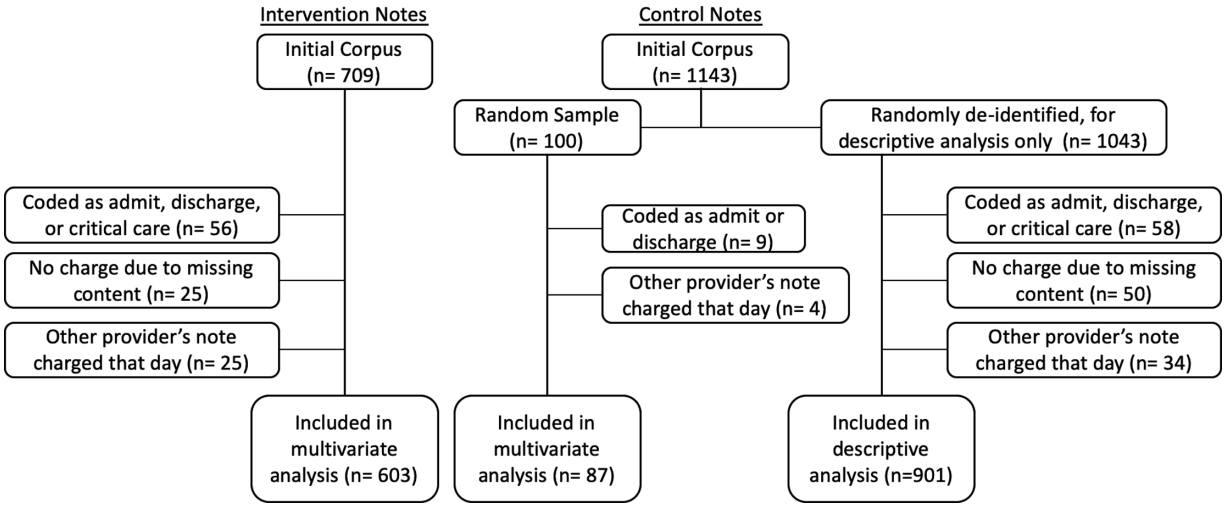

Fig. 1 Notes included or excluded in analysis.

Table 1 CPT codes assigned to inpatient progress notes in the VGEENS trial

CPT code	Control notes ($n = 988$) *n* (%)	Intervention notes ($n = 603$) *n* (%)	Total ($n = 1591$) *n* (%)
99231	60 (6.1)	25 (4.1)	85 (5.3)
99232	688 (69.6)	386 (64.0)	1074 (67.5)
99233	240 (24.3)	192 (31.8)	432 (27.2)

Abbreviations: CPT, current procedural terminology; VGEENS, voice-generated enhanced electronic note system.

hypotheses, ►Fig. 2 displays individual exam and ROS components by relative prevalence in intervention and control notes. The figure highlights the relative increase in documentation of constitutional, respiratory, and gastrointestinal ROS in intervention compared with control notes. All other exam and ROS components had less than 20% variation between intervention and control.

Discussion

This analysis of data from a randomized trial shows that notes created using VGEENS during or soon after rounds contain modestly more detailed history and ROS components. As a result, use of ASR as tested was associated with higher E/M coding. This data refutes our hypothesis that clinicians would document fewer 99233 notes than if they relied on template and keyboard. Unlike our finding about patient history, physicians using VGEENS did not document more detailed physical exam findings and may have decreased documentation of the neurologic exam. These results add to our findings that VGEENS, as a whole, had no effect on note timeliness or physician satisfaction[12,14] but indicate that ASR, as tested, may nonetheless subtly affect the content of progress notes. This study offers reassurance that portable ASR solutions similar to VGEENS are unlikely to cause a decrease in revenue. Because U.S. regulators had noticed increased professional fee billing associated with templates and copy/paste,[18] we hypothesized that voice with ASR might reduce that trend. We did not

Table 2 Regression: modeling intervention effects on progress note coding

Outcome	Effect size	Significance	
Linear regression models ($n = 690$)	**Beta**	**95% CI**	***p*-Value**
Model 1: HPI score	1.34	0.14–2.54	0.03
Model 2: ROS score	1.62	0.48–2.76	0.01
Model 3: exam score	0.09	−1.68–1.86	0.92
Logistic regression model ($n = 690$)	**OR**	**95% CI**	**p**
Model 4: exam coded as "comprehensive" or "detailed" rather than "problem focused," "expanded problem focused," or "none"	2.08	0.65–6.66	0.22
Multinomial logistic regression model ($n = 1,591$)	**RRR**	**95% CI**	***p*-Value**
Model 5: professional fee CPT code assigned and billed 99231 99232 99233	0.74 (ref) 1.43	0.46–1.20 1.14–1.79	0.23 <0.01

Abbreviations: CI, confidence interval; CPT, current procedural terminology; HPI, history of present illness; OR, odds ratio; ROS, review of system.
Note: Standard errors were adjusted for clustering on note author in all regression models.

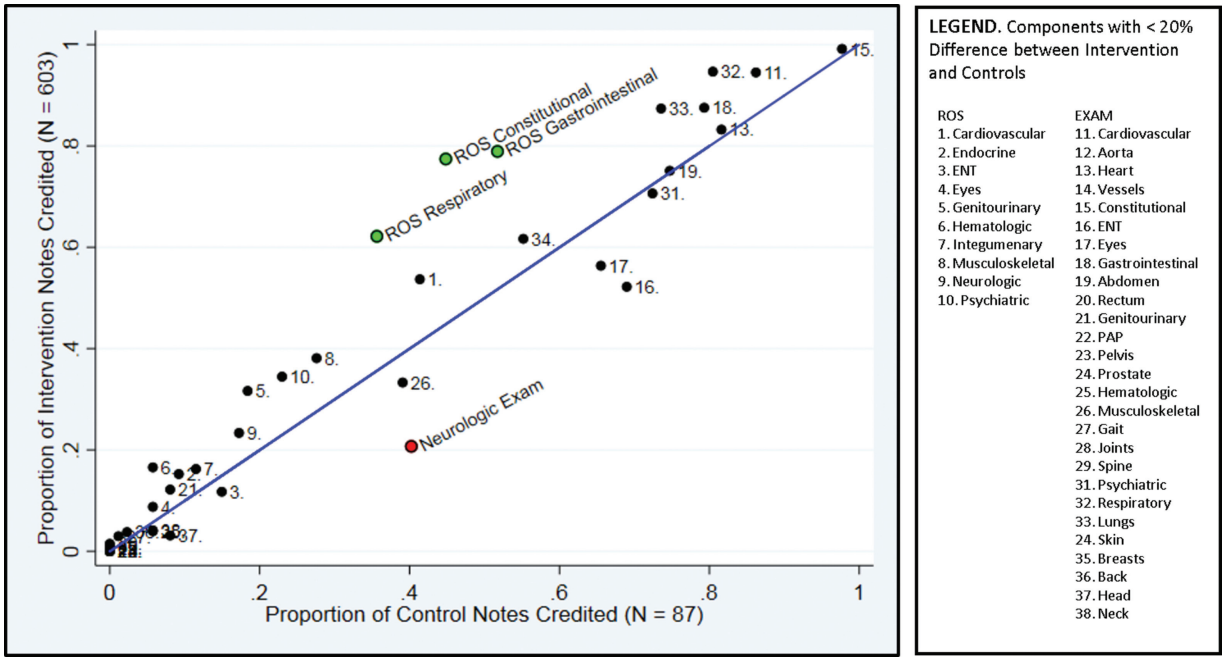

Fig. 2 Intervention Impact on ROS and exam coding. ROS, review of system.

find that to be the case and this raises questions of whether ASR adoption could affect healthcare costs by escalating physician billing. For an average hospital in the United States, we estimate that the shift in billing found in this manuscript would result in several tens of thousands of dollars of additional practice revenue, which would be significant enough to influence investment decisions made by practice managers.

A next step for research into the effect of ASR on physician documentation would be to complement this single-center randomized trial with a retrospective analysis of a large number of billing outcomes from a multicenter cohort of providers using voice with ASR. There are other advantages to using voice, including reduced incentive to use copy/paste and precompleted templates to create notes. However, it is possible that an optimal system would combine elements of voice with ASR, templates completed by physicians, information provided by patients, and copy forward. Each documentation modality has limitations that might be offset by software that minimizes physician effort without compromising document quality.

Strengths and Limitations

This study has notable strengths, including randomized design which should mitigate bias including patient acuity, the use of blinded coders, analytic methods to take into account clustering at the note author level. However, the study has important limitations. First, the findings may not be generalizable beyond the academic setting, more broadly to all forms of ASR, outside the United States, or to U.S. practices in which physicians code their own notes with E/M codes. However, we suspect that they are generalizable beyond VGEENs since VGEENs used a commonly used ASR engine, and there are now commercially available systems with functionality similar to VGEENs. Second, the data protections enacted for studying the primary outcomes prevented use of the entire note corpus, as well as human verification, of ASR errors that may have affected billing. Third, we did not analyze elements of medical decision making that could offer deeper insight into why intervention notes were more likely to be assigned a code of 99233. Fourth, we used natural language processing which may not be as accurate as human note assessment. However, the researchers did not have the resources for manual human counting of history and exam component and sought to mitigate this limitation of natural language processing by applying it to the most discrete and objective note components. Lastly, although we used a randomized trial to limit confounding, we did not measure severity of illness or other demographic characteristics at the patient level and cannot exclude the possibility that unmeasured patient variables account for the higher complexity of care documented in the VGEENS arm of the trial.

Conclusion

VGEENS appears to modestly affect physician documentation habits that affect E/M coding. Although decisions to implement or upgrade ASR solutions are likely influenced by many other factors, including physician preference and system acquisition costs, we offer evidence that efforts to help physicians diminish reliance on cut and paste do not diminish practice revenue. Because ASR adoption is growing, well-powered studies of its impact on health care costs and quality are warranted.

Clinical Relevance Statement

Physicians are increasingly using dictation with automatic speech recognition (ASR) to create inpatient progress notes. In a randomized trial of a handheld ASR system versus note

entry with keyboard and mouse, physicians included more information about patient symptoms and generated notes with higher professional fees. This may encourage physicians and hospital administrators to invest in similar systems.

Multiple Choice Questions

1. In a randomized trial of a smartphone-based progress note dictation system, how were notes authored by physicians using the intervention billed in comparison to notes written with keyboard and mouse?
   a. No difference was found.
   b. A higher proportion of low complexity notes were billed (99231).
   c. A higher proportion of medium complexity notes were billed (99232).
   d. A higher proportion of high-complexity notes were billed (99233).

   Correct Answer: The correct answer is option d.

2. In a randomized trial of a smartphone-based progress note dictation system, what changes were observed in note components in notes authored by physicians using the intervention billed in comparison to notes written with keyboard and mouse?
   a. Increased history and review of systems information.
   b. Additional social and family history.
   c. Decreased physical exam elements.
   d. Increase physical exam elements.

   Correct Answer: The correct answer is option a.

Protection of Human and Animal Subjects

This study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects and was reviewed by the University of Washington Institutional Review Board.

Conflict of Interest

None declared.

Acknowledgments

The authors would like to acknowledge K.K. Kailasam, Cerner Corporation for providing computer assisted coding using in this analysis.

References

1 Henry J, Pylypchuk Y, Searcy T, Patel V. Adoption of electronic health record systems among u.s. non-federal acute care hospitals: 2008–2015. ONC data brief, no.35, 2016. Available at: https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php. Accessed May 15, 2020
2 Amarasingham R, Plantinga L, Diener-West M, Gaskin DJ, Powe NR. Clinical information technologies and inpatient outcomes: a multiple hospital study. Arch Intern Med 2009;169(02):108–114
3 Chaudhry B, Wang J, Wu S, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med 2006;144(10):742–752
4 Friedberg MW, Chen PG, Van Busum KR, et al. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. Rand Health Q 2014;3(04):1
5 Sinsky CA, Willard-Grace R, Schutzbank AM, Sinsky TA, Margolius D, Bodenheimer T. In search of joy in practice: a report of 23 high-functioning primary care practices. Ann Fam Med 2013;11(03):272–278
6 Tsou AY, Lehmann CU, Michel J, Solomon R, Possanza L, Gandhi T. Safe practices for copy and paste in the EHR. Systematic review, recommendations, and novel model for health IT collaboration. Appl Clin Inform 2017;8(01):12–34
7 Abelson R, Creswell J. US warning to hospitals on Medicare bill abuses. Available at: http://www.nytimes.com/2012/09/25/business/us-warns-hospitals-on-medicare-billing.html. Accessed May 15, 2020
8 Centers for Medicare & Medicate Services, Department of Health and Human Services. Evaluation and management services guide. Available at: https://www.cms.gov/Outreach-and-Education/-Medicare-Learning-Network-MLN/MLNProducts/Downloads/eval-mgmt-serv-guide-ICN006764.pdf. Accessed May 24, 2019
9 Al Hadidi S, Upadhaya S, Shastri R, Alamarat Z. Use of dictation as a tool to decrease documentation errors in electronic health records. J Community Hosp Intern Med Perspect 2017;7(05):282–286
10 Johnson M, Lapkin S, Long V, et al. A systematic review of speech recognition technology in health care. BMC Med Inform Decis Mak 2014;14:94
11 Blackley SV, Huynh J, Wang L, Korach Z, Zhou L. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. J Am Med Inform Assoc 2019;26(04):324–338
12 Payne TH, Alonso WD, Markiel JA, et al. Using voice to create hospital progress notes: description of a mobile application and supporting system integrated with a commercial electronic health record. J Biomed Inform 2017;77:91–96
13 Payne TH, Alonso WD, Markiel JA, et al. Using voice to create inpatient progress notes: effects on note timeliness, quality, and physician satisfaction. JAMIA Open 2018;1(02):218–226
14 Office of Inspector General, Department of Health and Human Services. Not All Recommended Fraud Safeguards Have Been Implemented In Hospital EHR Technology. Available at: http://oig.hhs.gov/oei/reports/oei-01-11-00570.pdf. Accessed April 29, 2019
15 Payne TH, Perkins M, Kalus R, Reilly D. The transition to electronic documentation on a teaching hospital medical service. AMIA Annu Symp Proc 2006:629–633
16 Lybarger KJ, Ostendorf M, Riskin E, Payne TH, White AA, Yetisgen M. Asynchronous speech recognition affects physician editing of notes. Appl Clin Inform 2018;9(04):782–790
17 Payne TH, Garver-Hume A, Kirkegaard S, et al. Natural language processing improves coding accuracy. MGMA Connex 2011;11(09):15–17
18 Zigmond J.HHS, Justice Department warn hospitals on EHR-related payment fraud. Available at: https://www.modernhealthcare.com/article/20120924/NEWS/309249968/hhs-justice-department-warn-hospitals-on-ehr-related-payment-fraud. Accessed January 25, 2020
19 Obtaining robust variance estimates" Stata Corp. Stata 16 Base Reference Manual. College Station, TX: Stata Press; 2019
20 Correlated errors: Cluster–robust standard errors" Stata Corp. Stata 16 Base Reference Manual.College Station, TX: Stata Press; 2019
21 Graubard BI, Korn EL. Regression analysis with clustered data. Stat Med 1994;13(5–7):509–522
22 Watterson JL, Rodriguez HP, Aguilera A, Shortell SM. Ease of use of electronic health records and relational coordination among primary care team members. Health Care Manage Rev 2018. Doi: 10.1097/HMR.0000000000000222
23 McNeish D, Stapleton LM, Silverman RD. On the unnecessary ubiquity of hierarchical linear modeling. Psychol Methods 2017;22(01):114–140