Journal of Digital Imaging (2021) 34:1294–1301
https://doi.org/10.1007/s10278-021-00502-w

ORIGINAL PAPER

Radiology Dictation Errors with COVID-19 Protective Equipment: Does Wearing a Surgical Mask Increase the Dictation Error Rate?

Abiola Femi-Abodunde[1] · Kristen Olinger[1] · Lauren M. B. Burke[1] · Thad Benefield[1] · Ellie R. Lee[1] · Katrina McGinty[1] · Benjamin M. Mervak[1]

Received: 29 January 2021 / Revised: 1 August 2021 / Accepted: 4 August 2021 / Published online: 24 September 2021

Abstract
Our aim was to determine the effect of wearing a surgical mask on the number and type of dictation errors in unedited radiology reports. IRB review was waived for this prospective matched-pairs study in which no patient data was used. Model radiology reports ($n=40$) simulated those typical for an academic medical center. Six randomized radiologists dictated using speech-recognition software with and without a surgical mask. Dictations were compared to model reports and errors were classified according to type and severity. A statistical model was used to demonstrate that error rates for all types of errors were greater when masks are worn compared to when they are not (unmasked: 21.7 ± 4.9 errors per 1000 words, masked: 27.1 ± 2.2 errors per 1000 words; adjusted $p < 0.0001$). A sensitivity analysis was performed, excluding a reader with a large number of errors. The sensitivity analysis found a similar difference in error rates for all types of errors, although significance was attenuated (unmasked: 16.9 ± 1.9 errors per 1000 words, masked: 20.1 ± 2.2 errors per 1000 words; adjusted $p = 0.054$). We conclude that wearing a mask results in a near-significant increase in the rate of dictation errors in unedited radiology reports created with speech-recognition, although this difference may be accentuated in some groups of radiologists. Additionally, we find that most errors are minor single incorrect words and are unlikely to result in a medically relevant misunderstanding.

Keywords Dictation errors · Speech-recognition · Dictation software · Masks · Personal protective equipment · COVID-19

Introduction

The COVID-19 pandemic has presented many challenges to businesses across the world, including hospital systems, and has necessitated rapid changes to the daily practice of medicine. Public use of face masks has been one effective methods of source control recommended by the US Centers for Disease Control (CDC) [1]. Following this recommendation, most hospital systems have mandated the occupational use of masks to limit the spread of aerosols or droplets generated by activities like speaking [2–4].

In modern radiology practices, there is widespread use of speech-recognition dictation software as a means to generate radiology reports and assist with patient care. Although there have been significant advances in speech-recognition software over the last 30 + years, automated transcription of speech remains imperfect even in optimal situations, with varying reports on accuracy [5–10]. Prior studies have described the types of errors which can be introduced by speech recognition – wrong tenses, word substitutions, word omissions, nonsense/incomplete phrases, punctuation errors, incorrect measurements, laterality errors, and wrong dates, among others [7, 10] – and potentially confusing errors have been shown to occur in more than 20% of routine radiology reports dictated using speech-recognition software [10]. During the COVID-19 pandemic, a published study demonstrated a negative impact of personal protective equipment (PPE) on interpersonal healthcare communication in a clinical setting – including speech discrimination and understanding [11]. Experiments by Toscano and Nguyen et al. also illustrated the impact of mask-wearing on voice recognition at low and higher frequencies [12, 13]. Their results suggested that wearing masks does have varying effects on speech recognition. Toscano further demonstrated that there are varying degrees of sound dampening properties

✉ Benjamin M. Mervak
bmervak@med.unc.edu

[1] Department of Radiology, University of North Carolina School of Medicine, Chapel Hill, NC, USA

depending on the talker, level of background noise, and type of mask.

Radiologists may anecdotally feel that our accuracy has been affected by PPE, although to our knowledge, masks have an unknown effect on the accuracy of speech recognition and rate of dictation errors. The purpose of this study is to determine the effect of surgical masks on the number and type of dictation errors in unedited radiology reports.

Methods

Overview

A matched-pairs study design was used with no patient data included. A power analysis (detailed below) was conducted to plan the sample size, and a corresponding number of model radiology reports ($n = 40$) were created. Six participating radiologists used speech-recognition software to create dictations based on these model reports. Dictations ($n = 480$) were compared to the model reports and errors were manually tallied and classified according to type and severity. A statistical model was used to compare error rates for masked vs unmasked dictations. Before beginning the study, the Institutional Review Board (IRB) for our hospital system was consulted and determined that this project was exempt from a full review as no patient data was included.

Power Analysis

To determine the total number of dictations that would be required of our study participants, we conducted a power analysis using G*Power v3.1.9.2 (Heinrich Heine Universität, Düsseldorf, Germany) [14]. Existing literature was used as an estimate of the mean number of errors per report expected to occur when dictating without a mask (1.6 ± 1.1) [10]. The mean number of dictation errors with a mask was hypothesized to be 20% greater (1.9 ± 1.1). We found that with a matched-pairs study design, the upper bound would be 211 dictations in each group for 80% power at alpha = 0.05.

Generation of Model Reports and Dictations

Six radiologists agreed to participate as readers in the project: five attending diagnostic radiologists (four female and one male) each with at least eight years of experience dictating, and one female diagnostic radiology resident in her fourth year of postgraduate training (PGY-4) with more than 2 years of experience dictating.

To meet the target number of dictations, a total of 40 model radiology reports were fabricated by the radiology resident with oversight from one of the faculty radiologists, then validated by a second faculty radiologist to ensure that the reports approximated the structure and complexity commonly generated during a workday at our tertiary care center. No patient data was used. As the five participating attending radiologists were within the division of abdominal imaging and could be expected to have dictation voice models highly tuned to terms and conditions found in abdominal imaging reports, model reports were limited to varieties that would be reported by an abdominal imaging division. Reports were evenly balanced including ten each of computed radiography/radiofluoroscopy (CR/RF) reports, ultrasound (US) reports, computed tomography (CT) reports, and magnetic resonance imaging (MR) reports. Departmental structured templates served as a foundation for these reports, to which features were added including dates and times; indications; factitious comparisons; common, uncommon, and incidental imaging findings; biplanar/multiplanar measurements; and image/series numbers as commonly dictated at our institution. A variety of benign and malignant conditions were included. A summary of study indications and an example report are included in Fig. 1. The total number of words in each model report was counted to evaluate error rates per 1000 words.

Each radiologist was instructed to dictate word-for-word the contents of the 40 model reports twice: once while wearing a mask, and once without a mask, for a total of 80 dictations per reader and 480 dictations total. To control for bias from dictating the same reports twice, readers were randomized into two equal groups with one group dictating first masked then unmasked, and the other dictating first unmasked and then masked. Masks were provided to each radiologist by the radiology department as personal protective equipment and consisted of a standard disposable surgical mask attached to the face via elastic ear-loops. No N-95 masks were used. When dictating with a mask, participants were instructed to wear the mask tight to the face and fully cover both the nose and mouth.

To standardize the process of dictation, requirements for reading radiologists included: de-novo dictation of all section headers, words, numbers, dates, and punctuation exactly as written in the model report; no proofreading of reports during or after dictation (excepting an obvious manual error or accidental garbling of words due to something other than the mask itself); dictation at a natural pace, tone, and volume; dictation of all reports in the same physical location to minimize variation due to microphone, room noise, or other environmental factors. All reports were created using PowerScribe 360 v4.0-SP2 reporting software and a PowerMic III (Nuance Communications, Burlington, Massachusetts) and then copied directly from the reporting software into a separate text document. Radiologists used their own user account and associated personalized voice model, which had been attuned to their pattern of speech through daily use for

a

CT	MRI
• Abdominal pain • Trauma • Nephrolithiasis • Necrotizing pancreatitis • Occult cancer evaluation • Ovarian cancer follow-up • Complications post bowel resection • Hematuria (CT urogram) • Pancreatic cancer (resectability) • Cirrhosis / liver masses (multiphasic liver CT)	• Cirrhosis, liver mass on CT • Noncirrhotic, liver mass on US • Hepatocellular carcinoma post-ablation • Post liver transplant • Metastatic neuroendocrine tumor • Abdominal pain in pregnancy • Primary sclerosing cholangitis (MRCP) • Crohn's disease (MR enterography) • Fibroids, uterine artery embolization • Elevated PSA (prostate protocol) • Rectal cancer staging
US	**XR/FL**
• Nephrolithiasis • Abdominal pain • Cirrhosis, screening for liver lesions • Aortic aneurysm Screening • Thyroid nodules • Post liver transplant (doppler US) • Post kidney transplant (doppler US, plus native kidneys) • Hydrocele • Pelvic pain • Vaginal bleeding	• Enteric tube placement • Abdominal pain • Nausea / vomiting • Constipation (Sitz marker study) • Ureteral stent placement • Dysphagia • Chest pain • Intestinal leak • Esophageal perforation • Colovesical Fistula

b

DATE: 6/26/2020 2:11 PM

CLINICAL INDICATION: Staging of rectal adenocarcinoma.

COMPARISON: None.

FINDINGS:

The primary tumor is 9.6 cm from the anal verge, 4.8 cm from the top of the sphincter complex/anorectal junction, and measures 3.5 cm in length. Tumor straddles the anterior peritoneal reflection. There is approximately 1.5 cm of invasion through the muscularis propria and into the mesorectal fat, with focal abutment of the peritoneal reflection by tumor.

T stage on MRI is T4a.

No invasion of the genitourinary structures, pelvic sidewall, pelvic floor, sacrum, or nerves. However, note is made of extramural vascular invasion (EMVI) on 6:25.

No involvement of the anal sphincter complex.

Approximately 6 mesorectal lymph nodes measuring between 5 and 9 mm are present. These are suspicious in appearance, demonstrating a round morphology, heterogeneity, and indistinct margins. (6:15, 17, 19, 20, 22, 24). There is also a borderline IMA lymph node measuring 0.9 cm (6:10).

There is a T2 hyperintense collection along the lateral aspect of the left hip, most likely a bursitis.

IMPRESSION:

Rectal adenocarcinoma is staged as T4a, N0, Mx based on imaging. Note is made of focal abutment of the peritoneal reflection and EMVI.

Borderline enlarged 0.9 cm IMA lymph node, potentially metastatic.

Fig. 1 Summarized study indications for model reports (**a**) and example of a model report (**b**)

more than 2 years in each case. To simulate a real-world setting more closely, the dictation wizard was not run at the beginning of each dictation session, as this is not commonly done on a day-to-day basis.

Dictation Coding

Using a comparison feature in Microsoft Word (Microsoft Corporation, Redmond, WA) to highlight differences, model reports were compared side-by-side with dictations from the radiologists, and dictation errors were manually tallied and categorized by one attending radiologist and the participating PGY4 resident. Categories of errors (outcome variables) included: incorrect words, missing words, additional words, missing or incorrect phrases (defined as 3 + sequential words), incorrect terms of negation (e.g., errors in "no," "not," or "without"), sidedness errors, incorrect image numbers, incorrect measurements, incorrect dates/times, and punctuation errors.

Every error was counted, with no limit as to the maximum number of errors codified per report. Incorrect, missing, and additional-word errors were subclassified as minor, moderate, or major errors based on a subjective assessment of the potential to result in a clinically significant misunderstanding for the ordering provider or a future radiologist. Missing/incorrect phrases of 3 + words, errors in words of negation, sidedness errors, and incorrect measurements were all subclassified as being major errors that could result in a clinically significant misunderstanding. Incorrect image numbers and incorrect dates/times were all subclassified as being moderate errors which might result in misunderstanding.

All 480 dictations were codified, including 240 in the masked group and 240 in the unmasked group. To validate data coding and address inherent subjectivity, a selection of these dictations (20%; [96/480]) were separately coded by a second attending radiologist and were compared to the initial coding. The discrepancy rate was 6.3% (6/96).

Data Analysis

Graphical evaluation showed no evidence of overdispersion. Error rates were modeled for each outcome as a function of the presence/absence of a mask assuming a Poisson distribution with a log link. The number of words in each dictation report was included as an offset and the model controlled for the nuisance parameter of randomization order. The model included mixed effects to control for radiologist-level correlation and correlation within a study document. Predicted error rates per 1000 words were computed for the mask vs no mask group and compared using a *t*-test. *P*-values for these comparisons were adjusted using the false discovery rate method to control the Type 1 error rate [15].

Sensitivity and Subgroup Analyses

Following an initial data review, an approximately fourfold difference was seen in the total number of errors generated by one participant (1346 total errors for one reader vs. mean of 308 for other participants). This participant was notable for being the only trainee as well as the only participant having accented speech. Using the model above, predicted error rates per 1000 words were computed and compared for this individual vs. the other 5 readers for the "all errors," "major errors," "moderate errors," and "minor errors" outcomes variables. To reduce the potential for significant bias of study outcomes toward error patterns present for this individual, a sensitivity analysis using the same model described above was performed excluding this trainee.

A separate subgroup analysis was conducted to evaluate whether modality was associated with the "all errors" outcome variable. We implemented the same model described above with the addition of a modality indicator variable. Predicted error counts were computed for each modality and compared using a *t*-test. *P*-values for these comparisons were adjusted using the false discovery rate method to control the Type 1 error rate.

Results are described using model-based error rates per 1000 words with standard errors and associated adjusted *p*-values [15]. Adjusted *p*-values < 0.05 are considered statistically significant.

Results

When analyzing outcomes for all participants, the overall model-based error rate (per 1000 words) in reports dictated without masks was 21.7 ± 4.9 and with masks was 27.1 ± 6.0, a difference of 25% (adjusted $p < 0.0001$). Significant differences were also seen in the error rates for major errors (5.6 ± 1.6 unmasked vs. 7.3 ± 2.0 masked; $p = 0.008$), minor errors (11.9 ± 2.6 unmasked vs. 15.2 ± 3.2 masked; adjusted $p = 0.0002$), punctuation errors (0.4 ± 0.3 unmasked vs. 0.7 ± 0.6 masked; adjusted $p < 0.0001$), missing-word errors (3.5 ± 0.9 unmasked vs. 4.3 ± 1.1 masked; adjusted $p = 0.049$), and errors involving terms of negation (0.1 ± 0.05 unmasked vs. 0.2 ± 0.1 masked; adjusted $p = 0.018$). Significant differences were also seen in subsidiary outcomes including incorrect-word errors of major severity (3.9 ± 1.0 unmasked vs. 5.0 ± 1.3 masked; adjusted $p = 0.044$) and missing-word errors of moderate severity (0.3 ± 0.2 unmasked vs. 0.6 ± 0.4 masked; adjusted $p = 0.001$).

A significant difference was seen in the error rate for the one trainee participant for the "all errors," "major errors," "moderate errors," and "minor errors" outcomes variables (all $p < 0.0001$). Outcomes for the subgroup of

five attending radiologists differed from outcomes for the group inclusive of the radiologist in training.

The overall model-based error rate (per 1000 words) for the subgroup consisting of only attending radiologists was 16.9 ± 1.9 in reports dictated without masks and 20.1 ± 2.2 when wearing a mask, a difference of 19%; this difference was borderline significant (adjusted $p = 0.054$). Incorrect-word errors and the subsidiary outcome of incorrect-word errors of minor severity were also marginally significant (adjusted $p = 0.054$ and 0.066, respectively). Other types of dictation errors did not occur at a significantly different rate when wearing a mask versus when dictating unmasked.

The most frequent types of errors encountered were: incorrect word errors, with a model-based error rate of 14.3 ± 2.7 per 1000 words when unmasked and 15.9 ± 2.9 when masked; missing a word, with a model-based error rate of 3.5 ± 0.9 per 1000 words when unmasked and 4.3 ± 1.1 when masked; and mistakenly added words, with a model-based error rate of 1.7 ± 0.4 per 1000 words when unmasked and 2.0 ± 0.4 when masked. Errors in numerals (i.e., measurements, image numbers, or dates) were less frequent, with a total model-based error rate of 1.1 ± 0.4 per 1000 words when unmasked and 1.3 ± 0.5. Details of the model-based error rates per 1000 words for all outcome variables for the entire group of participants are listed in Table 1, and for the subgroup of attending radiologists in Table 2.

An analysis of the effect of modality on the all-type error rate revealed that MR and CR had significantly higher error rates than CT. When evaluating all participants, MR also had significantly more errors than US, although this difference was not significantly different in the attending radiologist subgroup. Other pairwise comparisons did not reach statistical significance for either group. Error rates per 1000 words by modality, pairwise comparisons of modalities, and adjusted p values are listed for the primary group and attending subgroup in Table 3.

Discussion

In the COVID-19 era, the use of facial coverings at workplaces is necessary for reducing aerosolized particles and is typically mandated in a hospital setting. However, masks add complexity to a radiologist's daily work practices, and it is important to better understand the effects of this physical barrier on the accuracy of speech recognition. Although the specific ways masks might affect dictation accuracy were not assessed in this study, we would hypothesize that this could be due to a combination of

Table 1 Model-based error rates for all outcome variables for all participating radiologists ($n = 6$), with associated model-based p-values; "a" indicates results presented as model-based error rates per 1000 words ± standard error

Error Rates - All Participants			
Error Type (Severity)	**Without Mask [a]**	**With Mask [a]**	**P-Value**
Incorrect Word (All)	14.3 ± 2.7	15.9 ± 2.9	0.11
Incorrect Word (Minor)	7.2 ± 1.2	8.5 ± 1.4	0.067
Incorrect Word (Moderate)	2.6 ± 0.9	2.2 ± 0.8	0.27
Incorrect Word (Major)	3.9 ± 1.0	5.0 ± 1.3	**0.044**
Missing Word (All)	3.5 ± 0.9	4.3 ± 1.1	**0.049**
Missing Word (Minor)	2.7 ± 0.6	3.0 ± 0.7	0.37
Missing Word (Moderate)	0.3 ± 0.2	0.6 ± 0.4	**0.001**
Missing Word (Major)	0.4 ± 0.1	0.4 ± 0.1	0.99
Additional Word (All)	1.7 ± 0.4	2.0 ± 0.4	0.39
Additional Word (Minor)	1.1 ± 0.2	0.1 ± 0.2	0.58
Additional Word (Moderate)	0.3 ± 0.08	0.5 ± 0.1	0.27
Additional Word (Major)	0.02 ± 0.02	0.07 ± 0.03	0.27
Erroroneous Phrase (Major)	0.6 ± 0.3	0.8 ± 0.3	0.29
Error in Term of Negation (Major)	0.1 ± 0.05	0.2 ± 0.1	**0.018**
Numeric Errors (All)	1.1 ± 0.4	1.3 ± 0.5	0.27
Incorrect Measurement (Moderate)	0.2 ± 0.1	0.3 ± 0.2	0.33
Incorrect Image Number (Moderate)	0.3 ± 0.1	0.4 ± 0.2	0.27
Incorrect Date/Time (Moderate)	0.3 ± 0.2	0.4 ± 0.2	0.63
Punctuation Errors (Minor)	0.4 ± 0.3	0.7 ± 0.6	**< 0.0001**
All Minor Errors	11.9 ± 2.6	15.2 ± 3.2	**0.0002**
All Moderate Errors	4.1 ± 1.3	4.3 ± 1.4	0.63
All Major Errors	5.6 ± 1.6	7.3 ± 2.0	**0.008**
All Errors	21.7 ± 4.9	27.1 ± 6.0	**< 0.0001**

Table 2 Error rates for all outcome variables for subgroup of attending radiologists ($n=5$), with associated model-based p-values; "a" indicates results presented as model-based error rates per 1000 words ± standard error

Error Rates—Attending Subgroup			
Error Type (Severity)	**Without Mask [a]**	**With Mask [a]**	**P-Value**
Incorrect Word (All)	10.9 ± 2.7	13.3 ± 1.4	*0.054*
Incorrect Word (Minor)	5.9 ± 0.9	7.5 ± 1.0	*0.066*
Incorrect Word (Moderate)	1.7 ± 0.4	1.6 ± 0.4	0.87
Incorrect Word (Major)	3.2 ± 0.9	4.1 ± 1.1	0.21
Missing Word (All)	2.9 ± 0.6	3.0 ± 0.7	0.86
Missing Word (Minor)	2.3 ± 0.5	2.2 ± 0.5	0.96
Missing Word (Moderate)	0.2 ± 0.2	0.3 ± 0.2	0.28
Missing Word (Major)	0.3 ± 0.1	0.3 ± 0.1	0.99
Additional Word (All)	1.4 ± 0.3	1.7 ± 0.3	0.74
Additional Word (Minor)	1.1 ± 0.3	1.1 ± 0.3	> 0.99
Additional Word (Moderate)	0.3 ± 0.08	0.4 ± 0.1	0.61
Additional Word (Major)	0.02 ± 0.02	0.06 ± 0.04	0.61
Erroroneous Phrase (Major)	0.4 ± 0.2	0.5 ± 0.2	0.49
Error in Term of Negation (Major)	0.08 ± 0.04	0.1 ± 0.05	0.74
Numeric Errors (All)	0.7 ± 0.2	0.9 ± 0.2	0.74
Incorrect Measurement (Moderate)	0.1 ± 0.08	0.2 ± 0.1	0.74
Incorrect Image Number (Moderate)	0.2 ± 0.1	0.2 ± 0.1	> 0.99
Incorrect Date/Time (Moderate)	0.2 ± 0.1	0.3 ± 0.1	0.74
Punctuation Errors (Minor)	0.3 ± 0.2	0.3 ± 0.2	0.74
All Minor Errors	9.8 ± 1.1	11.3 ± 1.3	0.2
All Moderate Errors	2.8 ± 0.6	3.1 ± 0.6	0.74
All Major Errors	4.3 ± 1.1	5.6 ± 1.3	0.11
All Errors	16.9 ± 1.9	20.1 ± 2.2	*0.054*

vocal dampening, impaired jaw/mouth motion when speaking, or generally being distracted by mask-wearing. To our knowledge, no published literature has evaluated the effects of surgical masks on dictation errors, as this was not a long-term issue before the current pandemic.

Table 3 Error rates and pairwise comparisons of "all errors" data subset, by modality; "a" indicates results presented as model-based error rates per 1000 words ± standard error

Modality	**Error Rate[a] (All Participants)**	**Error Rate[a] (Attending Subgroup)**
CT	19.7 ± 4.5	15.3 ± 2.0
MRI	29.3 ± 6.5	21.1 ± 2.6
US	21.9 ± 5.0	16.6 ± 2.4
XR	26.4 ± 6.1	22.2 ± 3.3
Comparison	**P-Value (All Participants)**	**P-Value (Attending Subgroup)**
CT vs MRI	**< 0.0001**	**0.043**
CT vs US	0.28	0.69
CT vs XR	**0.009**	**0.043**
MRI vs US	**0.003**	0.14
MRI vs XR	0.28	0.71
US vs XR	0.13	0.14

Interestingly, conclusions drawn from this study differ slightly depending on whether data from one reader with significantly higher error rates is considered. The precise reason(s) for this individual's high error rates is also outside of the scope of this study, although it is notable that this participant was the only resident in the study as well as the only individual with an accented pattern of speech. We believe that any conclusions drawn with this data included would be skewed toward error rates and patterns of errors for this individual, and that a subgroup analysis consisting of only the 5 attending radiologists is more broadly applicable to radiologists in the US. Data from this subgroup is used as the basis for discussion.

In the subgroup of attending radiologists, wearing a mask increased the overall error rate by approximately 19%, a finding which neared statistical significance ($p=0.054$). However, the majority of all errors were clinically inconsequential (minor errors) and most commonly the result of a single replaced word. Additional errors resulting from mask-wearing would therefore also be expected to be minor. On one hand, minor errors are a nuisance in that they can affect the perceived quality of radiology reports, including from a medicolegal standpoint, or put undue pressure on the reader when assessing the results of a radiologic study [16]. Transcription errors, irrespective of their effect on

interpretability, can also negatively affect the perception of the professionalism of its author and might affect a radiologist's professional relationships or referral patterns [16]. On the other hand, minor errors do not generally impact patient care, making them of far lesser clinical importance than moderate or severe errors. Moderate and severe errors—while of greater importance—occurred at a lower rate than minor errors, and no severity of error proportionally increased to a significant degree when wearing a mask.

This study also allows for an exploration of the relative incidence of different types of dictation errors encountered when using an automated dictation system. The most common type of error was the erroneous substitution of single words, which comprised 64–66% of all errors and was the only type of error to approach a significant difference when wearing a mask ($p = 0.054$). While single-word replacement errors might be of any severity, the majority were found to be minor, with many instances in which articles or conjunctions (e.g., "the," "of," or "and") were replaced with other articles/conjugations, or similar words were substituted (e.g., "duct" instead of "ductal" or "maximum" replacing "maximal"). Single missing words and single added words were the second and third most experienced errors. For all single-word error types, minor and clinically insignificant errors were predominant, and none were significantly affected by wearing a mask.

Errors involving numerals were infrequent, representing about 4% of errors, which is notable as radiologists are often tasked with measuring lesions, and may reference image numbers or measurements from prior studies when reviewing subsequent imaging. Missed or incorrect terms of negation (e.g., errors involving "no," "not," or "without") are also a constant worry for radiologists in that such errors are easily missed and can greatly affect patient management. These were also uncommon, representing about 0.5% of all errors. Again, none of these error types were significantly affected by a mask.

Significantly higher error rates were noted in MR and CR/RF compared to CT, while there was no significant difference between other pairwise comparisons of modalities. Specific reasons for these higher error rates were not explored in this study, although a hypothesis might include the presence of complex descriptions and pathologies on MR reports. The high error rate with CR/RF was unexpected and might be partially attributable to infrequently dictated terminology specific for fluoroscopy. It is worthwhile for radiologists to know which modalities may result in reports with more errors, as additional proofreading may be required when interpreting these imaging modalities.

There were several limitations to our study. First, participants were aware of the study objectives and were unable to be blinded to wearing or not wearing a mask. Next, radiologists were specifically instructed not to proofread their dictated reports, which is typically done before signing a clinical report during a workday. However, the goal of this study was to determine the effect of masks on dictation errors, and if editing were allowed, it would have been impossible to differentiate between errors resulting from a lack of editing versus errors resulting from a mask. As a result, the effect of masks on real-world reports in a patient's electronic medical record remains unknown, although would be expected to be lower than the number in these unedited reports due to proofreading.

There was also some inherent subjectivity in this study during the process of data coding, most importantly during subclassification of errors into minor, moderate, or major. Standardization was attempted to the extent possible by introductory meetings where data coding methods were discussed, including outlining definitions and reviewing examples for each severity classification, as well as by the process of data validation by a separate radiologist. Individual clinicians or radiologists may nonetheless differ in opinion as to what would constitute a minor versus moderate versus major error; this does not invalidate the conclusion that all-type error rates were significantly or near-significantly greater when radiologists wore masks.

Although we believe that these results apply to the majority of radiologists in the US, some factors may limit generalizability. In this study, we tested only one voice-recognition dictation system as it is the software available at our institution. However, the vendor used by our institution is the market leader, with an estimated 81% market share [17], and therefore findings apply to most radiologists. Minor variations might be expected for radiologists using other vendors for dictation software. Furthermore, demographics for participating radiologists may factor into generalizability. First, five participants in this study were female and one male. While gender is not expected to affect speech recognition, this question was not directly studied. More importantly, highly significant differences in error rates were seen between the attending radiologists and the single resident radiologist who also happened to be the only participant with accented speech. While the reasons for this were not specifically studied, we hypothesize that an accent may degrade voice recognition, although other conceivable reasons exist, for example, resident radiologists may not have as highly tuned dictation voice models as attendings given that residents dictate across multiple different subspecialties while they rotate through a radiology department. Of note, there were significantly more errors of negation for the resident radiologist, which would not be expected to depend upon the level of training or subspecialty. Implications of accented speech on the accuracy of speech recognition software have long been hypothesized, although current research is sparse. This may provide an interesting area for further study. Finally, this study focused on examinations generally

interpreted by abdominal imagers as participating faculty radiologists were within the division of abdominal imaging and would be expected to have dictation voice models highly tuned to terms and conditions found in abdominal imaging reports. If other types of exams (e.g., CT head, MR knee) were included, dictation errors could potentially have been due to weaker voice modeling for terms commonly found in those types of reports. Although not directly assessed, error rates for radiologists practicing using a personal voice model attuned to their study mix would not be expected to vary greatly.

We conclude that wearing a mask while dictating results in at least a near-significant increase in the rate of dictation errors in unedited radiology reports created with speech recognition, a difference which may be accentuated in some groups of radiologists. Notably, however, most errors are minor single incorrect words and are unlikely to result in a medically relevant misunderstanding.

Funding No funding was obtained for this study.

Data Availability Data is available on request from the authors.

Declarations

Ethics Approval The Institutional Review Board (IRB) for our hospital system was consulted and determined that this project was exempt from a full review as no patient data was included.

Consent to Participate All participants consented to participate in this study.

Consent for Publication All participants consent to publication of this manuscript.

Competing Interests No authors had conflicts of interest or competing interests.

References

1. CDC (2020) Coronavirus Disease 2019 (COVID-19). In: Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/cloth-face-cover-guidance.html. Accessed 12 Jul 2020
2. Leung NHL, Chu DKW, Shiu EYC, et al (2020) Respiratory virus shedding in exhaled breath and efficacy of face masks. Nat Med 26:676–680. https://doi.org/10.1038/s41591-020-0843-2
3. Anfinrud P, Stadnytskyi V, Bax CE, Bax A (2020) Visualizing Speech-Generated Oral Fluid Droplets with Laser Light Scattering. New England Journal of Medicine 382:2061–2063. https://doi.org/10.1056/NEJMc2007800
4. National Academies of Sciences E (2020) Rapid Expert Consultation on the Possibility of Bioaerosol Spread of SARS-CoV-2 for the COVID-19 Pandemic (April 1, 2020). National Academies Press (US)
5. Kanal KM, Hangiandreou NJ, Sykes AM, et al (2001) Initial evaluation of a continuous speech recognition program for radiology. J Digit Imaging 14:30–37. https://doi.org/10.1007/s10278-001-0022-z
6. Herman SJ (1995) Accuracy of a voice-to-text personal dictation system in the generation of radiology reports. AJR Am J Roentgenol 165:177–180. https://doi.org/10.2214/ajr.165.1.7785581
7. Hodgson T, Coiera E (2016) Risks and benefits of speech recognition for clinical documentation: a systematic review. J Am Med Inform Assoc 23:e169-179. https://doi.org/10.1093/jamia/ocv152
8. Madisetti V (2018) Video, Speech, and Audio Signal Processing and Associated Standards. CRC Press
9. Johnson M, Lapkin S, Long V, et al (2014) A systematic review of speech recognition technology in health care. BMC Med Inform Decis Mak 14:94. https://doi.org/10.1186/1472-6947-14-94
10. Quint LE, Quint DJ, Myles JD (2008) Frequency and Spectrum of Errors in Final Radiology Reports Generated With Automatic Speech Recognition Technology. Journal of the American College of Radiology 5:1196–1199. https://doi.org/10.1016/j.jacr.2008.07.005
11. Hampton T, Crunkhorn R, Lowe N, et al (2020) The negative impact of wearing personal protective equipment on communication during coronavirus disease 2019. J Laryngol Otol 134:577–581. https://doi.org/10.1017/S0022215120001437
12. Toscano JC, Toscano CM (2021) Effects of face masks on speech recognition in multi-talker babble noise. PLOS ONE 16:e0246842. https://doi.org/10.1371/journal.pone.0246842
13. Nguyen DD, McCabe P, Thomas D, et al (2021) Acoustic voice characteristics with and without wearing a facemask. Sci Rep 11:5651. https://doi.org/10.1038/s41598-021-85130-8
14. Faul F, Erdfelder E, Buchner A, Lang A-G (2009) Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 41:1149–1160. https://doi.org/10.3758/BRM.41.4.1149
15. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 57:289–300
16. Basma S, Lord B, Jacks LM, et al (2011) Error Rates in Breast Imaging Reports: Comparison of Automatic Speech Recognition and Dictation Transcription. American Journal of Roentgenology 197:923–927. https://doi.org/10.2214/AJR.11.6691
17. Speech Recognition in Radiology - State of the Market. In: Reaction Data. https://www.reactiondata.com/report/speech-recognition-in-radiology-state-of-the-market/. Accessed 31 Aug 2021