Assessing Quality of Care and Elder Abuse in Nursing Homes via Google Reviews

Jared Mowery[1], Amanda Andrei[1], Elizabeth Leeds Hohman[1], Jing Jian[1], Megan Ward[1]

1. The MITRE Corporation

Abstract

Background: It is challenging to assess the quality of care and detect elder abuse in nursing homes, since patients may be incapable of reporting quality issues or abuse themselves, and resources for sending inspectors are limited.

Objective: This study correlates Google reviews of nursing homes with Centers for Medicare and Medicaid Services (CMS) inspection results in the Nursing Home Compare (NHC) data set, to quantify the extent to which the reviews reflect the quality of care and the presence of elder abuse.

Methods: A total of 16,160 reviews were collected, spanning 7,170 nursing homes. Two approaches were tested: using the average rating as an overall estimate of the quality of care at a nursing home, and using the average scores from a maximum entropy classifier trained to recognize indications of elder abuse.

Results: The classifier achieved an F-measure of 0.81, with precision 0.74 and recall 0.89. The correlation for the classifier is weak but statistically significant: r = 0.13, P < .001, and 95% confidence interval (0.10, 0.16). The correlation for the ratings exhibits a slightly higher correlation: r = 0.15, P < .001. Both the classifier and rating correlations approach approximately 0.65 when the effective average number of reviews per provider is increased by aggregating similar providers.

Conclusions: These results indicate that an analysis of Google reviews of nursing homes can be used to detect indications of elder abuse with high precision and to assess the quality of care, but only when a sufficient number of reviews are available.

Keywords: social media, geriatric nursing, patient safety, natural language processing, supervised machine learning

Correspondence: jmowery@mitre.org

DOI: 10.5210/ojphi.v8i3.6906

Introduction

Studies of social media and healthcare phenomenon have explored a wide variety of applications, including a growing body of literature analyzing physician review websites (PRWs). Many PRWs contain reviews of physicians and medical facilities written by patients or relatives of patients, which often include both a text component and one or more numeric ratings.

The influence of PRWs is likely to grow over time. The number of PRW reviews has been increasing rapidly, with one study finding that Yelp reviews related to patient experiences grew at a rate of 1.5 times annually between 2007 and 2012 [1]. A survey of consumers found that 59% of respondents reported PRWs to be "somewhat important" or "very important", and that among consumers who sought online ratings, 35% reported selecting a physician based on good ratings and 37% reported avoiding a physician with bad ratings; meanwhile, 43% of respondents who did not seek online ratings reported a lack of trust for information on the websites [2]. A survey of 854 patients visiting a Pre-Operative Evaluation Clinic at Mayo Clinic in Minnesota showed that 84% had not visited a PRW, although 28% strongly agreed that a positive review alone would cause them to seek care from a provider, and 27% indicated a negative review would cause them to choose against using a provider [3]. The influence of PRWs on consumer decisions suggests that determining the aspects of care that are important to reviewers, and ascertaining the accuracy of information on PRWs regarding those aspects of care, would help consumers make more informed decisions.

Online reviews also create opportunities to improve the quality, safety, and efficiency of patient care, but only if accurate indicators can be extracted. Compared to paper surveys of patients or inspections of facilities, online reviews offer more timely, cost-effective information. However, online reviews are largely unstructured and are not subject to the same quality control measures as survey instruments, inspections, or clinical studies. This presents a need to better understand how online reviews relate to existing quality measures, such as the degree to which the online reviews are accurate, the number of reviews needed to achieve useful correlations with other quality measures, the aspects of care or customer service reflected in online reviews compared to existing quality measure data, and the efficacy of methods for extracting useful information from the text of an online review.

It is essential to validate the utility of online reviews as a measure of the technical quality of care. A survey of studies on PRWs observed that most information on PRWs is related to “structural quality and patient satisfaction” and is not risk-adjusted [4], which places doubt on whether PRWs accurately reflect a provider’s technical quality of care. Multiple studies have examined the relationship between online and offline patient reviews and quality measure data [5] for a variety of provider types. Multiple studies have examined hospitals based on a variety of PRWs and social media services, with many finding positive correlations or useful indicators, including the National Health Service (NHS) Choices website [6,7], Yelp [8-10], Twitter [11], two Korean web portals [12], and Facebook [13]. Studies focused on physicians have used NHS Choices [14], RateMDs.com [15,16], a set of nine PRWs [17], and two German PRWs [18].

The variety of study results suggests that finding sufficiently strong correlations between reviews and quality care measures is challenging, and that the degree of challenge varies as a function of the healthcare provider type, review type, PRW, and quantity of data available. For example, a

comparison of two German physician PRWs found that different correlations existed for each between reviews and other quality care measures [18].

Elder abuse in nursing homes is an area of particular concern, since it involves a vulnerable patient population, many of whom may be unable to report abuse. In addition, the technical quality of care may be difficult to assess from elderly patients' review data. A survey of 236 vulnerable elderly patients in two managed care institutions found that global ratings of care correlated with the quality of communication, but not with the technical quality of care [19]. A survey of 3,487 elderly patients at 18 general practices in England for treatment involving hypertension and influenza vaccinations also found no statistically significant correlation between patients' assessments and the technical quality of their primary care [20]. Although these two surveys did not use social media data, they suggest potential difficulties in evaluating elder care quality from patient reviews.

This paper examines nursing homes, using Google review ratings as well as a maximum entropy classifier trained to recognize indications of elder abuse in text. In each case, this study calculates the correlation with Centers for Medicare and Medicaid Services (CMS) inspection results from the Nursing Home Compare (NHC) data set [21].

Maximum entropy classifiers, a type of machine learning classifier, use hand-annotated data to learn how to classify their inputs as belonging to one of several output classes. In this case, the maximum entropy classifier learned to determine whether a Google review's text and rating were indicative of elder abuse, based on examples which a human being had labeled as either indicative or not indicative of elder abuse. Machine learning classifiers produce their estimate based on the presence or absence of features, such as consecutive pairs of words. For example, a review containing "soiled sheets", "call light", and "ignored calls" would be likely to indicate elder abuse, while a review containing "always attentive", "clean linens", and "polite staff" would be unlikely to indicate elder abuse.

This study uses the NHC data set, which contains CMS inspection results for nursing homes. The deficiencies found during inspections are categorized as either fire and safety deficiencies or health deficiencies. This study uses the health deficiencies as ground truth data. Therefore, for consistency, this study defines the "technical quality of care" as the nursing home's adherence to CMS' standards for care as represented by the set of health deficiencies in the NHC data. The health deficiencies cover a wide variety of potential problems, which include failing to maintain accurate clinical records, failing to grant patients access to their medical records, prohibiting patients from having visitors, failing to use licensed or certified staff, failing to notify family members of changes in a patient's condition, using unnecessary physical restraints, administering unnecessary medications, subjecting patients to abuse or physical punishment, failing to maintain a clean facility, giving incorrect medications to patients, failing to have a registered nurse on duty, and failing to ensure the call system is functional.

To our knowledge, this is the first study to use an automated analysis of online reviews for assessing the technical quality of care and the presence of elder abuse in nursing homes. Two recent studies are related to elder care providers. First, an exploratory study of Dutch social media data used search queries to locate 116 long-term elder care reviews relating to four safety risks and found that 72 reviews provided added value, according to inspectors from the Dutch

Healthcare Inspectorate [22]. Second, an analysis of 146 patients found that Nursing Home Compare (NHC) ratings of nursing homes do not include all aspects of care relevant to patients [23]. Since this study includes Google reviews obtainable for each provider without using keyword or phrase-based filtering, the reviews may also describe aspects of care not included in the NHC data set, which may limit the correlation achievable between the NHC data and online reviews.

Successfully extracting indications of the technical quality of care and the presence or absence of elder abuse from online review websites can benefit patients and the healthcare system in several respects, including (1) quantifying the utility of online reviews in measuring technical quality of care as opposed to other factors such as bedside manner, (2) aiding patients or family members of patients in choosing a facility, (3) helping CMS prioritize inspections of facilities to maximize the likelihood of uncovering and preventing abuse, and (4) supporting further studies analyzing the correlates of elder abuse to guide policy-makers.

Methods

Overview

This section describes the methods used to collect and analyze Google reviews and to correlate them with health deficiency data. The Data Collection section describes the NHC data set, as well as the methods for collecting Google reviews and splitting the review data into training and test sets. The Maximum Entropy Classifier section describes the definition of elder abuse developed in this study and the maximum entropy classifier which was trained to recognize elder abuse. The Ratings section discusses properties of the review ratings. Finally, the Correlations, Aggregations, and Analyses section discusses using correlation calculations to maximize statistical power, aggregating similar providers to extrapolate the correlation coefficients achievable with more reviews per provider, and performing several further analyses to understand factors which influence the results.

Data Collection

This study used the provider information and deficiencies spreadsheets from the Nursing Home Compare data set, updated December 17, 2015, as ground truth data. The deficiencies spreadsheet lists deficiencies reported by inspectors visiting CMS-certified nursing homes, including deficiencies that are likely to indicate elder abuse. The Deficiencies file includes 479,167 deficiencies for 15,584 providers. The deficiencies are split into 323,994 health deficiencies and 155,173 fire safety deficiencies. The data also includes inspection dates, correction dates, and other metadata, but this study uses only the counts of health deficiencies for each facility.

In addition to the facilities which received deficiencies, there are 77 facilities in the provider information spreadsheet which did not appear in the deficiencies data. These providers were assumed to have no deficiencies and are included in this study, resulting in a total of 15,661 providers.

The Google Maps API [24] provides a capability to search for a business by name and location, and returns up to 5 Google reviews, including the optional review text and a rating from 1 (worst) to 5 (best). When there are more than 5 reviews available for a nursing home, Google returns the 5 most helpful reviews. Querying the Google Maps API for each of the providers in the NHC data set yielded 16,160 reviews. Of those, 4,631 reviews did not include text and were discarded since the maximum entropy classifier requires text. This left 11,529 reviews spanning 5,516 providers. Since there were 15,661 total providers, 35.22% of providers in the NHC data set were included in this study.

A set of 2,500 reviews were hand-annotated to train the classifier, pulled from facilities which had a total number of deficiencies between 20 and 39, inclusive. The total deficiencies included both health and fire safety related deficiencies. Since reviews used for training were not included in testing and subsequent analysis, this choice preserved the providers with the greatest and least numbers of deficiencies for subsequent analysis. Hand-annotation of the 2,500 reviews resulted in 714 being marked as indicative of elder abuse (28.56%).

The distribution of providers, binned by the number of health deficiencies they received from CMS inspectors, was affected by the filtering used to generate the test set of providers (Figure 1). The first distribution shows all providers in the NHC data set, regardless of whether the providers had matching reviews. The second distribution includes only providers which had at least one review, even if that review lacked text. The third distribution counts providers only when they had at least one review which contained text, which is necessary for applying the maximum entropy classifier. This third distribution corresponds to the 35.22% of providers examined in this study. The training data for the classifier was selected from this distribution. The final distribution shows the providers after removing the providers which had at least one review included in the training data set: this final set of providers was used as test data for this study. Note that the provider bins are affected unevenly, since reviews for training were selected from providers whose total deficiency counts (including both health and fire safety deficiencies) were between 20 and 39, inclusive. Additionally, when a review was selected for training, any provider with a review having the same review text was removed—even if that specific review was not used for training—along with all reviews for that provider. This guaranteed that during testing, the classifier would never encounter a review whose text it had seen during training. Although additional providers and reviews could have been retained for the testing data set by including providers which had at least one review not used in training, doing so would have potentially skewed the test results in two ways:

- The classifier's score could have been biased if a provider's reviews had been split between the training and testing data, since reviews for a single provider could be conditionally dependent on one another due to shared features.
- The review count per provider in the testing data would have been disproportionately lower for providers whose reviews were used in training.

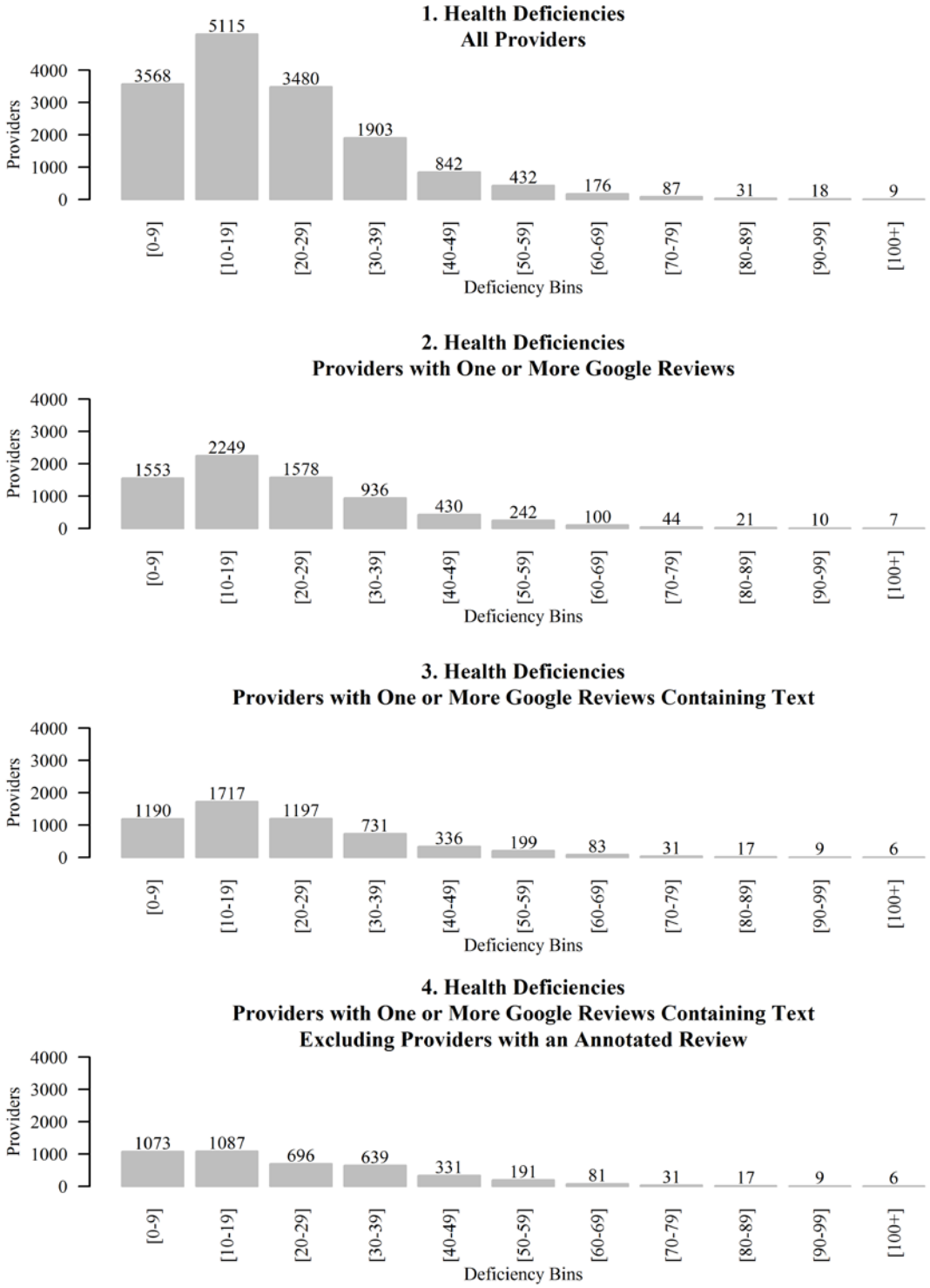

Figure 1: Number of providers binned by the number of health deficiencies each received. The sequence of plots shows the progression of changes in the provider distribution as providers were removed to form the final (bottom) plot of providers in the test set. The exclusion of providers with an annotated review excluded all providers which had the text of any of their reviews match the text of any review used in training.

Maximum Entropy Classifier

The maximum entropy classifier was trained on 2,500 reviews that were hand-annotated for indications of elder abuse. To test inter-rater agreement, a subset of 100 reviews was labeled by two additional annotators, resulting in a Krippendorff's Alpha Coefficient [25] of 0.79. For this study, elder abuse was defined as including both intentional and unintentional abuse, as well as neglect. Common examples included staff (1) failing to respond to call lights in a timely manner, (2) allowing patients to develop chronic bed sores, (3) leaving patients in soiled clothing or sheets for extended periods of time, and (4) demonstrating incompetence in recognizing or reporting residents' medical problems. Poor food quality was also a common complaint, but was only marked as abuse if either the review indicated the quality was so poor that it adversely impacted a resident's health, or descriptions of the food clearly indicated negligence, such as serving food that was still frozen. Rude, condescending, or dismissive behavior was only marked as abuse if it was directed toward a patient and seemed to be on-going, such that it could be regarded as psychological abuse. To focus the classifier on identifying reviews which could be the most helpful to inspectors, a review was only considered an indication of abuse if it provided reasonably specific information identifying an abuse. As a consequence, reviews describing a facility as "horrible", or advising readers that loved ones sent to the facility would die, were not marked as indicative of abuse unless the review also included a more specific complaint of abuse.

The maximum entropy classifier uses Apache's OpenNLP [26] implementation. For each review, Uniform Resource Locators (URLs) appearing in the review text were replaced with a URL token, and then unigram and bigram features were extracted. The classifier also uses the Google rating divided by 5.0 as a feature, as well as the review length in [0.0, 1.0], with 1.0 corresponding to a length of 2,000 characters (longer lengths are assigned a value of 1.0). The classifier uses Gaussian regularization with $\sigma = 1.0$ and 10,000 iterations to ensure convergence. The classifier's performance was tested using stratified 10-fold cross-validation.

Ratings

While the classifier was designed to detect specific references to elder abuse, the Google ratings were used as an indication of the technical quality of care. The distribution of Google ratings as a function of providers binned by their number of health deficiencies reveals that reviews typically exhibit extreme polarity between 1 and 5 star reviews, with the ratio of 1 to 5 star reviews correlating with the number of deficiencies found by CMS inspectors (Figure 2). This distribution includes all providers which had at least one review, regardless of whether the review contained text or was used in the training data set.

The distribution of provider counts as a function of the number of Google reviews available for the provider remains relatively unchanged between the original data and the test set (Figure 3). Although providers with fewer reviews were more likely to be removed due to having only non-text reviews, the distribution remained approximately the same due to providers whose review counts decreased after removing non-text reviews. The test set exhibits a distribution similar to the original data. However, the predominance of providers with one or two reviews presents a challenge for assessing the quality of care in nursing homes. The average rating in the test set is 3.04. There are 8,787 ratings with 4,161 nursing homes, which yields an average number of reviews per provider of 2.11.

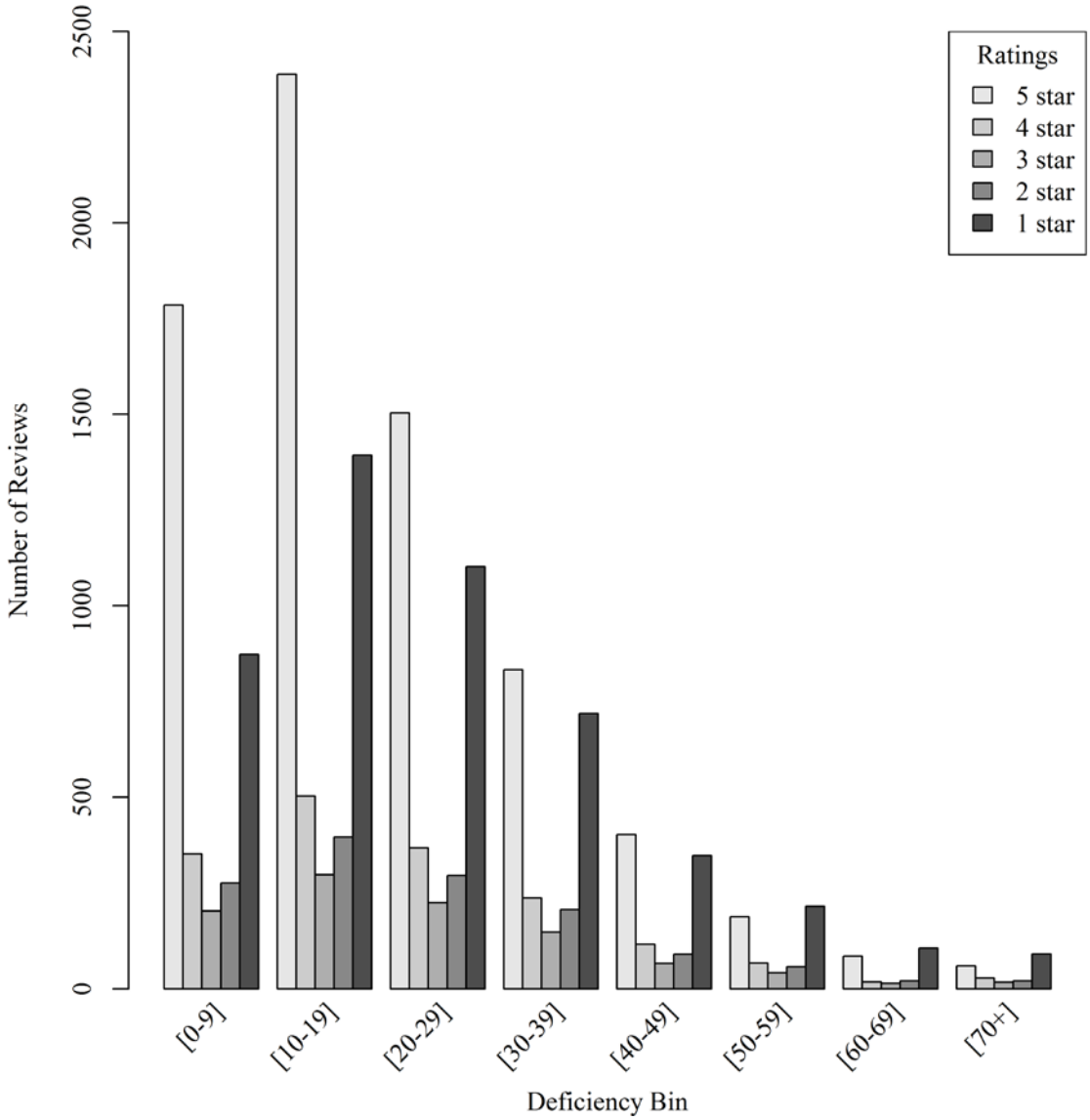

Figure 2: Google rating distributions for providers in each health deficiency bin. Provider bins corresponding to high health deficiency counts have a higher ratio of 1 star (bad) to 5 star (good) reviews than providers in bins corresponding to low health deficiency counts.

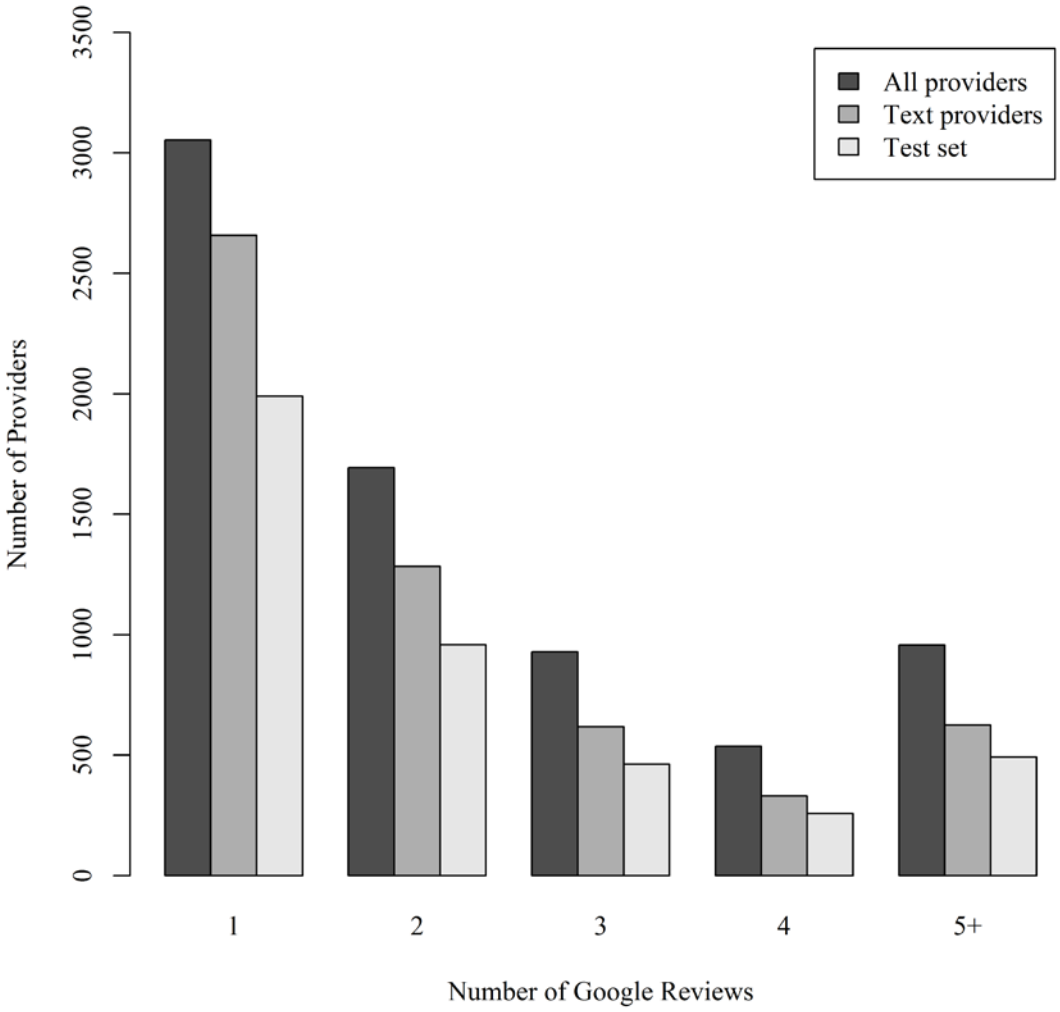

Figure 3: Number of providers as a function of the number of Google reviews each provider received. The text providers had at least one review which contained text. Test set providers had at least one text review and also had no reviews with text matching any review text used to train the classifier.

Correlations, Aggregations, and Analyses

This section discusses several methods to address the challenge of having few reviews per nursing home, including correlation calculations to maximize statistical power, and aggregating providers to simulate a larger number of reviews per provider without generating synthetic reviews. Additional analyses are described to examine the extent to which Google's selection of the top five reviews influences the results, the correlations as a function of the number of deficiencies received by providers, and the relationships between deficiency categories.

Maximizing the statistical power of correlation calculations is valuable since limited numbers of reviews are available per nursing home and past studies have found varying correlation strengths between reviews and other quality measure data. This study uses a Rank Inverse Normal transform-based Pearson correlation coefficient to maximize statistical power for continuous variables, which Bishara et al. [27] found yielded few Type I and Type II errors while maximizing statistical power for testing the significance of bivariate correlations involving continuous variables with reasonable sample sizes ($n \geq 20$). This study also uses the Henze-Zirkler test for multivariate normality, which was recommended by a Monte Carlo study of 13 tests for multivariate normality [28]. The MVN package [29] for R [30] is used for most of the statistical calculations. Both the per-provider average classifier scores versus health deficiency counts and the average ratings versus health deficiency counts fail the Henze-Zirkler test, indicating normalization or rank-based methods are required. Furthermore, Shapiro-Wilk's normality test shows that none of the deficiency count, average ratings, and average classifier score univariate distributions are normal. The Rankit equation [31], a type of Rank-Based Inverse Normal Transformation, was used to normalize each of the three univariate distributions since it was found to be an accurate normalization method [32]. After normalization, the classifier versus deficiency data passes the Henze-Zirkler test. However, the rating versus deficiency data fails the Henze-Zirkler test even after applying the RIN transformation, so the Spearman correlation coefficient may be preferable for correlations of ratings and deficiencies, especially since the ratings data is drawn from a discrete distribution.

Another method of overcoming the limited number of reviews per provider is to aggregate similar providers. Aggregating similar providers simulates having a larger average number of reviews per provider without generating any synthetic reviews. First, providers were sorted in order by the number of CMS deficiencies they received, so that similar providers were adjacent. Next, given a stride value s, each non-overlapping, consecutive group of s providers was merged to produce an aggregated provider whose deficiencies were the union of the deficiencies for each of the individual providers, and whose ratings and classifier scores were the average of the average ratings and classifier scores for the individual providers. Effectively, given the average number of reviews per provider $n = 2.11$, applying a stride s increases the average number of reviews per aggregated provider to $s \cdot n$. However, this extrapolation has limitations. For example, adjacent providers may differ in the distribution of their deficiency codes, which would reduce the correlation between the aggregated provider's reviews and the aggregated provider's deficiency distribution, yielding lower correlation test results. Fortunately, as will be discussed in the Results section, categories of deficiencies have moderately strong correlations with one another.

Two additional analyses can help in interpreting the results. First, comparing correlation coefficients between the full test set and the set of providers which received four or fewer reviews can help measure any influence on the results due to Google's selection of the top five reviews. Second, examining providers binned by their deficiency counts can reveal whether the correlations tested in this study vary as a function of the number of deficiencies received by providers.

Finally, examining correlations between the maximum entropy classifier, the ratings, and categories of health deficiencies—including categories of abuse-related deficiencies—can provide insights into their strengths and weaknesses. This study includes a correlogram of

Spearman correlation coefficients between all health deficiencies, the Google ratings, the classifier scores, a subset of deficiencies chosen to reflect severe quality of care issues, minor deficiencies (the health deficiencies remaining after excluding the severe deficiencies), three deficiencies which explicitly mention "abuse" in their descriptions, and the set of health deficiencies which do not include "abuse" in their descriptions. The severe deficiencies category is designed to capture the deficiencies which describe poor technical quality of care or indications of possible abuse, regardless of whether the deficiency description includes the word "abuse". Examples of severe deficiencies include using unnecessary physical restraints, using unnecessary drugs to restrain patients, not complying with legal requirements for providing care (such as having the necessary licenses), not allowing residents to accept visitors, not giving residents access to private phones, and not preventing dehydration. In addition, the three deficiencies which contain "abuse" in their description are also included in the severe deficiencies set.

Results

Overview

To measure the usefulness of online reviews for assessing the quality of care, both the online ratings and the classifier scores were compared to the number of CMS deficiencies per provider. To measure the usefulness of online reviews for detecting elder abuse, the maximum entropy classifier was tested using 10-fold cross-validation. The Maximum Entropy Classifier and Ratings sections present correlations with deficiency count data for the maximum entropy classifier and review ratings data, respectively. The Maximum Entropy Classifier section also reports the 10-fold cross-validation results for the classifier. The Correlations, Aggregations, and Analyses section presents results demonstrating each of the following: that the correlation coefficients increase significantly as the number of reviews per provider is increased by aggregating providers; that Google's selection of the top five reviews has little impact on the results of this study; that the average classifier score, average rating, and average deficiency counts for providers binned by their health deficiency counts are consistent across bins; and finally, that a correlogram analysis reveals strong correlations between deficiency categories, which results in the ratings correlating best with abuse related deficiency categories while the maximum entropy classifier's high precision makes it best-suited to supporting investigators searching for indications of elder abuse.

Maximum Entropy Classifier

The classifier achieved an F-measure of 0.81, with precision 0.74 and recall 0.89, based on 10-fold cross validation with 2,500 hand-annotated reviews and a Krippendorff's Alpha Coefficient of 0.79. The RIN Pearson correlation coefficient for the deficiency versus classifier data set indicates a weak but statistically significant correlation: r_{RIN} = 0.13, P < .001, and 95% confidence interval (0.10, 0.16). The Spearman correlation coefficient for the classifier is similar: r_S = 0.13, P < .001. Finally, the regular Pearson correlation coefficient is r_p = 0.14, P < .001, and 95% confidence interval (0.11, 0.17). These results show that the elder abuse classifier successfully locates indications of elder abuse in review text with high precision, and that the correlation with the overall quality of care at a facility, as represented by the CMS deficiency counts, is low. Although the weak correlation with CMS deficiency counts is expected since the

classifier is not intended to assess the overall quality of care, the Correlations, Aggregations, and Analyses section will demonstrate that there are substantial correlations between abuse-related deficiencies and other deficiencies.

Ratings

The ratings data achieves a slightly higher Spearman correlation with the deficiency data: r_S = 0.15, P < .001. For comparison purposes, the RIN Pearson correlation coefficient for the ratings is r_{RIN} = 0.15, P < .001, and 95% confidence interval (0.12, 0.18) and the regular Pearson correlation coefficient is r_p = 0.15, P < .001, and 95% confidence interval (0.12, 0.18). However, the Pearson correlation coefficients for the ratings should be treated with caution, since the ratings and deficiency data fail the Henze-Zirkler test for bivariate normality, which indicates the result may be imprecise. The correlation coefficients for the ratings are higher than the correlation coefficients for the classifier, which is expected since the deficiency counts represent an overall measure of a nursing home's quality of care, whereas the classifier is designed to identify abuse.

Correlations, Aggregations, and Analyses

The correlations for both the classifier scores and the ratings with deficiency counts are statistically significant but weak, due to the small number of reviews available per provider. However, the correlation strength steadily increases as similar providers are aggregated to effectively increase the number of reviews per provider, approaching a value of approximately 0.65 (Figure 4). At strides of 5 and above, both the deficiency versus classifier and deficiency versus rating data sets pass the Henze-Zirkler test. Therefore, the correlations shown are all RIN transform-based Pearson correlation coefficients. For each correlation with classifier scores at each stride, P < .001, and for each correlation with Google ratings at each stride, P < .001. Therefore, the set of statistical correlation tests remains valid under a Bonferroni correction.

Since Google selects the top five reviews to return when a provider has more than five reviews, the correlation results could be influenced by Google's selections. Restricting the analysis to use only providers with 4 or fewer reviews results in a deficiency versus classifier RIN Pearson correlation coefficient of r_{RIN} = 0.13 with (P < .001) and 95% confidence interval (0.10, 0.17), while the deficiency versus rating Spearman correlation coefficient becomes r_S = 0.16 with (P < .001). The similarity of these correlation results to results which include providers with five or more reviews suggests that the Google selection does not have a significant effect. This may be due to the relatively small number of providers with five or more reviews in this study (Figure 3).

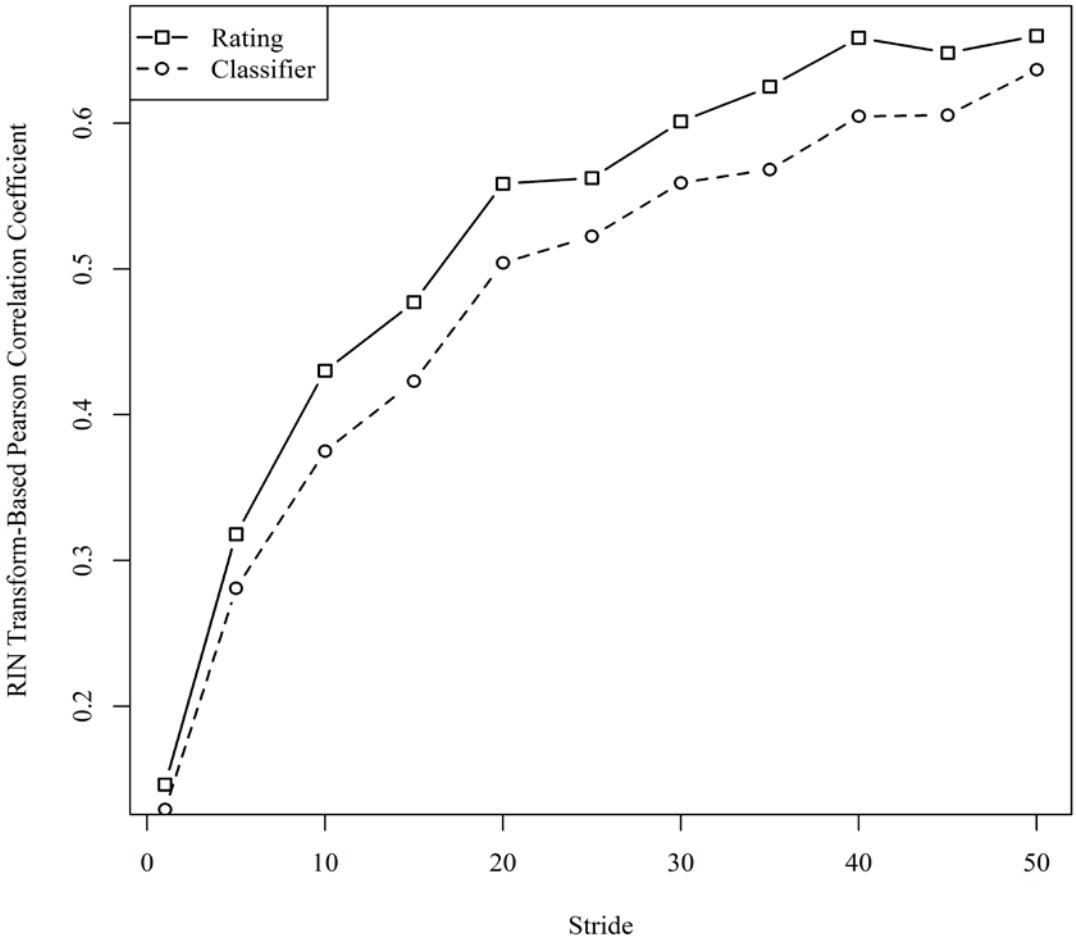

Figure 4: RIN transform-based Pearson correlation coefficients between Google ratings or classifier scores and CMS health deficiency counts as a function of stride. The stride is the number of providers with similar deficiency counts merged to produce an aggregated provider with an effectively larger number of reviews. Aggregating providers enables extrapolating the potential correlation coefficients achievable with more reviews per provider without generating synthetic reviews. Google ratings correlate better with health deficiency counts, while classifier scores correlate nearly as well, even though the classifier scores reflect references to elder abuse rather than overall quality of care.

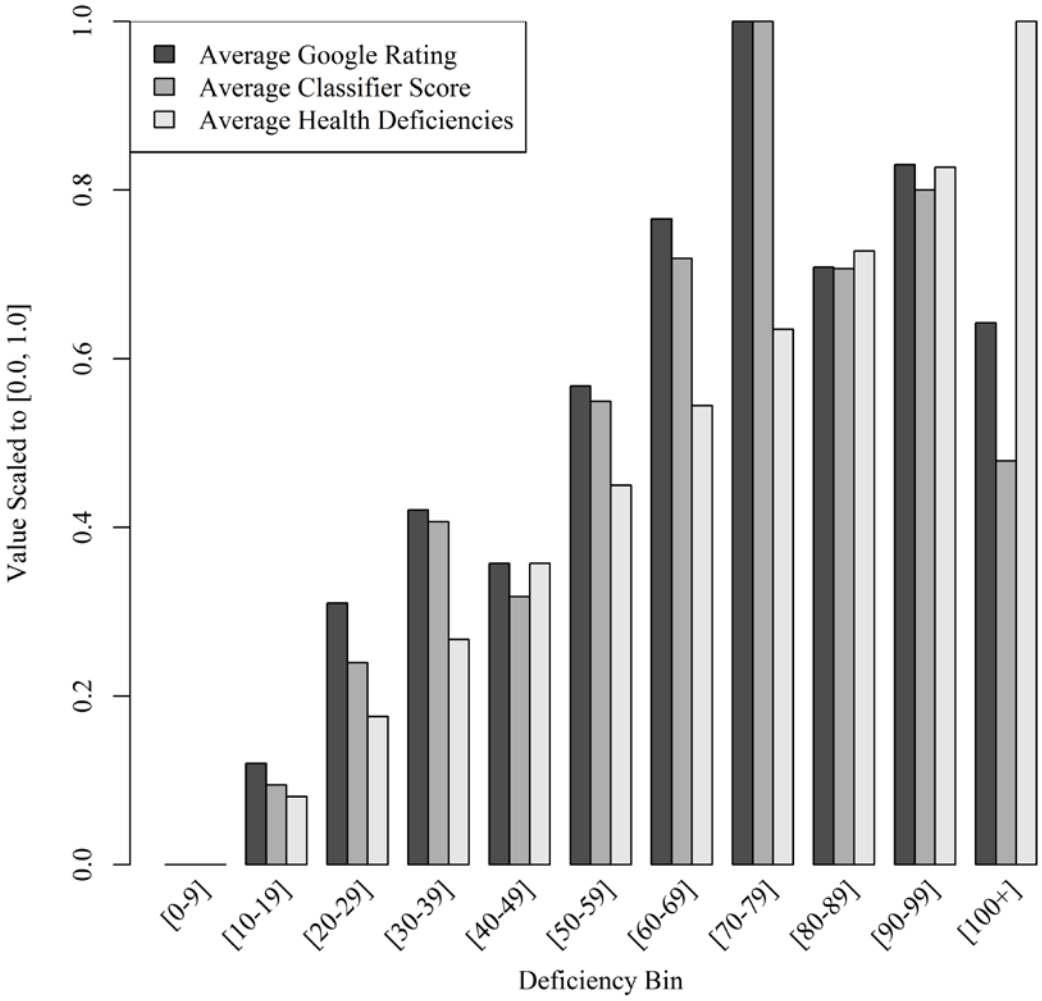

Figure 5: Illustration of correlations between average Google ratings, average classifier scores, and average health deficiency counts for providers in each health deficiency bin. The relationships between the averages remains stable across deficiency bins except for the [100+] bin, which contains only six providers.

Examining average classifier scores and average ratings across provider deficiency bins indicates that the correlations are consistent even as the provider quality of care varies (Figure 5). Each series has been normalized to have a minimum value of 0.0 and a maximum of 1.0. The ratings have also been reversed, such that a rating of 1 equates to a value of 1.0, while a rating of 5 equates to 0.0. This makes visual comparison of the series data easier, since it means higher values indicate poorer quality of care or a greater degree of abuse for each series. Although the deficiency bin range is shown on the x-axis, displaying the average number of deficiencies facilitates visual comparison of the ground truth data with the classifier and rating data, and it also shows that the average deficiency count for the [100+] bin is high. There are few providers in the bins on the right (Figure 1), which means those values will not be statistically significant.

The ratings and the classifier scores both consistently reflect the average number of deficiencies across deficiency bins.

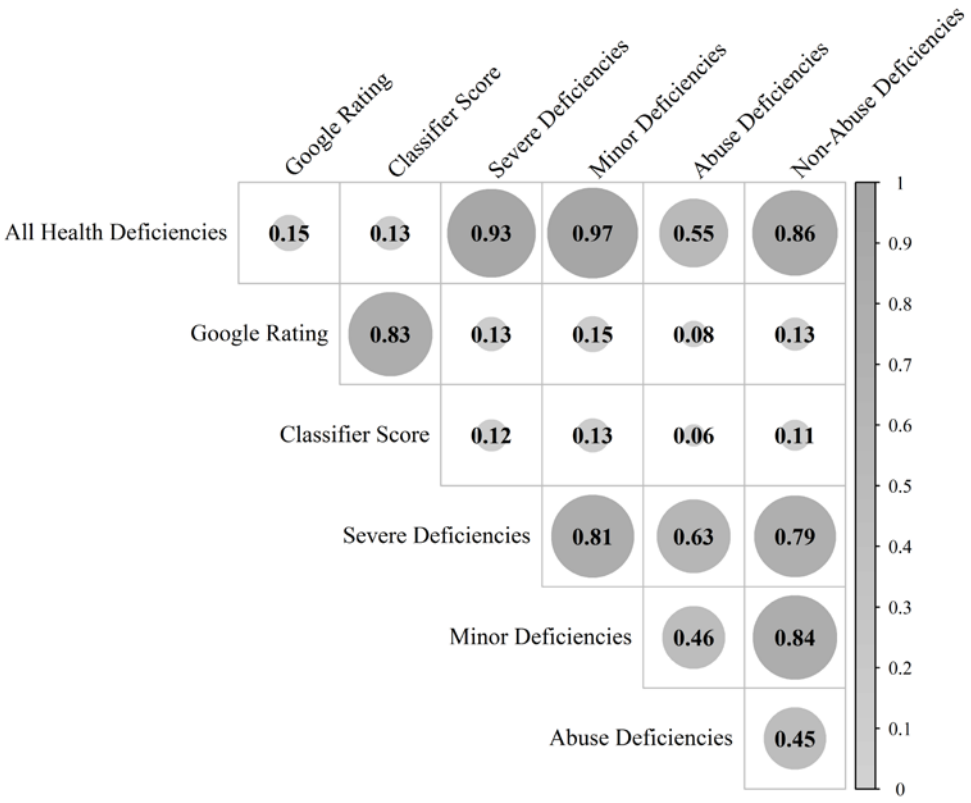

Figure 6: Spearman correlations between Google ratings, classifier scores, and categories of deficiencies. Abuse deficiencies include only deficiencies whose CMS descriptions included the word abuse, while the set of severe deficiencies include both abuse deficiencies and additional deficiencies chosen as indicators of very poor technical quality of care or possible elder abuse. The severe and minor deficiencies are disjoint sets, as are the abuse and non-abuse deficiencies. Deficiency categories correlate strongly with one another, which contributes to classifier scores correlating well with ratings.

The similarity between the rating and classifier results corroborates anecdotal experience from annotating the reviews: many negative reviews are either indicative of elder abuse or lack sufficient detail to determine whether abuse is taking place. There are moderate to strong Spearman correlation coefficients between the maximum entropy classifier, ratings, and several deficiency categories (Figure 6). For consistency, since high values for classifier scores and deficiency counts are both indicative of poor quality of care or elder abuse, the ratings were scaled to [0.0, 1.0], with 1.0 corresponding to a rating of 1 (worst) and 0.0 corresponding to a rating of 5 (best). The sum of *P*-values for all tests is < .001.

Notably, the Google ratings correlate better than the classifier scores with both the severe deficiencies and the explicit abuse deficiencies, even though the classifier is trained to identify indications of abuse and achieved a precision of 0.74. The higher correlation coefficient for the

ratings may be explained by the significant correlations observed between each of the four deficiency categories: facilities which provide poor technical quality of care are more likely to receive deficiencies indicative of abuse, and vice versa. As a result, the Google rating result benefits since it reflects reviewers' overall impression of a facility, which leverages the dependencies between deficiencies. Moreover, the annotation guidelines used to train the classifier are designed to identify references to abuse which contain sufficient detail to identify the type of abuse. This is useful for supporting inspectors (e.g. in a similar manner to [21]), but it also means the classifier will not benefit as much as the ratings from the dependencies between poor overall quality of care and deficiencies.

Discussion

Google review data exhibits a weak but statistically significant correlation with Centers for Medicare and Medicaid Services' inspection results for nursing homes listed in the Nursing Home Compare data set. The Spearman correlation coefficient for the ratings, r_S = 0.15 (P < .001), is slightly higher than the RIN transform-based Pearson correlation coefficient for the classifier scores r_{RIN} = 0.13, P < .001, and 95% confidence interval (0.10, 0.16). The classifier achieved an F-measure of 0.81, with precision 0.74 and recall 0.89, based on 10-fold cross-validation with 2,500 hand-annotated reviews. Comparisons of correlations between ratings, classifier scores, and deficiency categories revealed that ratings exhibit higher correlations with deficiency counts, even for categories of deficiencies limited to severe or abuse-related deficiencies. This indicates ratings are a better overall measure of the technical quality of care at a facility, while the classifier's high precision is better suited to supporting inspectors looking for reviews which contain sufficient detail to identify the type of elder abuse.

Aggregating providers with similar deficiency counts causes the RIN transform-based Pearson correlation coefficients, for both ratings and classifier scores with deficiency counts, to approach approximately 0.65 as the effective number of reviews per provider increases. This suggests that as the popularity of PRWs increases, the validity of online review data in assessing nursing homes will increase correspondingly. These correlations contribute to the body of literature which has already demonstrated correlations between online review data and quality of care for other types of healthcare providers, such as hospitals and physicians.

This study aggregated providers with similar total deficiency counts to perform the extrapolation, although slightly better results may be achievable by clustering providers based on the distribution of their deficiencies. In addition, since chains of nursing homes are sometimes investigated for poor quality of care or fraud, aggregating nursing homes by ownership may be a useful investigative tool.

The positive correlations found in this study also suggest that as the number of online reviews grows, the reviews and the NHC data could jointly enable consumers to make more informed decisions. The methods are complementary: CMS inspections provide detailed information on specific aspects of care but have limited timeliness, while online reviews provide subjective information in a timely manner on aspects of care noted by the reviewers. A study on consumer use of the NHC data set found that consumers have limited awareness of it, and the study authors suggest that including measures of "consumer satisfaction" could increase its usefulness [33]. A study of nursing home information and search capabilities on state websites found that less than

a quarter were perceived as easy to find [34]. Popular online review websites could incorporate the NHC data into their services, which would address both the limited awareness of the NHC data and the lack of consumer feedback in the NHC data.

Further research is needed to understand the factors that influence which studies find correlations between online review data and other quality care measures. The factors will naturally include the volume of data, PRW, and the provider type, but comparisons with other studies suggest additional possibilities. Two previous studies using non-social media surveys of elderly patients did not find a correlation between the survey data and the technical quality of care [19,20]. Since those studies surveyed elderly patients while this study used online reviews in which reviewers were likely to be younger, and to be relatives of patients rather than patients themselves, age and reviewer perspective may both have been factors. In addition, the type of healthcare service provided may be a factor, since the high correlations between categories of deficiencies in the NHC data suggest that a variety of aspects of the overall quality of care, many of which are readily discernable by visiting relatives, will be indicative of the technical quality of care and the likelihood of elder abuse. This may not generalize to other healthcare provider types, in which there may be a wider range of services offered to patients and in which the technical competence of staff may be harder to discern without a medical background.

Future research could also address potential sources of bias in online ratings. In one consumer survey, when participants were asked about the implications of writing a negative review, 34% expressed concern over having their identity disclosed, while 26% expressed concern that the physician might take action against them [2]. Although a survey of studies on PRWs found that approximately 90% of ratings were positive and that there was no evidence of “doctor-bashing” [4], the reviews of nursing homes in this study exhibited strong polarization. There could be many reasons for this difference, including a reviewer being more willing to choose a low rating for a facility than for a named physician, self-selection biases varying between PRWs, and whether a PRW allows anonymous reviews. The increasing popularity of PRWs may also increase providers’ incentives to generate positive fake reviews of themselves or negative fake reviews of their competitors. Examining factors taken into account by humans when using online review data, such as the number of reviews for a provider and the degree of emotion or factual information expressed in the reviews [35], may yield useful methods for automated assessments of the review data’s validity, in addition to direct methods for detecting fake reviews. Geography is another possible factor, since a study of general practitioners found that the practitioner’s location influenced correlations with referral volume and peer-nominated awards [36]. There may also be a self-selection bias, since users of PRWs can be characterized by psychographic variables, information-seeking behavior, and health status [37]. Finally, since consumers will vary in their prioritization of different aspects of nursing home care [22], efforts to isolate aspects of care through machine learning techniques such as clustering (e.g [9].) could provide consumers with information tailored to each consumer’s priorities.

Conclusion

Although the online review data has many potential sources of bias and there is ample room for further research to improve the accuracy of information extracted from online reviews, this study still found that both the maximum entropy classifier and the Google ratings approach a RIN Pearson correlation coefficient of approximately 0.65 as the effective number of reviews

increases. This correlation indicates that as more online reviews become available, they will become a valuable resource for assessing the technical quality of care and the prevalence of elder abuse in nursing homes. This could help patients choose nursing homes, help regulators protect patients from abuse, help inspectors work more efficiently, and help policy-makers make decisions by providing additional quantifiable data. Moreover, if legal and financial restrictions on collecting reviews from multiple PRWs could be overcome, it is likely that some nursing homes would have a sufficient number of reviews to use the correlations found in this study, benefiting patients, regulators, inspectors, and policy-makers in the near-term.

Acknowledgements

The authors would like to thank The MITRE Corporation for funding this research.

Conflicts of Interest

None declared. As a not-for-profit operator of federally funded research and development centers, The MITRE Corporation is not permitted to compete with industry.

References

1. Rastegar-Mojarad M, Ye Z, Wall D, Murali N, Lin S. 2015. Collecting and analyzing patient experiences of health care from social media. *JMIR Res Protoc*. 4(3), e78. PubMed http://dx.doi.org/10.2196/resprot.3433

2. Hanauer DA, Zheng K, Singer DC, Gebremariam A, Davis MM. 2014. Public awareness, perception, and use of online physician rating sites. *JAMA*. 311(7), 734-35. PubMed http://dx.doi.org/10.1001/jama.2013.283194

3. Burkle CM, Keegan MT. 2015. Popularity of internet physician rating sites and their apparent influence on patients' choices of physicians. *BMC Health Serv Res*. 15, 416. PubMed http://dx.doi.org/10.1186/s12913-015-1099-2

4. Emmert M, Sander U, Pisch F. 2013. Eight questions about physician-rating websites: a systematic review. *J Med Internet Res*. 15(2), e24. PubMed http://dx.doi.org/10.2196/jmir.2360

5. Verhoef LM, van de Belt TH, Engelen LJ, Schoonhoven L, Kool RB. 2014. Social media and rating sites as tools to understanding quality of care: a scoping review. *J Med Internet Res*. 16(2), e56. PubMed http://dx.doi.org/10.2196/jmir.3024

6. Greaves F, Pape UJ, King D, Darzi A, Majeed A, et al. 2012. Associations between internet-based patient ratings and conventional surveys of patient experience in the English NHS: an observational study. *BMJ Qual Saf*. 21(7), 600-05. PubMed http://dx.doi.org/10.1136/bmjqs-2012-000906

7. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. 2013. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *J Med Internet Res*. 15(11), e239. PubMed http://dx.doi.org/10.2196/jmir.2721

8. Kilaru AS, Meisel ZF, Paciotti B, Ha YP, Smith RJ, et al. 2016. What do patients say about emergency departments in online reviews? A qualitative study. *BMJ Qual Saf*. 25(1), 14-24. PubMed http://dx.doi.org/10.1136/bmjqs-2015-004035

9. Ranard BL, Werner RM, Antanavicius T, Schwartz HA, Smith RJ, et al. 2016. Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. *Health Aff (Millwood)*. 35(4), 697-705. PubMed http://dx.doi.org/10.1377/hlthaff.2015.1030

10. Bardach NS, Asteria-Peñaloza R, Boscardin WJ, Dudley RA. 2013. The relationship between commercial website ratings and traditional hospital performance measures in the USA. *BMJ Qual Saf*. 22(3), 194-202. PubMed http://dx.doi.org/10.1136/bmjqs-2012-001360

11. Hawkins JB, Brownstein JS, Tuli G, Runels T, Broecker K, et al. 2016. Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Qual Saf*. 25(6), 404-13. PubMed http://dx.doi.org/10.1136/bmjqs-2015-004309

12. Jung Y, Hur C, Jung D, Kim M. 2015. Identifying key hospital service quality factors in online health communities. *J Med Internet Res*. 17(4), e90. PubMed http://dx.doi.org/10.2196/jmir.3646

13. Glover M, Khalilzadeh O, Choy G, Prabhakar A, Pandharipande P, et al. 2015. Hospital evaluations by social media: a comparative analysis of Facebook ratings among performance outliers. *J Gen Intern Med*. 30(10), 1440-46. PubMed http://dx.doi.org/10.1007/s11606-015-3236-3

14. Greaves F, Pape UJ, Lee H, Smith DM, Darzi A, et al. 2012. Patients' ratings of family physician practices on the Internet: usage and associations with conventional measures of quality in the English National Health Service. *J Med Internet Res*. 14(5), e146. PubMed http://dx.doi.org/10.2196/jmir.2280

15. Wallace BC, Paul MJ, Sarkar U, Trikalinos TA, Dredze M. 2014. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *J Am Med Inform Assoc*. 21(6), 1098-103. PubMed http://dx.doi.org/10.1136/amiajnl-2014-002711

16. Gao GG, McCullough JS, Agarwal R, Jha AK. 2012. A changing landscape of physician quality reporting: analysis of patients' online ratings of their physicians over a 5-year period. *J Med Internet Res*. 14(1), e38. PubMed http://dx.doi.org/10.2196/jmir.2003

17. Segal J, Sacopulos M, Sheets V, Thurston I, Brooks K, et al. 2012. Online doctor reviews: do they track surgeon volume, a proxy for quality of care? *J Med Internet Res*. 14(2), e50. PubMed http://dx.doi.org/10.2196/jmir.2005

18. Emmert M, Adelhardt T, Sander U, Wambach V, Lindenthal J. 2015. A cross-sectional study assessing the association between online ratings and structural and quality of care measures: results from two German physician rating websites. *BMC Health Serv Res*. 15(1), 414. PubMed http://dx.doi.org/10.1186/s12913-015-1051-5

19. Chang JT, Hays RD, Shekelle PG, MacLean CH, Solomon DH, et al. 2006. Patients' global ratings of their health care are not associated with the technical quality of their care. *Ann Intern Med*. 145(8), 635-36. PubMed http://dx.doi.org/10.7326/0003-4819-145-8-200610170-00020

20. Rao M, Clarke A, Sanderson C, Hammersley R. 2006. Patients' own assessments of quality of primary care compared with objective records based measures of technical quality of care: cross sectional study. *BMJ*. 333(7557), 19. PubMed http://dx.doi.org/10.1136/bmj.38874.499167.7C

21. Centers for Medicare and Medicaid Services. 2016. Nursing Home Compare Data Archive. https://data.medicare.gov/data/archives/nursing-home-compare. Archived at: http://www.webcitation.org/6iC1K4y4G

22. van de Belt TH, Engelen LJ, Verhoef LM, van der Weide MJ, Schoonhoven L, et al. 2015. Using patient experiences on Dutch social media to supervise health care services: exploratory study. *J Med Internet Res*. 17(1), e7. PubMed http://dx.doi.org/10.2196/jmir.3906

23. Mukamel DB, Amin A, Weimer DL, Sharit J, Ladd H, et al. 2016. When patients customize nursing home ratings, choices and rankings differ from the government's version. *Health Aff (Millwood)*. 35(4), 714-19. PubMed http://dx.doi.org/10.1377/hlthaff.2015.1340

24. Google. 2016. Google Maps APIs. https://developers.google.com/maps/. Archived at: http://www.webcitation.org/6iCGXJ6TW

25. Hayes AF, Krippendorff K. 2007. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas*. 1(1), 77-89. doi:http://dx.doi.org/10.1080/19312450709336664.

26. Apache Software Foundation. 2016. The Apache OpenNLP Library. https://opennlp.apache.org/. Archived at: http://www.webcitation.org/6hvXhTr5U

27. Bishara AJ, Hittner JB. 2012. Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychol Methods*. 17(3), 399-417. PubMed http://dx.doi.org/10.1037/a0028087

28. Mecklin CJ, Mundfrom DJ. 2005. A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *J Stat Comput Simul*. 75(2), 93-107. doi:http://dx.doi.org/10.1080/00949650420001932333.

29. Korkmaz S, Goksuluk D, Zararsiz G. 2014. MVN: An R Package for Assessing Multivariate Normality. https://cran.r-project.org/web/packages/MVN/vignettes/MVN.pdf. Archived at: http://www.webcitation.org/6iCHMjdZZ

30. R Core Team. 2016. A Language and Environment for Statistical Computing R Foundation for Statistical Computing. https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf. Archived at: http://www.webcitation.org/6iCHaoYyS

31. Bliss CI. Statistics in Biology. New York, NY: McGraw-Hill Higher Education; 1967. ISBN:0070058954

32. Solomon SR, Sawilowsky SS. 2009. Impact of rank-based normalizing transformations on the accuracy of test scores. *J Mod Appl Stat Methods*. 8(2), 448-62. http://digitalcommons.wayne.edu/coe_tbf/5.

33. Konetzka RT, Perraillon MC. 2016. Use of nursing home compare website appears limited by lack of awareness and initial mistrust of the data. *Health Aff (Millwood)*. 35(4), 706-13. PubMed http://dx.doi.org/10.1377/hlthaff.2015.1377

34. Liu D, Lu C-J. 2015. An evaluation of web-based nursing home finders. *J Consum Health Internet*. 19(2), 77-92. doi:http://dx.doi.org/10.1080/15398285.2015.1026701.

35. Grabner-Kräuter S, Waiguny MK. 2015. Insights into the impact of online physician reviews on patients' decision making: randomized experiment. *J Med Internet Res*. 17(4), e93. PubMed http://dx.doi.org/10.2196/jmir.3991

36. Wiley MT, Rivas RL, Hristidis V. 2016. Provider attributes correlation analysis to their referral frequency and awards. *BMC Health Serv Res*. 16(1), 90. PubMed http://dx.doi.org/10.1186/s12913-016-1338-1

37. Terlutter R, Bidmon S, Röttl J. 2014. Who uses physician-rating websites? Differences in sociodemographic variables, psychographic variables, and health status of users and nonusers of physician-rating websites. *J Med Internet Res*. 16(3), e97. PubMed http://dx.doi.org/10.2196/jmir.3145