Journal of Information Systems Applied Research

Volume 16, Issue 1
March 2023
ISSN: 1946-1836

In this issue:

Information Systems & Computing Academic Professionals

The **Journal of Information Systems Applied Research** (JISAR) is a double-blind peer reviewed academic journal published by ISCAP, Information Systems and Computing Academic Professionals. Publishing frequency is three to four issues a year. The first date of publication was December 1, 2008.

JISAR is published online (https://jisar.org) in connection with CONISAR, the Conference on Information Systems Applied Research, which is also double-blind peer reviewed. Our sister publication, the Proceedings of CONISAR, features all papers, panels, workshops, and presentations from the conference. (https://conisar.org)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%), other journal papers (top 30%), unsettled papers, and non-journal papers. The unsettled papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in the JISAR journal. Currently the acceptance rate for the journal is approximately 35%.

Questions should be addressed to the editor at editor@jisar.org or the publisher at publisher@jisar.org. Special thanks to members of ISCAP who perform the editorial and review processes for JISAR.

Journal of Information Systems Applied Research

Text Prediction Using Artificial Intelligence: An Analysis of Two Text Prediction Systems

Hayden Wimmer
hwimmer@georgiasouthern.edu
Georgia Southern University
Statesboro, GA 30460, USA

Jie Du
dujie@gvsu.edu
Grand Valley State
Allendale, MI 49401, USA

Peter Hui
pheri@georgiasouthern.edu
Georgia Southern University
Statesboro, GA 30460, USA

Abstract

Natural language Processing is a discipline under artificial intelligence that involves interaction between human language and computer systems. It involves analyzing and representation of natural language, the ability to comprehend both text and spoken words. Natural language processing has evolved to the extent of having the ability to give useful responses to human beings. Large language models have been making landmark advances with new more efficient algorithms and improved hardware and processing power. Models like Google's BERT power predictive text in search predictions. Recently, companies have been training models on billions of parameters, a task that was not feasible just a few short years ago. OpenAI is the market leader in this technology; however, competitors have emerged. This research project aims to investigate perceptions of OpenAI versus an emerging competitor, AI21 on the ability to answer questions and predict text. We developed two web applications that allowed users to key in any questions or text in the textbox, the web application will then answer the user query as a response. The web applications were embedded with Jurassic-1 language model API and the GPT-3 language model APIs. Subjects asked the AI systems questions and rated their perceptions of the results. Furthermore, we investigate perceived privacy of AI systems via a post survey.

Keywords: Artificial Intelligence, Natural Language Processing, Text Prediction.

1. INTRODUCTION

Processing and extracting meaning from human text and speech has challenged AI researchers. Both large corpuses of data and mass processing power are required to construct even the smallest and simplest models and sentient AI remains a distant horizon. Organizations have leveraged the power of the cloud and large computing resources and modern AI techniques to train large language models using billions of parameters. The most familiar application of large language models is on Google search's type ahead search suggestions and now many chat bots. One of the most famous language models is the Generative Pre-trained

Transformer 3 (GPT3) developed by OpenAI. It is the third-generation language prediction model in the GPT-n series. GPT-3 model has 175 billion parameters while its predecessor GPT-2, has 1.5 billion parameters. Recently, more organizations have entered this competitive marketplace of large language models offering their own pre-trained models.

In this work, we examine two of the leading pretrained language models, OpenAI and AI21. We seek to answer which system has better perceived performance as well as the impact of demographics and familiarity with AI on perceptions of privacy of AI systems. We develop two identical systems on the Python/Django framework which submits to its respective web API. We ask survey participants to ask each system a few questions and then answer a brief survey on that system. Which system is presented first is randomized to account for the learning effect. Following testing both Web applications, participants answer a brief post-survey on perceived privacy. We seek to answer a) which system has better performance as indicated by human subjects and 2) do demographics have any influence on participants perceived privacy.

The remainder of this paper is organized as follows, next we provide a brief review of related works, then we discuss our methodology in detail, following methodology we present our results to the aforesaid research questions, and then conclude our work.

2. RELATED WORKS

Language Models in NLP

Natural Language Processing (NLP) has been on the rise in recent years. Several language models perform NLP-related tasks. One of the language models is GPT which can be fine-tuned for a specific task such as sentiment analysis, text summarization. Panchbhai & Pankanti (2021) applied a GPT-2 model for a movie review task. A dataset from the IMDB website was used. 50 reviews of the dataset were used where 46% of the reviews were taken positive and 54% marked negative. Special character prompt and natural language prompt were used to test the GPT-2 model. The results indicated that GPT-2 performed better in the natural language prompt sentiment analysis compared to the character prompt.

Bidirectional Encoder Representations from Transformers (BERT) is another language model. BERT was developed by Google and has an encoder that generates contextual embedding from any kind of sequence. This capability is essential for sentiment analysis or question answering. González-Carvajal and Garrido-Merchán (2020) empirically tested a BERT model using different scenarios compared to the traditional TF-IDF (Term Frequency - Inverse Document Frequency) vocabulary fed machine learning algorithms like Ridge Classifier and Voting Classifier. The BERT emerged the best with an accuracy of 0.9387, voting classifier followed with an accuracy of 0.9007, and the Ridge classifier with an accuracy level of 0.8990 (González-Carvajal and Garrido-Merchán (2020).

Wang and Cho (2019) showed that BERT is a Markov random field language model and compared BERT to traditional left-to-right language models. A pre-trained BERT model was taken on a mix of Toronto Book Corpus and Wikipedia with their Pytorch implementation. The quality of the generations was evaluated by computing the BLEU (Bilingual Evaluation Understanding) between the generations and the original data distributions. The BERT generations were found to be of low quality but very diversified compared to GPT.

Yu, Su, and Luo (2019) claimed that BERT lacks task-specific knowledge and domain-related knowledge. They proposed a BERT-based text classification model BERT4TC via constructing auxiliary sentences to address the task-awareness problem. The language model was evaluated on different datasets and the results showed that BERT4TC with suitable auxiliary sentences significantly outperforms both typical feature-based methods and finetuning methods.

Xu, Liu, Shu, and Yu (2019) explored turning customer reviews into a large knowledge base that can be used to answer user questions, which is referred to as Review Reading Comprehension (RRC). A novel post-training approach was used to fine-tune the BERT language model and to increase performance. Restaurant reviews and amazon shopping reviews were used as datasets for the RRC. The experimental results showed that the proposed post-training is highly effective (Xu et al. (2019).

Word Prediction

Trnka and McCoy (2007) evaluated word prediction methods using the keystroke savings and found that a larger amount of out-of-domain language is more beneficial than a smaller amount of in-domain language for training any model and the topic model can be fine-tuned even when there is dissimilarity in text. It was suggested that adaptive language models have the potential to outperform both in-domain and out of domain language models.

Grujić and Milovanović (2019) investigated word prediction based on their associative relations using a neural network. The dataset was collected from publicly available books in the Serbian language, which includes dictionaries, wiki pages, and historical books. The results showed that the number of times a word appears in the training set does not affect the outcome of this particular test. It was also found that in the case of best-performing associations, most clues are not close to the target while the worst-performing associations have all clues together into the target area.

Rojan, Alias, Rajan, Mathew, and Sudarsan (2020) applied the BERT language model into Malayalam. The pre-trained language model task is to predict the original vocabulary id of the masked word based only on its context. The proposed model has the capability of solving general tasks like next sentence prediction and mask language modeling. The model showed outstanding results of accuracy of 83% for next sentence prediction compared to the original BERT with an accuracy of 54% for the same set of data.

Risk and Trust in AI

Crockett, Garratt, Latham, Colyer, and Goltz (2020) investigated whether there was a difference when comparing perceived risk and trust in AI between the general public and college students studying computer science. A questionnaire was developed to collect people's perceptions of trust and risk in AI applications. The participants were asked to first rate the risk of the system on a scale of 0-10 and then answer a set of opinion questions on trust bias, ethics, and whether they support the development of such systems. They found that in specific applications such as medical imaging and diagnosis, there was a significant difference of opinions between the two groups with regards to risk. Both groups scored a medium risk regarding AI taking greater control over cybersecurity and strongly agreed that education in how AI works was significant in building trust. Perception of risk was found to be greater when the outcome of an error is more personal or serious.

Hengstler, Enkel, Duelli, and Change (2016) investigated trust in applied AI. Semi-structured interviews were conducted with the workers working in the transportation and medical technology industries. Informal follow-ups via emails and short phone calls were also used. The results showed that operational security, data security, and purpose were the eminent factors influencing trust in the technology. A domestic development process for applied AI was proposed which include stakeholder alignment, transparency in development, and early, proactive communication.

Kim, Ferrin, and Rao (2008) developed a theoretical framework for the trust-based decision-making process. The framework was tested using the structural equation modeling technique on Internet consumer purchasing behavior data collected via a web survey. The results suggested that a consumer's trust has a strong positive impact on purchasing a product and a strong negative effect on consumer's perceived risk. The factors that influence a consumer's trust were also identified, including consumer disposition to trust, reputation, privacy concerns, security concerns, the information quality of the Website, and the company's reputation.

4. METHODOLOGY

During the summer of 2022, we conducted an IRB-approved study of analyzing two text prediction systems. This section gives an overview of the study procedure and describes the text prediction systems development, survey administration, and data analysis procedure.

Study Procedure

The participants were asked to follow a 6-step process to complete the study (see Figure 1). In Step 1, they took a pre-task survey to answer several demographics questions including gender, age, education level, occupation, and familiarity with AI. In Step 2, participants were asked to complete a task of asking a few questions (created by the user) via two web applications developed on Python and the Django framework and hosted on PythonAnywhere. The two systems utilize the OpenAI model and the AI21 model, respectively. The details about the system development are presented later. The interface for the tasks was identical between the OpenAI and AI21 language models and which language model they interacted was randomized by the Qualtrics survey software to mitigate any learning effect. In Step 3, the participants completed a post-task survey to rate the system. In Step 4, the participants performed the same task for the second system. To further emphasize, the systems were identical with the exception of the API with which they communicated, and the tasks were randomized. In Step 5, the participants answered the same questions in the post-task survey for the second system. During the last step, the participants were surveyed on their perceptions of privacy when using the prediction systems.

Figure 1 The Study Procedure

Development of Two Prediction Systems
To test information usefulness, reliability, sufficiency, satisfaction, and quality of artificial intelligence in text prediction, two web applications were developed which will allow the user to key in any text or a series of words in a text box, the web application will then predict the following words or sentences. The web application was developed using Python Django framework, JavaScript, HTML, and CSS programming languages. Django framework was used to develop the application because of its high-level Python web framework attribute, and its rapid development of secure web applications. JavaScript was used to create div elements in the application, the div elements were used to display the predictions after the user types in any series of words or texts. HTML was used to develop the structure of the web application while CSS was used to modify the design of the application. Among the two web applications, one uses AI21's Jurassic-1 language model API with the following parameters: maximum tokens of 15, a top K return of 0, and a temperature of 0.0. The second web application is embedded with OpenAI's GPT-3 language model using the same parameters as the Jurassic-1 language model. The user interface of the two systems is identical (see Figure 2).

To interact with the prediction systems, the user will be making a GET request to the APIs embedded in the web applications by entering any string of text or words. The API will then predict the following strings of texts or words and return them in form of JavaScript Object Notation. To be specific, the user will create a question and type it in the textbox below "Ask a Question" and click the "Search" button when he is done typing. The question typed in by the user is then displayed first in a blue box and the prediction from the system is displayed in a grey box just below the question box. Two sample questions and the predicted text from the system is shown in Figure 2.

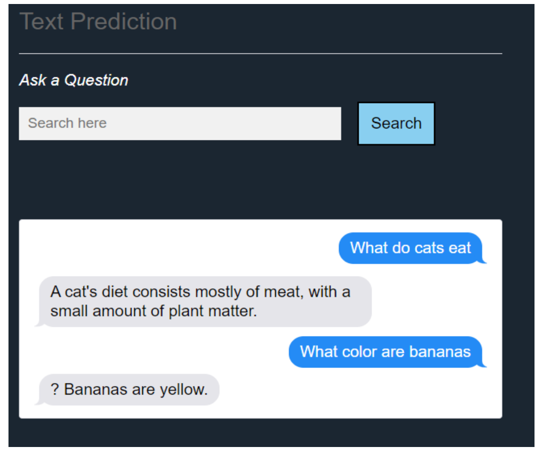


Figure 2 The UI of the Text Prediction System(s)

Survey Administration
An electronic Likert-scale questionnaire was implemented to survey the participants to gather their ratings on the text prediction systems and their perceived privacy when using the systems. The survey contains three sections: the pre-task section, the post-task section, and the perceived privacy section with a target completion time of less than 15 minutes. The pre-task section contains seven demographics questions including gender, age, education level, occupation, and familiarity with AI. The pos-task section consists of five questions about the overall ratings of the text prediction system. The five questions for the OpenAI are indexed as O1 to O5 and the five questions for the AI21 are indexed as J1 to J5. The perceived privacy section has nine questions about people's perception of privacy when they use the text prediction system. Survey participants were recruited from the Amazon MTurk service with a restriction to a location in the US. There were 39 total respondents and 8 had to be removed due to non-completion of the survey.

All items (see Appendix) except the demographics items are anchored on 7-point Likert scales, with the following weights on each response, Strongly Disagree = 7, Somewhat Disagree = 6, Disagree = 5, Neutral = 4, Agree = 3, Somewhat Agree = 2, and Strongly Agree = 1. A survey service from Qualtrics was used to administer the survey questions. The data collection yielded 31 usable survey response sets. sets. Table 1 summarizes the demographics of the sample.

Demographic Category	Percentage
Gender	Male: 24 Female: 14 No binary/third gender: 0 Prefer not to say: 1
Marital Status	Married: 13 Windowed:1 Divorced:7 Never married:18
Age Representation	Under 18: 18-24: 1 25-34: 8 35-44: 17 45-54: 6 55-64: 6 65-74:0 75-84:0 85 or older:1
Education Level	High school graduate: 6 Some College: 12 2-year degree: 6 4-year degree 13 Professional Degree: 1 Doctorate: 1
Occupation	Employed full time: 30 Employed part time: 3 Unemployed looking for work: 2 Unemployed not looking for work: 2 Retired: 1 Student: 0 Disabled: 1
Familiarity with AI	Strongly agree: 8 Agree: 11 Somewhat agree: 16 Neither agree or disagree: 1 Somewhat disagree: 2 Disagree: 0 Strongly disagree 1
Familiar with text prediction	Strongly agree: 2 Agree: 10 Somewhat agree: 12 Neither agree or disagree: 1 Somewhat disagree: 6 Disagree: 5 Strongly disagree 3

Table 1: Survey Demographics Profile

Data Analysis
We use SPSS to perform a two-step analysis on the data collected via the survey. First, a paired samples t-test was run to examine whether our participants evaluated the two systems differently. A significant t-value will indicate a significant difference between the ratings on the two systems. Second, we employed ANOVA to test the effects of the demographics items on the participants' perceived privacy when using the prediction system. In our study, we have seven demographics items and would like to understand whether those demographics items have an effect on people's perceptions of privacy when using the prediction systems.

5. RESULTS

This section breaks down the results for each step of the data analysis process. The results of a paired samples t-test were reported first to compare the two systems. T-test was chosen because of its ability to adapt to smaller sample sizes (De Winter, 2013). The results of ANOVA between different types of demographics and perceived privacy were reported next.

Comparison of Two Systems
To test the difference between OpenAI and AI21 language models, we employed a paired samples t-test using SPSS to compare the 5 rating questions (J1 – J5 to O1 – O5) between the two systems (see Appendix). Table 2 shows the descriptive statics of each pair. The means from the Open AI model are less than the ones from the AI21 Jurassic model, which means the ratings to the Open AI model are more positive compared to the AI21 model.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	J1	4.55	31	1.690	.304
	O1	2.32	31	1.579	.284
Pair 2	J2	4.71	31	1.865	.335
	O2	2.81	31	1.833	.329
Pair 3	J3	5.26	31	1.751	.314
	O3	2.71	31	1.697	.305
Pair 4	J4	5.03	31	1.703	.306
	O4	2.58	31	1.876	.337
Pair 5	J5	4.94	31	1.731	.311
	O5	2.68	31	1.815	.326

Table 2: Paired Samples Statistics

The results indicate a significant difference on each rating item between the two systems (see Table 3). We found that in all cases, our participants preferred the responses generated by the OpenAI model over the responses generated by the AI21 model. To be specific,

- For the pair J1-O1, the results indicate a significant difference regarding information usefulness between the two systems ($t = 5.668$, $p = .000$). The participants believe that the OpenAI model provides more useful information than the AI21 model.
- For the pair J2-O2, the results indicate a significant difference regarding information

reliability between the two systems (t = 4.522, p = .000). The OpenAI model was found to provide more reliable information compared to the AI21 model.

- For the pair J3-O3, the results show a significant difference regarding information sufficiency between the two systems (t = 6.722, p = .000). The OpenAI model was rated higher by our participants regarding information sufficiency.
- For the pair J4-O4, the results show a significant difference on user satisfaction between the two systems (t = 5.917, p = .000). The participants were more satisfied with the information provided by the OpenAI model than the one provided by the AI21 model.
- For the pair J5-O5, the results indicate a significant difference on information quality between the two systems (t = 5.740, p = .000). The participants rated the information provided by the OpenAI model with a higher quality compared to the one provided by the AI21 model.

We, therefore, state that there is a difference regarding overall ratings between the two systems and conclude that the OpenAI model is rated significantly higher than the AI21 model.

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	J1 - O1	2.226	2.186	.393	1.424	3.028	5.668	30	.000
Pair 2	J2 - O2	1.903	2.343	.421	1.044	2.763	4.522	30	.000
Pair 3	J3 - O3	2.548	2.111	.379	1.774	3.323	6.722	30	.000
Pair 4	J4 - O4	2.452	2.307	.414	1.605	3.298	5.917	30	.000
Pair 5	J5 - O5	2.258	2.190	.393	1.455	3.062	5.740	30	.000

Table 3: The Paired Samples t-test Results

Effects of Demographics on Perceived Privacy

We did an ANOVA test between different types of demographics, which include gender, marital status, age, education level, occupation, familiarity in comparison to the nine items of perceived privacy (dependent variables)(see Appendix). To be specific,

- Utilizing gender (D1) as an independent variable, the results showed that there was a statistically significant difference in the responses to four items (P2, P3, P5, and P9) between the gender groups (P2: F (1, 29) = 4.176, p = .050; P3: F (1, 29) = 5.877, p = .022; P5: F (1, 29) = 5.100, p = .032; P9: F (1, 29) = 8.209, p = .008).
- Utilizing marital status (D2) as an independent variable, the results revealed that there was a statistically significant difference in the responses to P9 between the marital status groups (F (3, 27) = 5.717, p = .004).
- Utilizing age (D3) as an independent variable, there was no significant difference in any perceived privacy items between the age groups.
- Utilizing educational level (D4) as an independent variable, the results revealed that there was a statistically significant difference in the responses to P9 between the education groups (F (3, 27) = 4.065, p = .017).
- Utilizing occupation (D5) as an independent variable, the results revealed that there was a statistically significant difference in the responses to three items (P1, P3, and P9) between the occupation groups (P1: F (3, 27) = 3.370, p = .033; P3: F (3, 27) = 3.006, p = .048; P9: F (3, 27) = 3.007, p = .048).
- Utilizing familiarity with AI (D6) as an independent variable, there was no significant difference in any perceived privacy items between these groups.
- Utilizing familiarity with prediction systems (D7) as an independent variable, the results showed that there was a statistically significant difference in the responses to five items (P5, P6, P7, P8, and P9) between the groups (P5: F (6, 24) = 5.178, p = .002; P6: F (6, 24) = 3.592, p = .011; P7: F (6, 24) = 3.150, p = .020; P8: F (6, 24) = 4.175, p = .005; P9: F (6, 24) = 4.067, p = .006).

We conclude that at least some demographics including gender, marital status, education, occupation, and familiarity with text prediction have an effect on people's perceived privacy when using prediction systems.

6. DISCUSSIONS AND LIMITATIONS

The paired samples t-test reveals a significant difference between the two systems on the five overall rating items including usefulness, reliability, sufficiency, user satisfaction, and quality. The OpenAI model was rated significantly higher than the AI21 model by our participants for all the five items. A comparison between the two systems when using the same input is shown in Figure 3. The responses from both systems seem reasonable. When comparing them, the responses from OpenAI are much more detailed. For instance, when a user types in "life is like", the AI21 model will respond with "a box of chocolates", a well-known quote from the movie Forrest Gump which follows by "you never know what you're going to get." It is assumed that people who have watched the movie understand the meaning of this phrase. But for those who never watched the movie, it may be hard for them to comprehend this phrase. It is likely that those people would rate the information provided by the AI21 model not useful. On the other hand, the responses from the OpenAI model are more straightforward. The responses are stand-alone and easy to understand. The comparison between the two systems next to each other using the same input helps explain the findings of our study. Users are more satisfied with the information provided by the OpenAI model which is more useful, sufficient, reliable, and high quality compared to the one provided by the AI21 model.

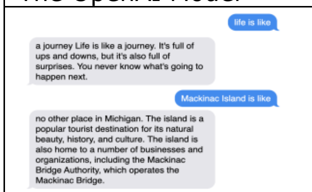
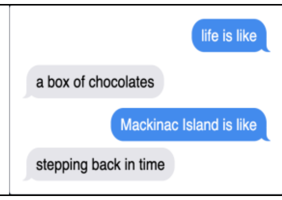
The OpenAI Model	The AI21 Model
	

Figure 3 A Comparison between the Two Systems Using the Same Inputs

We also examined the effects of demographics on perceived privacy. The results show that there is a significant difference on people's perception of privacy between the demographics groups including gender, marital status, education, occupation, and familiar with text prediction. Between different gender groups (D1), the means of the responses to four items (P2, P3, P5, and P9) were significantly different. Although ANOVA cannot tell us which groups are different from each other, the results suggest that gender plays a role in people's perceptions of privacy when using prediction systems. When comparing the effects of marital status (D2) and education (D4) on perceived privacy, we found that there was a significant difference on people's perception whether the AI web application has their best interest in mind (P9) between those demographic groups. Other than that, we did not find any difference on other perception items between those groups. A possible explanation might be that our sample does not have enough representatives for each group in D2 or D4 and thus no difference was detected between the groups. Between different occupation groups (D5), the means of the responses to three items (P1, P3, and P9) were significantly different. The results suggest that different occupation groups have significantly different trust on the system using their personal information properly. When comparing the effect of familiarity with prediction system (D7), we found that there was a significant difference on perceptions of saving time (P6), convenience (P7), and risk (P8 and P9). The results indicate that familiarity with prediction system plays a role in people's perceptions when using such a system. It seems reasonable that user's familiarity of technology impacts their perceived benefits of using technology.

It is interesting to find that age (D3) or familiarity with AI (D6) does not have any effect on the perception items at all. This might be caused by our sample. Our sample did not spread out in the age groups. 90% of the participants indicated that they were familiar with AI, which means only 10% of the participants belong to the group labeled with unfamiliarity with AI. Such unbalanced groups may contribute to the findings that D6 does not influence perceived privacy.

There are several implications. For academics, we developed two text prediction systems using different language models. The systems were tested and evaluated. The findings shed light on how to improve the performance of text prediction systems. There are also implications for practitioners in the field of NLP. The effects of demographics on people's perceived trust, risk, and privacy on a text prediction system can be further explored to better understand how to design text prediction systems.

There are some limitations in this study. While the sample size is acceptable, a much larger sample size would give more reliable statistical results. The ANOVA results show that there is a significant difference between demographics groups on people's perceived privacy. But one of the limitations of ANOVA is that it won't tell you which statistical groups were different from each other. For instance, we found a significant difference on perceived risk between groups based on familiarity with prediction system. But it is unclear which groups had a different mean unless additional test is conducted. We will address such limitation by running additional test to better understand the effects of demographics on people's perceived privacy when using prediction systems. Additionally, testing the antecedents and factors and assigning weights to these factors to study why participants favored on set of responses is a direction for future research.

7. CONCLUSION

Natural language processing is a large domain, text prediction is one of the major tasks natural language processing presents. This research project has presented two language models;

Jurassic-1 and the GPT-3 language model, tested their efficiency, accuracy, reliability, and trust in artificial intelligence. Participants were able to interact with the two language models and later on asked a few questions about how their experience was in using the application. Paired sampels t-test and ANOVA were then done to test which language model is better and if demograogics influence perceived privacy in text prediction.

8. REFERENCES

Crockett, K., Garratt, M., Latham, A., Colyer, E., & Goltz, S. (2020, 19-24 July 2020). *Risk and Trust Perceptions of the Public of Artifical Intelligence Applications.* Paper presented at the 2020 International Joint Conference on Neural Networks (IJCNN).

De Winter, J. C. (2013). Using the Student's t-test with extremely small sample sizes. Practical Assessment, Research, and Evaluation, 18(1), 10.

González-Carvajal, S., & Garrido-Merchán, E. C. J. a. p. a. (2020). Comparing BERT against traditional machine learning text classification.

Grujić, N. D., & Milovanović, V. M. (2019). Natural Language Processing for Associative Word Predictions. Paper presented at the IEEE EUROCON 2019-18th International Conference on Smart Technologies.

Hengstler, M., Enkel, E., Duelli, S. J. T. F., & Change, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. 105, 105-120.

Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. Decision Support Systems, 44(2), 544-564. doi:https://doi.org/10.1016/j.dss.2007.07.001

Panchbhai, A., & Pankanti, S. (2021, 28-29 Jan. 2021). Exploring Large Language Models in a Limited Resource Scenario. Paper presented at the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence).

Rojan, A., Alias, E., Rajan, G. M., Mathew, J., & Sudarsan, D. (2020). Natural Language Processing based Text Imputation for Malayalam Corpora. Paper presented at the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC).

Trnka, K., & McCoy, K. F. (2007). Corpus studies in word prediction. Paper presented at the Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility.

Wang, A., & Cho, K. J. a. p. a. (2019). Bert has a mouth, and it must speak: Bert as a markov random field language model.

Xu, H., Liu, B., Shu, L., & Yu, P. S. J. a. p. a. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis.

Yu, S., Su, J., & Luo, D. (2019). Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge. IEEE Access, 7, 176600-176612. doi:10.1109/ACCESS.2019.2953990

APPENDIX

Constructs	Measurement items
Demographics	Gender (D1) Marital Status (D2) Age (D3) Educational level (D4) Occupation (D5) Overall, I am familiar with Artificial intelligence(D6) I am familiar with text prediction using artificial intelligence(D7)
Overall Rating	Overall, I think the website provides useful information (J1/O1) This website provides reliable information (J2/O2) This website provides sufficient information when I try to type any text (J3/O3) I am satisfied with the information this website provides (J4/O4) Overall, the information that website provides is of high quality (J5/O5)
Perceived Privacy	I am concerned that: The web vendor will use my personal information for other purposes without authorization (P1) This website is collecting too much information from me (P2) The web vendor will sell my personal information to others without my permission(P3) This website vendor gives the impression that it predicts texts accurately (P4) This text prediction site is trustworthy (P5) I can save time using this website(P6) I think using this web application for text prediction is convenient (P7) How would you rate your overall perception of risk from this site (P8) I believe the artificial intelligence web application has my best interest in mind (P9)

QR Code Hacking – Detecting Multiple Vulnerabilities in Android Scanning Software

Joe Homan
joehoman3@gmail.com
Former President & COO
Stephenson Technologies Corporation

Jennifer Breese
jzb545@psu.edu
IST and Cybersecurity Analytics and Operations
Pennsylvania State University

Abstract

The COVID-19 pandemic created a need that increased the use of Quick Response (QR) codes. The need to minimize contact with items handled by multiple people, for example, drove restaurants to replace menus with QR codes on tabletops. Scanning a QR code with a smartphone camera presents the user with the option to open a website URL displaying something simple as a restaurant menu or to check-in to health care and trusted payment applications. As with other technology solutions, QR codes come with risks. Accessing a website from a smartphone creates the potential for a cybersecurity hack. For this project the team conducted primary research and investigated cyber risks associated with QR codes. We tested multiple QR code scanning software and documented multiple threat vectors present in many of the publicly available QR code scanning software products. Initial research focused on Android smartphone applications.

Keywords: QR code, Internet of Things hacking, cybersecurity risks, Android smartphones, security awareness.

1. INTRODUCTION

A diner sits down at their favorite restaurant for a nice meal. Thanks to the pandemic, the physical menu is replaced by a Quick Response (QR) Code on a sticker at the corner of the table or a pop-up placard. The hungry diner scans the QR code with their smartphone camera and click on the website to open the menu. As they peruse the menu to make their dinner selection, their personal information is downloaded in the background. The diner's entire contact list, calendar schedule, and other data is being transmitted to a remote site. Without knowing it, the unsuspecting consumer has just been hacked.

Through a project funded by the US Air Force Research Laboratory, researchers at Stephenson Technologies Corporation, a non-profit affiliate of Louisiana State University conducted primary research on hacking of various devices and technologies. In part, this research effort aimed to identify potential threats created by IoT devices, inform the manufacturers of the vulnerabilities, and provide awareness to users and consumers of these devices and technologies.

While QR codes have been around for almost 30 years (Pandey, 2008), their ability to transmit data in a contactless and touchless manner has increased their use during the COVID pandemic. Yet, the risks they pose from a cybersecurity standpoint should be seriously considered before use. This may sound like science fiction or a scare tactic to get consumers to avoid using QR codes, but it is a real threat. QR codes are finding increased use in more industries and media outlets including "commercial tracking systems, entertainment, in-store product labels, marketing, traditional print newspapers, television broadcasting, traditional book publishing and websites" (Akta, 2017, p. xxiii) This paper will present the primary research conducted on QR code vulnerability, the results of analysis of QR code software, identification of specific software applications with inherent vulnerabilities, and recommendations for further research.

Figure 1 1D Barcode

2. QR Codes in Use

What is a QR code? Essentially it is like the barcode that is commonly used on grocery items and scanned at the checkout register. Barcodes (Figure 1) code information in only one direction (one dimension), while QR codes store information in 2D, that is in two directions: left/right and up/down (Figure 2). The advantage of 2D QR codes over 1D barcodes is that they can hold much more information. A 1D barcode can hold 85 characters while a 2D barcode "can contain 7,000 digits of characters at maximum including Kanji characters" (Pandey, 2008, p. 60) Barcodes continue to be prevalent in grocer and other applications and 2D codes have become more popular in advertisements (as witnessed on television during the 2022 Super Bowl), automated systems, transportation, and restaurants.

Newer QR code or similar 2D code designs have been developed or proposed to increase data storage. For example, "a new high capacity color barcode, named HCC2D (High Capacity Colored 2-Dimensional), which use colors to increase the barcode data density" (Querini et al., 2011, p. 136) has been proposed and the authors considered the hacking and security aspects of such barcodes. Other research into designer QR codes (which may include images like corporate logos) has studied use of color and contrast on the ability of the QR code to be recognized but did not consider hacking or security aspects (Berisso, 2013).

Whether the QR code hacking is related to the release of personally identifiable information (PII) in the healthcare arena (Ang, 2021), in high-tech artificial intelligence applications (Wahsheh & Al-Zahrani, 2021) or something as mundane as opening a menu in a restaurant, the risks are real, and consumers and users should be aware of the dangers associated with QR code use. Depending upon the security features of the software application that reads the QR code, data on the device may be safe – or exposed (Wahsheh & Luccio, 2020).

Figure 2 2D QR code

3. The Problem

QR codes can be used to hack users' electronic devices, most notably – smartphones and tablets. Both QR codes and smartphones have become ubiquitous, the former especially during the pandemic. Increased usage of any technology that touches personal, commercial, industrial, or military systems creates an opportunity for hackers. Most cybersecurity attacks revolve around hackers attempting to improperly access information (personal or corporate) on individual devices (computers, smartphones, tablets, etc.), i.e., phishing. Previous research theorized about the "dangers of possible attacks utilizing manipulated QR codes" and the expectation "that this kind of attack will receive more and more attention by the hacking community in the future. (Kieseberg et al., 2012, p. 37) These authors outlined the concepts that may lead to hacking but did not directly test various methods or prove the feasibility of hacking a QR code.

The focus of this research was to determine if it is possible to perform more advanced attacks with QR codes in smartphone applications, specifically those with the Android operating system. These attacks included executing malicious commands on a victim's smartphone. Such commands could be used to carry out more advanced attacks (e.g., command injection). Based on our initial research, it appeared there was little research done in this domain. As noted by (Atka, 2017) "[n]ot much literature exists on QR (Quick Response) Codes and their applications in the emerging digital society." One early study noted "it is not likely that users will be able to find out easily the content encoded, typically URLs, until after they scan QR codes. This makes QR codes a perfect medium for attackers to conceal and launch their attacks based on malicious URLs." (Yao & Shin, 2013, p. 341) Other researchers have suggested that QR codes can be used as an attack vector for SQL injection or command injection. (Focardi, 2019) Previous research documented attacks as largely theoretical, experimental, or simulated rather than actually testing the limitations and risks associated with QR codes. (Zhang et al.,2021; Mavroeidis & Nicho, 2017)

Beyond academic research and theoretical suggestions of hacking dangers, our literature review revealed minimal direct testing and awareness surrounding the real risks associated with QR codes. The initial focus of our research was on achieving various types of command injection with QR codes on the Android smartphone platform. Our hypothesis was that QR codes could be used to inject commands into applications that do not perform proper input sanitization to prevent malicious activity. Our project focused on achieving various types of command injection with QR codes on the Android platform

4. Research Approach

QR codes only store data. They are a static image and do not do any processing of the data. Software on the device that scans the QR code reads the data and directs the device to the appropriate website (Pandey, 2008). The software should have appropriate controls in place to prevent phishing attempts and not execute malware that a QR code could point to. If the scanning software application is not secure, it can execute malicious commands packaged inside of the QR code (Focardi, Luccio, & Wahsheh, 2019).

The approach of our research effort was to identify potential methods of hacking using QR codes to execute malicious commands on various devices. We obtained and tested multiple QR code software applications to evaluate them for vulnerabilities to multiple hacking methods. Having the highest market share globally makes Android a target for attacking by exploiting vulnerabilities. Further, due to the complexity and specialization of the vulnerabilities, few users can relate them to their mobile devices (Wu et al., 2015). This research began with Android devices and could be applied to other operating systems including those on Windows and IOS smartphones.

5. Methodology

Multiple categories of potential hacks and risks were identified to test for potential security issues. These attack vectors and a brief description of each is as follows:

- HTML injection: In this hack, the application gets tricked into parsing html within the QR code, causing unintended appearance modifications. HTML (HyperText Markup Language) is used for formatting text on websites, so during our testing, we could do things such as change the color, size, and style of text inside of the application. (Figure 3)

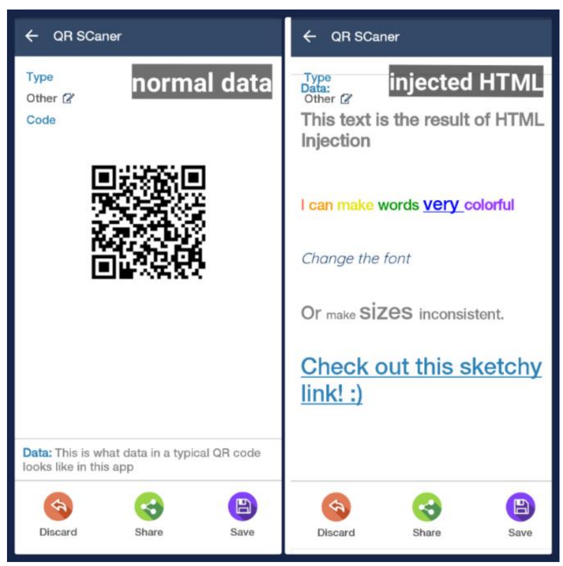


Figure 3 HTML injection

- SQL injection: In this attack, the application gets tricked into executing SQL (Structured Query Language) commands on a database. SQL can be used to insert and delete data inside of a database. With this attack, we were able to insert entries into the user's "scan history" database, which is a database that shows them all previous scan results from QR codes they have scanned. By putting an insert statement in the QR code, we were able to insert dozens of malicious websites directly into their scan history. (Figure 4)

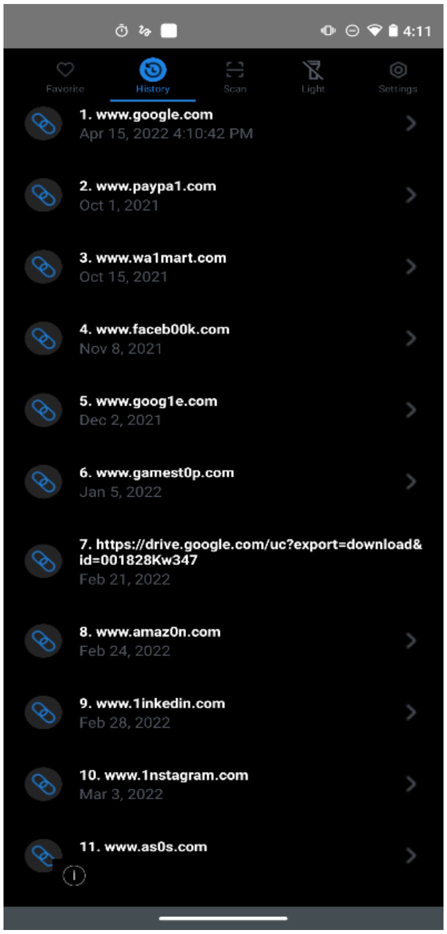


Figure 4 SQL injection

- Javascript injection: This hack only works for vulnerable web based QR code scanners, but the app gets tricked into executing JavaScript within the QR code. It is important to note that our research did not find an Android-based device application published in the Google PlayStore that was vulnerable to this type of injection. Web-based scanners are less common but to prove the vulnerability is real, our researchers wrote a demo QR Code scanner that contained this vulnerability. This allowed us to change the HTML using JavaScript, causing drastic appearance changes. Future research will focus on investigating a spyware concept using Javascript injection.

- Unsupported encoding modes: By standard, QR codes only support numeric, alphanumeric, kanji, and byte encoding modes. We created a few methods that allowed other types of characters (such as Unicode) to work inside of QR codes, but the QR code scanner must include support for it. We experimented with putting some of these Unicode symbols inside of the QR code to see if scanning them would bring about unintended behavior. This would allow us to bypass any input checking the QR code software was doing.

- Special symbols: The collection of QR codes that we created didn't target any specific type of injection but included some symbols that had significance across a wide range of programming languages. This is the group of codes we used to test and prove that the above attacks that were possible.

Research included symbols that have significance in various languages. The basic testing methodology involved the following steps:

- Feed QR codes into a sample of scanning applications and observe behavior
- Determine if behavior was as expected or if abnormal or inappropriate actions took place
- Once suspicious behavior was found, pass the application through reverse engineering and debugging tools
- Analyze logs, decompile Java, and evaluate Java Virtual Machine (JVM) bytecode to pinpoint the issues and potential vulnerabilities
- Craft more QR codes to exploit the vulnerabilities, once identified

The following process was followed in conducting this research:

1. Created a folder filled with QR codes that contained special symbols. These are the same “special symbols” discussed above. They contained characters like “;”, “<”, “ ‘ ”, and others that have significance in programming languages.

2. Scanned these QR codes with a wide range of QR code scanners (both default scanners preinstalled on the phone and ones downloaded from the Google PlayStore) and observed how they responded.

3. Observed that a collection of scanners was not responding to the "<" symbol, and a group of other scanners would crash when the camera scanned the " ' " symbol.

4. Ran the problematic applications through a decompiler, which is a software that tries to reconstruct the original application code. Decompilers work like a compiler, but in the reverse direction. This let us pinpoint exactly where the abnormal behavior was occurring, why, and whether it was due to a vulnerability.

5. It was determined that the applications that weren't responding to "<" had an HTML injection vulnerability. The app was reading the symbol as an HTML tag which has the general format <tag>contents</tag>. The QR code software interpreted the symbol in the QR code as a part of its own app code. The software did not recognize anything inside of the QR code for it to process and therefore did not take the appropriate action.

6. The applications that were crashing as a result of scanning " ' " had a SQL injection vulnerability. The symbol is used inside of SQL statements. The application was carelessly passing the QR code text directly into a SQL statement. When the QR code text contained " ' ", it created a syntax error in the statement and caused the app to crash. As in the HTML injection case, the app was tricked into thinking the symbols we inserted in the QR were a part of its application code and was trying to execute it.

7. As we learned more about the vulnerabilities, we constructed QR codes that exploited these vulnerabilities as demonstrated in Figures 3 and 4.

For the JavaScript injection, we created the demo web-based scanner as a proof of concept. We did so by writing a web application that was susceptible to JavaScript injection, then loading it into an android app we created using WebView. WebView is a technology used by Android application developers to turn an existing website into an application. Many developers are attracted to this since they can create a mobile phone application by recycling website code. Once this was set up, we were able to prove that arbitrary JavaScript could be injected through QR codes if scanners are designed this way.

6. Results

Based on initial research, there were no observed vulnerabilities in native QR code scanning software (i.e., the applications built into Android devices). However, multiple applications downloaded from the Google PlayStore were susceptible to HTML and SQL injection. Figure 5 graphically depicts the types of QR code software vulnerability types.

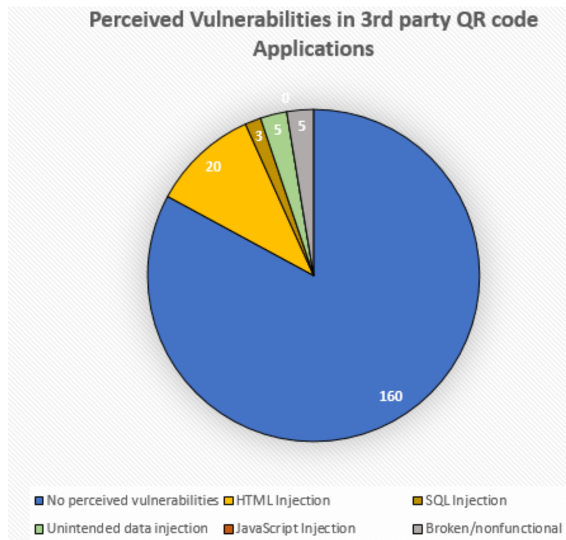


Figure 5 QR Code Software Vulnerability Types

Using the methodology described above, the following steps and results were observed:

- Wrote an experimental QR code scanner that executes JavaScript commands
- At least one of the scanners in the Google PlayStore appeared vulnerable to a JavaScript injection
- A small number of apps crashed or gave unexpected behavior
  - For some, due to poorly written code
  - For others, due to injection vulnerabilities
- After more reverse engineering, crafted QR codes to achieve injection
- Types achieved: HTML, SQL, JavaScript
  - No vulnerabilities found in native QR code scanners
  - Some third-party scanners were found to have HTML and SQL injection vulnerabilities

- These vulnerabilities were a result of not sanitizing input properly and passing it directly to Android application programming interface (API) calls such as from Html and exec SQL.
- If the application uses an Android Webview, it is susceptible to extensive HTML injection and even Javascript injection, if the webpage it uses is susceptible to these things.

- innerHTML: allows insertion of arbitrary html tags
- JavaScript enabled: allows for JS injection
  - Extensive HTML injection was found in one app on the Google PlayStore. JavaScript injection vulnerabilities weren't found in the wild, but we were able to achieve it experimentally

7. Responsible Disclosure

Responsible disclosure is a process that cybersecurity researchers typically follow to report vulnerabilities discovered during their research to the appropriate parties. The intent is to allow software manufacturers (in this case) to address vulnerabilities prior to research being published and the risks being disclosed to the public (and, of course, hackers). Although there is no standard period defined between notification and publication of research, a minimum of 30 days is recommended.

App Name	Company Name
AiScanq	ViewAI Lab
Code Scanner	QR Code Apps
QR & Barcode Scanner	Rajshah5599
QR &Barcode Scanner	BondRen
QR and Barcode Scanner	Apps360 Team
QR and Barcode Scanner	Solution App Technology
QR Code	Faysal INC
QR Code Copy and Save	FalsinSoft
QR Code Scan	ZipoApps
QR Code Scanner	bestdeveloperteam
QR Code Scanner	Hotapp Studio
QR Code Scanner Plus	Disneysoft
QR Scanner	1MB Apps Studio
QR Scanner	App Karo
QR Scanner	TOH Talent Team
QR X2 - Scan QR & Barcode	Easy To Use (OnMobi)
QR/Barcode Scanner Pro	Smart Scanner
Quick-QRScan	Madhivanan

Figure 6 QR Code Software Vulnerable to HTML Injection

All manufacturers of QR code software that demonstrated a vulnerability in the three categories identified in this paper were notified of one or more vulnerabilities detected in their software. The notifications were sent in February 2022, and all were given the opportunity to contact our researchers for additional details associated with their vulnerability.

Figures 6, 7, and 8 identify companies and their application in the three categories of vulnerability – namely HTML injection, SQL injection, and data injection.

App Name	Company Name
QReader3	Dmitry Ventures
QRMagic	Boris Expert
QR Code Scanner	MobMatrix Apps

Figure 7 QR Code Software Vulnerable to SQL Injection

App Name	Company Name
QReader3	Dmitry Ventures
QRMagic	Boris Expert
QR Code Scanner	MobMatrix Apps

Figure 8 QR Code Software Vulnerable to Data Injection

8. Conclusion

QR code users may not see a cybersecurity risk from that small square patch of dots. Many who have become accustomed to using them to view restaurant menus do not think twice about continuing to do so. However, the software reading QR codes may pose a bigger threat than users ever imagined. If scanning applications do not have the necessary security protocols in place, the user may be at risk of data loss.

An easy way for a hacker to attack a smartphone would involve a simple switch as follows:

1. The hacker uses the correct QR code to access a restaurant's menu.
2. The menu is copied and duplicated on a different website.
3. A new QR code is created that directs the scanning software to the fake website.
4. The new QR code is pasted over the restaurant's QR code at the table.
5. The diner scans the QR code and sees the menu on the fake website.
6. The malicious site uses HTML, SQL, or data injection to infect the smartphone and download personal data like contacts, calendar events, email, etc.
7. The user orders their meal completely unaware that they have been phished.

It is vital to educate developers about the importance of software development using secure development operations. The SecDevOps process prioritizes delivery speed, integrated security, and system quality. The process is a continuous integration/delivery pipeline that weaves security awareness, practices, and testing through the fabric of the development, testing, and deployment process instead of incorporating it into just one phase like a DevOps process. Software developers must design their code to check inputs to a system to prevent injection of malicious code. This will help strengthen security of current APIs offered to consumers. Equally, consumers need to understand the risks associated with third party applications for software like QR code readers.

This primary research was conducted by engineers, analysts, and a student intern at Stephenson Technologies Corporation in Baton Rouge, LA. The research, partially funded by a US Air Force Laboratory contract, uncovered the vulnerabilities identified in this paper. As noted, we followed the responsible disclosure protocol by informing the manufacturers of the QR code software containing vulnerabilities so that they could address the issues before the publication of this paper.

9. Future Work

This research was limited to the Android operating system. Future research is needed to repeat this work for Windows and IOS QR code scanners. Also, this research focused on smartphone software applications. Future research is also needed to assess QR code scanners used in other applications such as those used in airports, public transportation, banks, factories, hospitals, warehouses, etc.

10. References

Ang, R. J. (2021). Encrypted QR Code in Healthcare Applications. *Canadian Journal of Nursing Informatics*, *16*(2).

Akta, C. (2017). *The Evolution and Emergence of QR Codes* (1st. ed.). Cambridge Scholars Publishing, Newcastle upon Tyne, GBR.

Berisso, K., (2013), Designer QR Codes; Ensuring the “beep,” Ohio University Department of Engineering Technology and Management White Paper.

Focardi, R., Luccio, F. L., & Wahsheh, H. A. M. (2019). Usable security for QR code. *Journal of Information Security and Applications*, *48*, 102369. https://doi.org/10.1016/j.jisa.2019.102369

Kieseberg, P., Schrittwieser, S., Leithner, M., Mulazzani, M., Weippl, E., Munroe, L., & Sinha, M. (2012). Malicious pixels using QR codes as attack vector. *Atlantis Ambient and Pervasive Intelligence*, 21–38. https://doi.org /10.2991/978-94-91216-71-8_2

Mavroeidis, V., Nicho, M. (2017). Quick Response Code Secure: A Cryptographically Secure Anti-Phishing Tool for QR Code Attacks. In: Rak, J., Bay, J., Kotenko, I., Popyack, L., Skormin, V., Szczypiorski, K. (eds) Computer Network Security. MMM-ACNS 2017. Lecture Notes in Computer Science, vol 10446. Springer, Cham. https://doi.org/10.1007/978-3-319-65127-9_25

Mohammed S. Al-Zahrani, Heider A. M. Wahsheh, Fawaz W. Alsaade, "Secure Real-Time Artificial Intelligence System Against Malicious QR Code Links," *Security and Communication Networks*, vol. 2021, Article ID 5540670, 11 pages, 2021. https://doi.org/10.1155/2021/5540670

Pandey, D. (2008). Synthesis Journal, Retrieved from https://www.academia.edu/31427962/-Three_QR_Code

Park, H., Kim, T., Kim, G., Jang, W., Lee, K., & Lee, S.-Y. (2019). Improvement of QR code access control system based on Lamport Hash Chain. *Innovative Mobile and Internet Services in Ubiquitous Computing*, 824–833. https://doi.org/10.1007/978-3-030-22263-5_79

Querini, M., Grillo, A., Lentini, A., & Italiano, G. (2021). 2D Color Barcodes for Mobile Phones, *International Journal of Computer Science and Applications*, Technomathematics Research Foundation, Vol. 8 No. 1, pp. 136 – 155.

Wahsheh, H. A. M., & Al-Zahrani, M. (2021). Secure real-time computational intelligence system against malicious QR code links. International Journal of Computers, Communications and Control, 16(3). https://doi.org/10.15837/ijccc.2021.3.4186

Wahsheh, H. A. M., & Luccio, F. L. (2020). Security and Privacy of QR Code Applications: A Comprehensive Study, General Guidelines and Solutions. *Information*, *11*(4), 217. MDPI AG. Retrieved from http://dx.doi.org/10.3390/info11040217

Wu, J., Wu, Y., Yang, M., Wu, Z., Luo, T., & Wang, Y. (2015). QRCloud: Android Vulnerability Query and Push Services Based on QR Code in Cloud Computing. *Journal of Computational Information Systems, 11*(11), 3875–3881.

Yao, Y and Shin, D. (2013). Towards preventing QR code based attacks on android phone using security warnings. In Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security (ASIA CCS '13). Association for Computing Machinery, New York, NY, USA, 341–346. https://doi.org/10.1145/2484313.2484357

Zhang, J., Liu, S.-J., Pan, J.-S., & Ji, X.-R. (2018). Digital Certificate based security payment for QR code applications. *Proceedings of the Fourth Euro-China Conference on Intelligent Data Analysis and Applications*, 88–97. https://doi.org/10.1007/978-3-319-68527-4_10

Short Stay Healthcare Quality in Skilled Nursing Facilities: Occupancy, Nurse Staff Mix, and COVID-19

Elaine Winston
Elaine.r.winston@hofstra.edu
Hofstra University
Hampstead, NY

Jason Xiong
xiongjj@appstate.edu

Dawn Medlin
medlinbd@appstate.edu

Appalachian State University
Boone, NC

Alex Pelaez
Alexander.Pelaez@hofstra.edu
Hofstra University
Hampstead, NY

Abstract

Recently, there has been a shift in the Skilled Nursing Facility (SNF) industry to provide short stay health care services, rather than long-term care. The shift to short-term care and the availability of government public domain databases of quality measures have become focal points of healthcare discussions during the COVID-19 pandemic. This study explores COVID-19 pandemic outcomes by analyzing government short-stay quality metrics, occupancy rates, and nurse staff mix levels in SNFs. A statistical analysis provides results as predictive factors of short-term quality. These factors are used to explore COVID-19 outcomes. The study indicates that a higher number of Certified Nursing Assistants (CNAs) and Licensed Practical Nurses (LPNs) in relation to the number of Registered Nurses (RNs) suggests better COVID-19 healthcare outcomes. A higher occupancy rate also suggests better COVID-19 outcomes. Additionally, nurse staff who understand Information and Communication Technologies (ICTs) may help improve short-stay quality healthcare outcomes by providing critical information to healthcare administrators and governmental policymakers.

Keywords: COVID-19, Skilled Nursing Facilities, Nurse Staff Mix, Short Stay Quality Measures.

1. INTRODUCTION

The skilled nursing facility (SNF) provides both long-term care and short-term care (i.e., stroke and surgery rehabilitation) and has become a focal point for discussions concerning

government quality standards. An SNF can provide a long-term residence for people who require continual nursing care and have significant difficulty coping with activities of daily learning (ADL). Residents may include the elderly as well as younger adults with physical or mental disabilities. Additionally, SNFs can provide short-term post-acute care (PAC) for multi-faceted medical conditions.
The SNF industry is composed of 67% privately owned, 7% government-managed, and 24% non-profit organizations (Centers for Medicare & Medicaid Services, 2020). Although privately-owned organizations dominate the industry, the government controls the reimbursement of services based on government-determined quality metrics.

The number of residents in SNFs for long-term care has been declining in the last several years, with the percentage of people 65 years and older in nursing facilities decreasing by 24.5% between 2004 and 2010 (Laes-Kushner, 2018). Some suggest that the decline in occupancy is due to the low-quality of care received by patients at SNFs (Winston et al., 2019; Winston et al., 2021). Current studies indicate that occupancy rates are lower than pre-pandemic rates and that there are many empty beds (Kauffman, 2020). An essential problem of low occupancy rates is that it reduces the revenue that an SNF receives.

Recently, there has been a growing shift in SNFs to provide short-term care rather than long-term care. The increasing need for SNFs to achieve high-level short-stay quality measures (SSQM) relies on a network that is comprised of SNFs, hospitals, and community services. Further, government policies have been shown to impact the quality provided at SNFs through incentives. For example, The Improving Medicare Post-Acute Care Transformation (IMPACT) Act, is a result of the increase in the number of patients transferred from an acute hospital to SNF care. This transformation depends on information communication technology (ICT). SNFs rely on data-driven healthcare services.

First, literature on the SNF industry transformation from providing activities of daily living (ADL) care to rehabilitative post-acute care (PAC) is presented. This includes government technologies and policies that directly affect the quality of healthcare services. Occupancy rate and nurse staff mix (i.e., certified nurse assistants, licensed practitioner nurses, and registered nurses) are reviewed in the context of short-stay healthcare quality. Information on COVID-19 in SNFs is presented. A statistical analysis is then conducted to identify any relationship among occupancy rate, nurse staff mix, and SSQM. Results are presented. Results, which are based on the pre-pandemic data, the control, are used to explore COVID-19 outcomes at SNF during the first year of the pandemic. Conclusions and potential future research are presented.

2. LITERATURE REVIEW

Public domain healthcare databases provide online data and reports of quality. This is particularly important for the highly competitive SNF industry.

Information Communication Technology

Information Communication Technology (ICT) (i.e., public domain and open-source health care databases) provide information on a myriad of factors indicating the quality of health care in organizations. These technologies incorporate methodologies for collecting and scrubbing data of personal characteristics. Concepts such as collection bias, data integrity, and accuracy help identify deviations from government standards. Currently, registered nurse (RN) staff in SNFs must have an understanding and knowledge of ICTs to enable government-required real-time reporting of occupancy levels, nurse staff hours per resident, and health progress.

Healthcare administrators who understand database and reporting technologies may analyze data to obtain high reimbursement rates by meeting the government-established metrics for quality, such as the SSQM. The Centers for Medicare and Medicaid Service's public domain provides readily accessible resources that serve as objective evidence of healthcare organization performance. Additionally, the government pays for superior performance through SNFs with exceptional outcomes.

The capability to find and compare SNFs within a community is aided by an official website of the government which provides five-star quality ratings on over 15,000 SNFs in the United States (Medicare, n.d.). Hospitals and communities use the digitized 5-star quality ratings to filter the SNF industry's overall assessment of quality.

Occupancy

Low occupancy has a negative economic impact on SNFs that operate on low profit margins. Research indicates that the financial performance of an SNF can affect the quality of healthcare services (Weech-Maldonado et al., 2019). There has been a growing shift in SNFs to provide short-term care rather than long-term care. Medicare reimburses only rehabilitative short-term care and at higher rates than Medicaid, which only reimburses for long-term care. The use of incentives to help direct healthcare service and effectiveness has been used in rural, impoverished areas, as well as, outside the United States (Baumann et al., 2021). SNFs that can provide post-acute care services have an economic advantage in the long-term healthcare industry.

SNF and acute care hospitals have become critical partners. Most recently, twenty percent of all hospitalized Medicare beneficiaries were discharged to an SNF for PAC (Hakkarainen et al., 2016; Yurkofsky et al., 2020). The significant increase in PAC services at the SNF requires attention to government policy that has a direct impact on short stay quality.

In October 2019, the Centers for Medicare and Medicaid Services (CMS) changed its pay-for-service (PPS) reimbursement rate methodology from resource utilization group (RUG) IV to a patient-driven payment model (PDPM). RUGs provided revenue primarily based on length of stay. The PDPM funds more complex nursing care and is focused on the type of rehabilitative services the patient requires (Harrington et al.,2020). This change to payment reimbursement led to a shift of high acuity (i.e., nurse intensive) attention patients from hospital to SNF. The hospital reduces the length of stay of patients (reduced length of stay is incentivized) while the SNF increases PAC services.

The change to PAC from long-term care has revenue and cost implications for the SNFs. Patients with complicated short-term medical needs (e.g., hip surgery and stroke) have higher reimbursement rates than long-term residents who require ADL (Li, 2020). However, the additional revenue of short-term care patients often does not cover the costs that are increased due to the requirements of care (Grabowski et al., 2012), such as nursing staff trained to provide complex healthcare services. Further, rehabilitative (short-term) patients result in greater turnover at SNFs. This results in a lower occupancy rate that eventually reduces revenue.

In addition, there is a growing disparity between quality metrics and healthcare because of the increasing number of new PAC patients being transferred from hospitals to SNFs, according to an Assistant Director of Nursing in a large metropolitan SNF. A new PAC patient describes a patient who is discharged to an SNF the day after an acute care procedure. These patients have a distinct set of concerns that differ from those patients who are discharged after a 3-day hospital stay.

The government provides a monetary incentive to hospitals for the transfer of fresh PAC patients to an SNF through two programs: 1) Bundled Payment - hospitals can apply for a waiver that will address a 3-day hospital stay required for SNF services post-discharge, and 2) Comprehensive Care for Joint Replacement that allows hospitals to use a 3-day stay waiver.

Other studies explored the risk of hospital readmission within 30 days of SNF admission and found a significant correlation with SNF quality ratings (Rahman, 2016), hours of physical therapy (Jung et al., 2016), days at the skilled nursing facility (Hansen et al., 2011), and a mismatch between patient needs and the services provided by SNFs (Stone et al., 2010; Hansen, et al., 2011).

The Hospital Readmissions Reduction Program reduces payments to hospitals with excess readmissions. Residents who return to a hospital following acute care are often placed under observation (OS) rather than readmitted. In 2015, 2.1 million people were in hospitals under OS status (Bunis 2015). The status of OS does not meet Medicare’s criteria for further reimbursement at an SNF. Therefore, only a small percentage of residents after observation can go back to the SNF. From a data-specific view, these residents cannot be considered as discharged back to the community from their short stay at the SNF, which is rewarded.

Before COVID-19, the SNF industry reported that 48.57 % of patients returned from an SNF to the community (Saliba et al., 2018). This means that less than half of the residents who use an SNF for short-term rehabilitation made it home. This low rate of return necessitates more study on the disparity between actual short-term health care residents receive at SNFs and the SSQM.

Nurse Staff Mix

In this study, nurse staff mix refers to the combination of three key categories of nursing personnel used in the SNF: registered nurses (RNs), licensed practical nurses (LPNs), and certified nursing assistants (CNAs). There is a positive relationship between the number of nursing staff who provide long-term ADL care to residents and the quality of care (Harrington, et al., 2020a), and higher levels of RN staffing led to better resident long-term outcomes (i.e., fewer pressure ulcers); improved ADL and lower mortality rates (Backhaus et al., 2014).

Government staffing metrics require that SNFs have enough RNs, LPNs, and CNAs to provide adequate nursing care for long-term care activities. The LPN performs nursing tasks related to ADL. LPNs are limited in the tasks they may perform for PAC patients (e.g., supervising the distribution of prescriptions and wound bandage maintenance). CNAs help with ADL activities. These assistants are trained to help the residents eat, stay clean and move, as required in the resident care reports. CNAs are not trained to perform any medical service.

"The government overall staff quality rating requires SNFs to have only one RN for at least 8 consecutive hours a day, 7 days a week, and a second designated RN to serve as the director of nursing on a full-time basis (except for low occupancy and/or small SNF)" (Public Health, 2022).

Only the RN is licensed to perform a diagnostic evaluation. This means as patients arrive at the SNF the RN with a medical advisory team (from the hospital) determines the level of PAC.

Inadequate staffing levels have significant consequences. Lower staff levels in facilities before the pandemic made these SNFs more vulnerable to the coronavirus, resulting in more than 28,000 U.S. nursing home residents and worker deaths reported at the end of May 2020 (Yourish et al., 2020; Mathews et al., 2020).

COVID-19 in SNFs

COVID-19 deaths in nursing homes assisted living and other long-term care facilities made up over a third of all U.S. deaths in 2020, based on the COVID Tracking Project (CTP) data, which includes (The COVID-Tracking Project, 2021). Yet, studies found no association between COVID-19 death rates and the SNF overall five-star quality rating or infection control deficiencies (Figueroa et al., 2020; He et al., 2020). A positive association was reported between COVID-19 deaths and low nurse staff ratings (Harrington et al., 2020b), for-profit ownership (Harrington et al., 2020b), and a high percentage of minority residents (Chidambaram et al., 2020).

Research indicates that adequate sick pay, minimal use of outside agency staff, increased staff-to-bed ratio, and staff cohorts (Shallcross et al., 2021) reduce the transmission of COVID-19 among nurses. The combination of high occupancy rates and low staffing levels was associated with increased risks of infection (Ochieng et al., 2021). Nurse staff levels are a critical aspect of COVID-19 care (Desroches et al., 2021). The number of new admissions to the facility and poor compliance with isolation procedures resulted in high viral loads and transmission rates (Gibson et al., 2020; Abrams et al., 2020).

Data indicates that SNFs do not have the healthcare professionals or the capabilities (i.e., PPE and technology) to manage COVID-19. The COVID-19 death rate in SNFs increased from 0.48 to 1.88, per 100 residents and new COVID-19 cases increased from 2.6 to 10.8, per 100 from the second to the fourth quarter of 2020 (Paulin, 2021). The increase in infection rates is instead of policies that after first requiring SNFs to admit COVID-19 patients (Khimm, 2020a) then prohibited SNFs from accepting residents diagnosed with COVID-19 (Khimm, 2020b).

The research questions for the following study are:

1: Is there an association between occupancy rate and short stay quality measures (SSQM)?

2: Is there an association between nurse staff mix, which is the ratio (CNA + LPN) / RN hours per resident per day, and short stay quality measures (SSQM)?

3: Do predictions of SSQM also predict COVID-19 healthcare outcomes?

3. RESEARCH METHODOLOGY

The study uses ordinal logistic regression to investigate factors of short-term healthcare quality related to occupancy rates and the nurse staff mix. In this study, the goal is to understand factors of occupancy and nurse staff mix that suggest a likelihood of an increase or decrease in healthcare quality at the SNF. The dependent variable is the government-determined short-stay quality measure (SSQM). See Appendix A for the list of measures the government uses to calculate the short-stay quality measure. The government derives and reports a five-star SSQM rating based on these measures.

The predictor variables are 1) occupancy rate and 2) (CNA + LPN)/RN hours per resident per day. These factors are not part of the calculation of SSQM.

Ordinal logistic regression was used in both industry and state-level:
Regression Equation: SSQM = occupancy rate + (CNA & LPN)/RN hours per resident per day

The dataset used in this study was submitted by SNFs and reported by the Center for Medicare and Medicaid Services (CMS) as of March 1, 2020. The dataset includes approximately 15,000 SNFs. Note, 13,001 SNFs observations, certified by CMS, were used in this study. SNFs that had unreported data for any study variable were removed from the dataset.

The data variables used in this study are the number of certified beds, the average number of residents per day, reported CNA, LPN, and RN hours per resident per day, and SSQM.

Factors of PAC

Occupancy rate is used as a predictor factor in the ordinal logistic regression of SSQM. In this study, the occupancy rate is based on two variables of data collected. 1) The average number of residents per day and 2) the number of certified beds. Certified beds in an SNF are validated by government inspection. The occupancy rate is calculated by taking the average number of residents per day in a facility divided by the number of certified beds at the facility. The average number of residents per day was calculated from the SNF record of residents in bed count on every nursing shift and reported daily.

Occupancy rate = Number of resident days/Number of available bed days.

The occupancy rate considers daily bed use that fluctuates. Therefore, it reflects the SNF's use of physical resources.

A second predictor factor for SSQM is nurse staff mix. The nurse staff mix separates nursing assistants and practitioners from the registered nurse, who administers medical care. CNA, LPN, and RN hours per resident per day were used to predict SSQM. Note, the factor nurse staff mix is a ratio of nurse staff hours. (Importantly, the nurse staff mix is not a single nurse category, and the ratio requires a calculation of three nurse categories). The nurse staff mix used as a predictor of SSQM is shown below as a ratio:

(CNA + LPN) reported hours per resident per day / RN reported hours per resident per day

The CNA and the LPN reported hours per resident per day are combined so that the impact of the RN level of resident care is isolated. Nurse staff mix assesses the levels of different types of nurses by the SNF, which is based on government recommendations.

The short-stay quality metric (SSQM) is the dependent variable and is used as a surrogate measure of SNF short-stay post-acute healthcare quality. The SSQM ranges from poor to excellent quality. This SSQM (see Appendix A) does not include the number of certified beds, the average number of residents per day, and staff hours resident per day (i.e., RN, LPN, or CNA). A focus of the quality metric is on the discharge of residents to the community as well as readmission to a hospital. The use of SNFs for short-stay PAC has only recently received attention from government reporting.

The government reports five levels of short stay quality. The association between the five-star government ratings (i.e., levels) and occupancy rate and nurse staff mix were statistically tested. There was no statistically significant difference between SSQM ratings 2 and 3 and SSQM ratings 4 and 5. Ratings were merged to represent low, medium, and elevated levels of SSQM. They are categorized into three ordinal ratings: 1 low, 2 medium, and 3 high. The SSQM categories were tested for Type 1 error using Tukey HSD. These 3 ratings were used, rather than the five-star government ratings.

4. DATA ANALYSIS

The results provide an understanding of short-term healthcare quality in SNFs. First, the SNF industry-level results are reported. Second, state-level results are presented.

The SNF industry-level logistic regression was run using R, with the occupancy rate* measured in multiples of 10% to detect changes more readily in the rate. The coefficients, t statistic, and p-values are shown in Table 1 below. P-values are computed by comparing the t-value against the standard normal distribution, (n =13,0001, in this study).

a. Increasing occupancy rate by 10% results in a .119 log of odds of achieving a higher SSQM rating (i.e., move from low to medium quality rating). Computing the natural e raised to this exponent provides the odds ratio, a more interpretable measure. Thus, e to the power of

.119 equals 1.126.

That is for every 10% increase in occupancy rate the odds (i.e., likelihood) of improving the SSQM rating is 12.6% (i.e., moving to a higher-level quality rating). Table 1 below summarizes the results of the logistic regression analysis (it does not show the natural e computation).

Table 1. Results of Ordinal Logistic Regression

Occupancy rate* coefficient	t stat.	P value	Ratio (CNA+LPN)/RN coefficient	t stat.	P value
.119	10.37	0	-.109	-25.15	0

b. Increasing the (CNA + LPN) hours per resident per day more than increasing the RN hours per resident per day (e.g., the ratio increases) results in a -.109 log of odds of achieving a lower SSQM rating (i.e., move from medium to low-quality rating). Once again, we raise e to the coefficient in the model, resulting in e to the power of -.109 which equals .897. Since this number is below 1, the result is a lower likelihood and thus we obtain .103 (1 - .897).

That is for every 1 unit increase in the ratio of (CNA + LPN) / RN hours per resident per day the odds of getting a lower SSQM is 10% (i.e., moving to a lower quality rating).

Table 2 presents the summary of results that includes occupancy rate and nurse staff mix. It identifies how the statistical result relates to short stay quality measures.

This result has implications for SNF administrators. The occupancy level is a statistically significant factor that helps to predict short-term PAC quality (i.e., a higher occupancy rate predicts a higher SSQM rating). There is an association between occupancy rate and SSQM.

Table 2. Summary of Results for Predicting SSQM

Factor	Statistically Significant for predicting SSQM	Association with Improving SSQM
Occupancy Rate	Yes	Positive
(CNA +LPN)/RN hours per resident per day	Yes	Negative

The ratio of (CNA + LPN) /RN care is a statistically significant factor that helps predict short-term PAC quality (i.e., a higher ratio, that is more CNAs and LPNs predict a lower SSQM rating). This indicates that there is an association between nurse staff mix and SSQM.

In the next section, an exploratory correlation between factors presented in table 2 above and COVID-19 deaths in SNF is presented. This study considers the care of COVID-19 residents in SNFs as an extraordinary health care service. Although COVID 19 is short-term in nature, it is different than SSQM, which is focused on post-acute care (i.e., strokes and hip replacement surgeries). This preliminary exploration may help SNFs better understand adjustments that may be required in occupancy levels and nurse staff mix for viral outbreaks.

Exploratory correlation between COVID-19 outcomes and SSQM

To explore a relationship between SSQM and COVID-19 outcomes we use the logistic regression equation used at the SNF industry level by state level. We explore this relationship by using CMS public data for total COVID-19 deaths in SNFs per State, as of Jan 31, 2021 (February 3, 2021). This presents almost a year of experience with COVID-19. Although the CMS COVID-19 deaths dataset changes monthly a goal of this study is to better understand the viability of using both the occupancy predictor and nurse staff mix predictor coefficients determined for SSQM. This exploration may have the potential to help guide SNFs in better management of pandemic healthcare conditions. The dataset is used at the state level because of nuances in each state (i.e., demographics and policy characteristics).

At the state level, ordinal logistic regression (see equation above) was run for both occupancy rate and (CNA+LPN)/RN hours per resident per day as predictors for SSQM. Using this ordinal regression, the COVID-19 deaths per state, regression coefficients, and statistically significant p-values for state coefficients are shown in Appendices B and C.

The ordinal logistic regression was run as a

hierarchical multi-level regression model with the occupancy rate coefficients representing the likelihood of moving to a better SSQM rating (see Appendix B). The ratio (CNA + LP) / RN hours per resident per day coefficient, on the other hand, represents a likelihood of moving to a lower SSQM rating (see Appendix C). The predictions of SSQM therefore may also help predict COVID-19 healthcare outcomes.
The statistically significant state coefficients of the predictor variables of SSQM are correlated with the total number of COVID-19 deaths by state. The number of COVID-19 deaths is used as a surrogate for COVID-19 outcome in the SNF. This was done separately for occupancy rate and nurse staff mix ratio. The correlation analysis was computed using the Spearman method. There is a negative correlation between the occupancy coefficient, as a predictor of higher SSQM, and the total number of COVID-19 deaths. Graph 1 below shows this correlation. The Spearman correlation coefficient is -.52 and at the .05 significance level with a power of .80 (r-.52, n = 26, alpha =.05, Power =.80).

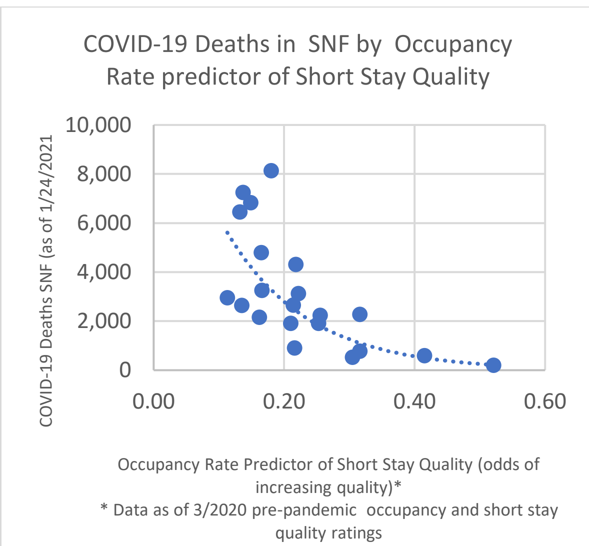


Figure 1. COVID-19 Deaths in SNF by Occupancy Rate predictor of Short Stay Quality

A possible suggestion for this relationship is that at higher occupancy rates there will be less room for COVID-19 admissions. The SNF becomes less condensed with COVID-19 viral load. Further, SNFs that have higher occupancy rates receive more revenue from the government that pays for the short-term care requirement per resident case. This higher revenue may improve the ability of the SNF to manage viral infection outbreaks with the purchase of personal protection equipment, and resources to protect staff and their family and implement new procedures for improved safety.

b. There is a moderate positive correlation between the ratio of (CNA + LP) / RN hours of resident per day coefficient, as a predictor of lower SSQM, and the total number of COVID-19 deaths by state. Graph 2 below shows this correlation. The Spearman correlation coefficient is .41 and at alpha = .10 significance level with power of .70 (r= .41, n = 27, alpha =.10, Power =.7).

Fewer COVID-19 deaths have a relationship with a larger ratio of (CNA + LPN) /RN hours per resident per day. This result is surprising because a larger ratio was a predictor of lower short-term post-acute care quality. This seems counter-intuitive that increasing the nurse staff mix (which means increasing certified nurse assistants and licensed practitioners relative to the number of RNs) predicts lower SSQM yet correlates to increasing COVID-19 outcomes as shown through the surrogate measure of COVID-19 deaths at the SNF. In other words, this exploration may indicate that a better COVID-19 healthcare outcome has an association with a higher nurse staff-mix ratio. Yet, the higher ratio is a predictor of lower SSQM.

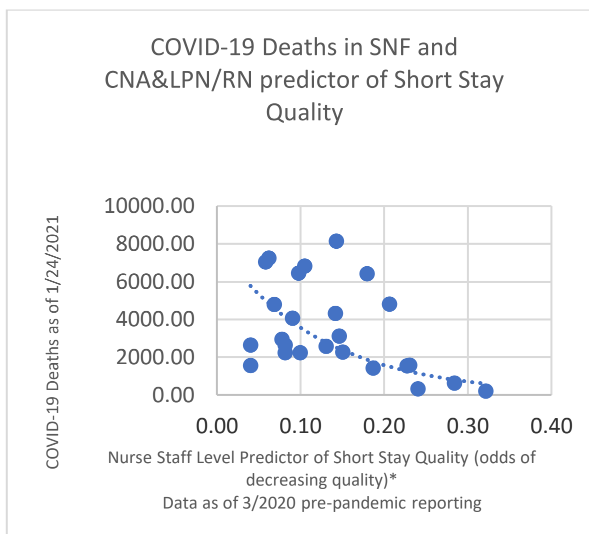


Figure 2. COVID-19 Deaths in SNF and CNA&LPN/RN predictor of Short Stay Quality

It is possible that in the case of infectious disease care when there are more CNAs, each CNA will have fewer patients per shift as well as a higher likelihood that the CNA will stay in one unit in the SNF. If the SNF has fewer CNAs, then each CNA will have to take care of more patients, as well as a float to other units and floors. In this situation, the CNA is more likely to help spread COVID-19 in the SNF.

Table 3. Summary of Results: SNF Factors and Results

SNF Factors	Environment Before COVID-19	Regression Result on SSQM	Environment COVID-19 Era	Correlation of regression result with COVID-19 outcomes*
Occupancy rate	Converting of long-term beds to short-stay PAC to maintain occupancy.	A higher Occupancy rate predicts higher SSQM.	Bed vacancies are filled with COVID-19-positive patients.	Occupancy Rate as a predictor of SSQM has a negative correlation with COVID-19 deaths (Higher occupancy may be correlated with better COVID-19 healthcare outcomes).
(CNA+LPN)/RN hours per resident per day	Changing staff healthcare tasks from ADL to PAC.	A higher ratio predicts lower SSQM.	CNAs and LPNs & RNs help with viral outbreaks.	Nurse Staff Mix Ratio as a predictor of SSQM has a positive correlation with COVID-19 deaths. (Higher nurse staff mix ratio may be correlated with better COVID-19 healthcare outcomes).

*COVID 19 deaths at SNF may be considered a surrogate for COVID-19 healthcare outcome.

5. CONCLUSIONS

The focus of the paper was on short-stay PAC quality. CMS public domain data was used for occupancy, nurse staff mix based on hours per resident and SSQM. COVID-19 deaths were reported in SNFs at the state level. Table 3 presents the summary of the results.

The COVID-19 pandemic emphasizes the importance of occupancy. The pre-pandemic occupancy rate may have limited the number of COVID-19-positive patients that could be admitted to an SNF. Also, focusing on occupancy rates may improve the likelihood of better short-term health care quality because of the increased revenue from more residents. The occupancy rate depends on PAC services, which are incentivized by the government. Therefore, a higher occupancy may also reflect hospital and community reliance on the SNF for short-term rehabilitative care. This necessitates ICT to help assist in the transfer of patients with their electronic medical records.

Further, the government established fixed nurse staff levels for SNFs may hinder short-term healthcare quality. The government assigns quality stars to SNFs that achieve predetermined nurse staff level metrics. Guidance and government incentives may begin to help define a process for adjusting nurse staff mix during extraordinary healthcare crisis conditions, such as COVID-19. The ability of SNFs to provide high-quality short-stay healthcare relies on agile processes and relationships with hospitals and communities. A future paper that explores resource agility in SNFs as a competitive asset may help guide administrators and managers. Nurse staff metrics could become flexible and ad-hoc based on emerging healthcare situations.

A study on the use of ICT to monitor and connect the healthcare organization network (i.e., SNFs, hospitals, and community) may be valuable to improve short stay quality as SNF provides increasingly complex medical procedures. Overall ratings for SNFs should focus on SSQM.

6. REFERENCES

Abrams, H. R., Loomer, L., Gandhi, A., & Grabowski, D. C. (2020). Characteristics of US nursing homes with COVID-19 cases. *Journal of the American Geriatrics Society, 68*(8), 1653-1656.

Backhaus, R., Verbeek, H., van Rossum, E., Capezuti, E., & Hamers, J. P. (2014). Nurse staffing impact on quality of care in nursing homes: A systematic review of longitudinal studies. *Journal of the American Medical Directors Association, 15*(6), 383-393.

Baumann, A., & Kaspar, W. (2021). The shift from inpatient care to outpatient care in Switzerland since 2017: Policy processes and the role of evidence. *Health Policy, 125*(4), 512-519.

Bunis, D. (2015). Need rehab? Beware of Medicare's 'observation status'. AARP. Retrieved June 6, 2022 from https://www.aarp.org/health/medicare-insurance/info-2018/medicare-observation-rule-fd.html

Centers for Disease Control. People who live in a nursing home or long-term care facility. CDC. Retrieved July 12, 2022 from https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-in-nursing-homes.html

Centers for Medicare & Medicaid Services. (2020, March 20). Data that helps you better understand CMS programs. Retrieved June 10, 2022 from https://data.cms.gov

Centers for Medicare & Medicaid Services. (2021, February 3). Data that helps you better understand CMS programs. Retrieved June 10, 2022 from https://data.cms.gov

Chidambaram, P., Neuman, T., & Garfield, R. (2020). Racial and ethnic disparities in COVID-19 cases and deaths in nursing homes. San Francisco: Kaiser Family Foundation.

Desroches, M. L., Ailey, S., Fisher, K., & Stych, J. (2021). Impact of COVID-19: Nursing challenges to meeting the care needs of people with developmental disabilities. *Disability and Health Journal, 14*(1), 101015.

Figueroa, J. F., Wadhera, R. K., Papanicolas, I., Riley, K., Zheng, J., Orav, E. J., & Jha, A. K. (2020). Association of nursing home ratings on health inspections, quality of care, and nurse staffing with COVID-19 cases. *Jama, 324*(11), 1103-1105.

Gibson, M., & Greene J. (2020). State actions and shortages of personal protective equipment and staff in US nursing homes. *Journal of the American Geriatrics Society 68*(12), 2721-2726.

Grabowski, D. C., Huckfeldt, P. J., Sood, N., Escarce, J. J., & Newhouse, J. P. (2012). Medicare post-acute care payment reforms have the potential to improve the efficiency of care, but may need changes to cut costs. Health Affairs, 31(9), 1941-1950.

Hakkarainen, T. W., Arbabi, S., Willis, M. M., Davidson, G. H., & Flum, D. R. (2016). Outcomes of patients discharged to skilled nursing facilities after acute care hospitalizations. Annals of Surgery, 263(2), 280.

Hansen, L. O., Young, R. S., Hinami, K., Leung, A., & Williams, M. V. (2011). Interventions to reduce 30-day rehospitalization: A systematic review. *Annals of Internal Medicine, 155*(8), 520-528.

Harrington, C., Dellefield, M. E., Halifax, E., Fleming, M. L., & Bakerjian, D. (2020a). Appropriate nurse staffing levels for US nursing homes. *Health Services Insights,* 13.

Harrington, C., Ross, L., Chapman, S., Halifax, E., Spurlock, B., & Bakerjian, D. (2020b). Nurse staffing and coronavirus infections in California nursing homes. *Policy, Politics, & Nursing Practice, 21*(3), 174-186.

Heeks, R. (2006). Theorizing ICT4D research. *Information Technologies & International Development, 3*(3), pp. 1-4.

Heeks, R. (2008). ICT4D 2.0: The next phase of applying ICT for international development. *Computer, 41*(6), 26-33.

Heeks, R. (2016). Seeking a new link between ICTs and human development. *Information Technologies & International Development, 12*(1), 51-54.

Jung, H. Y., Trivedi, A. N., Grabowski, D. C., & Mor, V. (2016). Does more therapy in skilled nursing facilities lead to better outcomes in patients with hip fracture? *Physical Therapy, 96*(1), 81-89.

Kauffman, B. (2021). Skilled nursing occupancy increased at slower rate in August 2021. *NIC Notes.* Retrieved June 12, 2022 from https://blog.nic.org/skilled-nursing-occupancy-increased-at-slower-rate-in-august

Kay, M., Santos, J., & Takane, M. (2011). mHealth: New horizons for health through mobile technologies. *World Health Organization, 64*(7), 66-71.

Khimm, S. (2020a). Coronavirus spreads in a New York nursing home forced to take recovering patients. *NBC News*. Retrieved June 1, 2022 from https://www.nbcnews.com/news/us-news/coronavirus-spreads-new-york-nursing-home-forced-take-recovering-patients-n1191811

Khimm, S. (2020b). New York will no longer require nursing homes to take COVID-19 patients from hospitals. *NBC News*. Retrieved June 1, 2022 from https://www.nbcnews.com/news/us-news/new-york-will-no-longer-require-nursing-homes-take-COVID-n1204586

Laes-Kushner, R. (2018). Skilled nursing facilities: Too many beds. *Commonwealth Medicine*. UMass Medical School. Retrieved July 11, 2022 from https://escholarship.umassmed.edu/commed_pubs/193

Li, Y., Thomas, M. A., Stoner, D., & Rana, S. S. J. B. (2020). Citizen-centric capacity development for ICT4D: The case of continuing medical education on a stick. *Information Technology for Development, 26*(3), 458-476.

Mathews, A. W., Fuller, A., & De Avila, J. (2020, May 1). Thinly-staffed nursing homes face challenges in pandemic. *Wall Street Journal*.

Mauro, M., & Giancotti, M. (2021). Italian responses to the COVID-19 emergency: Overthrowing 30 years of health reforms? *Health Policy (Amsterdam, Netherlands), 125*(4), 548–552. Retrieved January 12, 2022 from https://doi.org/10.1016/j.healthpol.2020.12.015

Medicare. Find & compare nursing homes, hospitals & other providers near you. Medicare. Retrieved June 12, 2022 from https://www.medicare.gov/care-compare/?providerType=NursingHome&redirect=true

Novotney, A. (2020). COVID-19 pushed SNF occupancy below 75% in June: NIC. *McKnight's Senior Living*. Retrieved February 1, 2022 https://www.mcknightsseniorliving.com/home/news/business-daily-news/COVID-19-pushed-snf-occupancy-below-75-in-june-nic/

Ochieng, P. C. (2021). Factors associated with COVID-19 cases and deaths in long-term care facilities: Findings from a literature review. *KFF*. Retrieved June 9, 2022 from https://www.kff.org/coronavirus-COVID-19/issue-brief/factors-associated-with-COVID-19-cases-and-deaths-in-long-term-care-facilities-findings-from-a-literature-review/

Orr, N. M., Forman, D. E., De Matteis, G., & Gambassi, G. (2015). Heart failure among older adults in skilled nursing facilities: more of a dilemma than many now realize. *Current Geriatrics Reports, 4*(4), 318-326.

Paulin, E. (2021). COVID Deaths in nursing homes continue alarming climb. *AARP*. Retrieved June 10, 2022 from https://www.aarp.org/caregiving/health/info-2021/nursing-homes-COVID-death-record-december.html

Payne, K. F. B., Wharrad, H., & Watts, K. (2012). Smartphone and medical related App use among medical students and junior doctors in the United Kingdom (UK): A regional survey. *BMC Medical Informatics and Decision Making, 12*(1), 121. doi: 10.1186/1472-6947-12-121.

Public Health. 42 CFR § 483.35 (amended 2022). Retrieved June 12, 2022 from https://www.ecfr.gov/current/title-42/chapter-IV/subchapter-G/part-483/subpart-B/section-483.35

Rahman, M., Grabowski, D. C., Mor, V., & Norton, E. C. (2016). Is a skilled nursing facility's rehospitalization rate a valid quality measure? *Health Services Research, 51*(6), 2158-2175.

Ryskina, K. L., Konetzka, R. T., & Werner, R. M. (2018). Association between 5-star nursing home report card ratings and potentially preventable hospitalizations. *Inquiry: The Journal of Health Care Organization, Provision, and Financing*, *55*.

Saliba, D., Weimer, D. L., Shi, Y., & Mukamel, D. B. (2018). Examination of the new short-stay nursing home quality measures: Rehospitalizations, emergency department visits, and successful returns to the community. *Inquiry: The Journal of Health Care Organization, Provision, and Financing, 55*, doi: 10.1177/0046958018786816

Sen, A. (2013). *A survey of sustainable development: Social and economic dimensions*. Island Press.

Shallcross, L., Burke, D., Abbott, O., Donaldson, A., Hallatt, G., Hayward, A., & Thorne, S. (2021). Factors associated with SARS-CoV-2 infection and outbreaks in long-term care facilities in England: A national cross-sectional survey. *The Lancet Healthy Longevity, 2*(3), e129-e142.

Silva, B. M., Rodrigues, J. J., de la Torre Díez, I., López-Coronado, M., & Saleem, K. (2015). Mobile-health: A review of current state in 2015. *Journal of Biomedical Informatics, 56*, 265-272.

Stone J., & Hoffman, G. (2010, Sept. 21). Medicare hospital readmissions: Issues, policy options and PPACA. Congressional Research Service Report for Congress, 1-37.

The COVID-Tracking Project. (2021, March). About 8% of people who live in US long-term-care facilities have died of COVID-19—nearly 1 in 12. For nursing homes alone, the figure is nearly 1 in 10. *The Atlantic.* https://COVIDtracking.com/nursing-homes-long-term-care-facilities

Weech-Maldonado, R., Lord, J., Pradhan, R., Davlyatov, G., Dayama, N., Gupta, S., & Hearld, L. (2019). High Medicaid nursing homes: Organizational and market factors associated with financial performance. *Inquiry: The Journal of Health Care Organization, Provision, and Financing, 56*, doi: 0046958018825061.

Wickramasinghe, N., Qureshi, S., & Boyer, E. (2020). Introduction to the Minitrack on information technology innovations to overcome health disparities. (Ed.),(Eds.). Proceedings of the 53rd Hawaii International Conference on System Sciences.

Winston, E., Marquis, R., & Pelaez, A. (2019). Improving the communication of medical information in the skilled nursing home industry: The case for a technology partnership. *Management in Healthcare, 4*(1), 21-31.

Winston, E. R., Pelaez, A., & Medlin, B. D. (2021). Will quality measures debunk quality care in the nursing home industry? *International Journal of Public and Private Perspectives on Healthcare, Culture, and the Environment, 5(*1), 18-28.

Yourish, K., Lai, K. R., Ivory, D., & Smith, M. (2020, May 9). One-third of all US coronavirus deaths are nursing home residents or workers. *The New York Times*. Retrieved June 2, 2022 from https://www.nytimes.com/interactive/2020/05/09/us/coronavirus-cases-nursing-homes-us.html

Yurkofsky, M., & Ouslander, J., (2018). Medical care in skilled nursing facilities (SNFs) in the United States. UpToDate. Retrieved July 9, 2022 from https://www.uptodate.com/contents/medical-care-in-skilled-nursing-facilities-snfs-in-the unitedstates

Appendix A. Short Stay Quality Measures

(https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/NursingHomeQualityInits/NHQIQualityMeasures

Percentage of short-stay residents who were re-hospitalized after a nursing home admission. Reporting began 01/01/20
Percentage of short-stay residents who have had an outpatient emergency department visit. Reporting began 01/01/20
Percentage of short-stay residents who got antipsychotic medication for the first time. Reporting began 01/10/20
Percentage of SNF residents with pressure ulcers that are new or worsened (SNF QRP). Reporting began 10/24/18 then changed to skin integrity reporting began FY 2020
Rate of a successful return to home and community from an SNF (SNF QRP). Reporting began 10/24/18
Percentage of short-stay residents who improved in their ability to move around on their own. *Not clear when reporting began
Percentage of short-stay residents who needed and got a flu shot for the current flu season. .*Not clear when reporting began
Percentage of short-stay residents who needed and got a vaccine to prevent pneumonia. .*Not clear when reporting began
Percentage of SNF residents who experience one or more falls with a major injury during their SNF stay (SNF QRP). Reporting began 10/24/18
Percentage of SNF residents whose functional abilities were assessed, and functional goals were included in their treatment plan (SNF QRP). Reporting began 10/24/18
Rate of potentially preventable hospital readmissions 30 days after discharge from an SNF (SNF QRP). Reporting began 10/24/19
Medicare Spending Per Beneficiary (MSPB) for residents in SNFs (SNF QRP). Reporting began 10/24/18

Appendix B. Occupancy Rate

Occupancy Rate			
Total Resident SNF COVID Deaths	State	Occupancy Rate Predictor SSQM	p-value
1907	AR	0.25	0.004
909	AZ	0.22	0.031
7037	CA	-0.23	0
2649	GA	0.21	0.015
4795	IN	0.17	0.003
2641	LA	0.14	0.081
4316	MA	0.22	0.012
1906	MD	0.21	0.045
200	ME	0.52	0.048
3124	MI	0.22	0.003
2153	MN	0.16	0.075
3249	MO	0.17	0.003
1576	MS	0.72	0.001
2952	NC	0.11	0.076
590	ND	0.42	0.088
782	NE	0.32	0.002
533	NM	0.31	0.026
327	NV	-0.34	0.049
6829	NY	0.15	0.057
6452	OH	0.13	0.014
8138	PA	0.18	0.012
1426	SC	-0.16	0.048
2275	TN	0.32	0
7243	TX	0.14	0
2232	VA	0.26	0.019
78	VT	1.14	0.009

Appendix C. (CNA+LPN)/RN Hours Per Resident Per Day

(CNA+LPN)/RN hours per resident per day			
Total Resident SNF COVID Deaths	State	(CNA+LPN)/RN hours per resident per day predictor SSQM	p-value
7037	CA	-0.06	0
1537	CO	-0.26	0.002
2579	CT	-0.14	0.013
341	DE	-0.82	0.01
4063	FL	-0.1	0
2649	GA	-0.09	0.001
6415	IL	-0.2	0
4795	IN	-0.07	0.017
2243	KY	-0.11	0.015
2641	LA	-0.04	0.017
4316	MA	-0.15	0.001
200	ME	-0.39	0.007
3124	MI	-0.16	0
1576	MS	-0.26	0
2952	NC	-0.08	0.013
4803	NJ	-0.23	0
533	NM	0.26	0.03
327	NV	-0.28	0.023
6829	NY	-0.11	0.002
6452	OH	-0.1	0
1564	OK	-0.04	0.045
8138	PA	-0.15	0
1426	SC	-0.21	0
2275	TN	-0.16	0
7243	TX	-0.06	0
2232	VA	-0.09	0.019
626	WV	-0.33	0.002

An Exploration of the Benefits of Certifications and their Relationship to Salaries in IS/IT

Kevin D. Matthews
mattewskd@uncw.edu

Jeff Cummings
cummingsj@uncw.edu

Thomas Janicki
janickit@uncw.edu

Information Systems
University of North Carolina Wilmington
Wilmington NC 28403

Abstract

This research extends previous studies that explore the ever-changing landscape of jobs in the IS/IT (Information Systems / Information Technology) field. To enhance prior research on skills needed and requirements of certifications, we investigated industry certifications in IS/IT fields and having a certification's relationship with salary ranges. Our analysis is from a data set with over 550 responses during the first quarter of 2022. From the survey, the top certifications in rank order were: Microsoft CSE, AWS Cloud Practitioner, AWS Solutions Architect, CompTIA A+ Tech, and CompTIA Network+ in most regions of the US. The survey also investigated salaries by key occupations. On a nationwide basis we found that having a certification versus not having a certification did not relate to any significant salary differences; however, this may depend on region of the US. For those in the Southwest, South Central, and Northeast regions of the US, the data indicated some salary ranges affected by obtaining certifications.

Keywords: job skills, certifications, salaries, economic regions

1. INTRODUCTION

The importance of jobs skills and how they affect curriculums in higher education courses has been a topic of interest for IS (Information Systems) and IT (Information Technology) researchers in the past (e.g., Cummings & Janicki, 2021, 2020; Janicki et al., 2014; Legier et al., 2013). Some have also stressed the importance of "technical certifications" such as A+, Cisco Certified Security Professional (CSSP), Microsoft Certified System Engineer (MSSE). Nevertheless, little research has been done to investigate the preferences of IS/IT professionals regarding certifications and their effects on salaries. This paper summarizes the results of over 550 survey responses throughout the United States to enhance our understanding of the perceived importance of specific certifications and detail any differences in salaries for IS/IT professionals with and without certifications. This research was conducted in the first quarter of 2022.

2. LITERATURE REVIEW

The literature over the past ten years has reported the demand for IS/IT professionals as well as common salaries, skills desired, and certifications. The driving impetus for these studies is to understand the ever-changing nature of IS/IT fields and the associated skill requirements (Aasheim et al., 2012; Prabhakar et al., 2005; Todd et al., 1995). Results of such studies have come from a variety of methods including reviewing IS/IT industry job ads (Burns et al., 2018; Dong & Triche, 2020; Lee & Han, 2008), reviewing the course content at universities (Mills et al., 2016), surveying alumni and recent graduates (Legier et al., 2013; Wilkerson, 2012), employers (Dillon & Kruck, 2008; Zaheer et al., 2021), recruiters (Saia, 2011), or employees in various job roles (Aasheim et al., 2012; Tastle & Russell, 2003).

Recently, Wierschem and Mendez Mediavilla (2018) also dove into exploring how such skills are acquired and the perceived value by employers of different modes of skill acquisition. They reported that over 80% of IT employers do not require certifications for entry-level positions. Interestingly, employer size impacted the desire of an employer for its employees to have certification. The highest desire for certification was for employers with 100 to 499 employers. In their study, of the 15%+ that do require certifications, the most popular certifications were A+, Cisco and Microsoft. Cummings and Janicki (2021) found similar results. From their survey, Microsoft CSE, CompTIA A+, and Cisco CAN were the top three certifications held by IS/IT professionals. They also reported that 89% of survey respondents had completed at least one certification.

Since the IS/IT fields are ever-changing, certification may be a way for professionals to stay abreast of those changes while employed. This may be more evident in different sub-fields of IS/IT. For example, with the increasing concern for cybersecurity knowledge and skills Knapp, Maurer, and Plachkinova (2017) argue for increased undergraduate and graduate courses in cybersecurity and to provide the opportunity for students to achieve certifications in the cybersecurity field.

Given that prior studies have shown the importance on and proliferation of certifications in IS/IT fields, we expect this same importance to be shown by employers; however, the challenge comes from determining how employers value such certifications. Human capital theory (Becker, 1992, 1993) states that the value of an employee is reflected by their wages and thus, certain factors exist that increase human capital and can lead to variances in wages of different employees. Education and training are factors that have been shown to influence salary based on human capital theory. Advice to employers by classic researchers has been to invest in education and training of employees and the returns in productivity will be seen (Schultz, 1961). The way employers have shown this value in the job market is through salaries. In other words, increasing a person's education and training should increase their human capital which will result in an increased valuation and higher salaries. These thoughts were extended by Quan, Dattero, and Galup (2007) to include certifications as a form of both education and training. Their conclusions were that certifications are valuable in general, mixed with education and experience effects, and that their worth varied by job category.

Our research sets out to supplement the findings of previous authors and shed light on the specifics of certifications in the IS/IT field. We do this by undertaking an exploratory look at certifications in IS/IT fields. Our goals are to uncover the current state of certifications and how they affect salaries of IS/IT professionals. We investigate the following certification and salary questions:

- What are the top certifications acquired by IS/IT professionals?
- How do the distributions of these certifications vary across the economic regions of the US?
- Does having a certification or the number of certifications have an effect on salary range?

3. METHODOLOGY

The results presented in this research come from a survey developed to investigate skills acquired and technologies used by IS/IT professionals. The survey was developed through a multi-phase process which utilized IS/IT faculty, an advisory board of industry professionals, and enhancements as suggested from professionals who responded to previous versions of the survey. The current survey was the sixth iteration of the survey and included extensions specifically designed to investigate our certification and salary questions. We have included more details concerning the survey and its development at https://csbweb01.uncw.edu/people/cummingsj/techskills.html.

A major takeaway from the advisory board, faculty, and survey responses was a list of job categories within the IT field. Due to the ever-changing nature of the field, this process can be cumbersome, but is necessary to ensure the most up-to-date categories are used during data collection efforts. In this iteration of the survey, the following job categories were utilized:

- Business/Systems Analyst
- Database Admin
- Networking
- Project Management
- Software Development
- Security
- Analytics

Specifically for the certifications interests of the survey, the advisory board, faculty, and previous survey results were utilized to also develop a list of common certifications that are seen as valuable by professionals. In the current iteration of the survey, the following certifications were explicitly queried:

- AWS Solutions Architect
- AWS Cloud Practitioner
- Certified Risk IS Control
- CISM
- CISSP
- CIS Auditor
- CISCO CAN
- CompTIA A+ Tech
- CompTIA Network+
- Security+
- Microsoft CSE
- PMP
- Professional Cloud Architect
- Scrum Master

After updates to the survey were made, a pilot test was performed to ensure the clarity of questions and responses, an appropriate coverage of the domains in question, and a response time below 10 minutes.

Previous efforts of snowball and convenience sampling limited the total number of response and coverage area of responses across the US. For this reason, a nationwide survey company was utilized to distribute and pay respondents to complete our survey.

4. SUMMARY STATISTICS

The survey was distributed across the US to IT professionals. The survey was completed in the first quarter of 2022 and received 566 total responses. After scrubbing data for issues such as abnormally long response times and incomplete responses, the resulting dataset had a total of 555 usable responses. The average time to complete the survey was 8 minutes, 42 seconds.

Firm Size	**N**	
< 11*	7	1%
11-20*	12	2%
21-100	46	8%
101 - 499	103	19%
500 - 999	160	29%
1000 - 9999	188	34%
10000+	39	7%
Organization Type	**N**	
Corporation	313	56%
Education	26	5%
Government	33	6%
Healthcare*	23	4%
LLC	54	10%
Non or Not for Profit	40	7%
Sole Proprietor or Partnership	66	12%
Region of US	**N**	
Northwest*	17	3%
Southwest	117	21%
North Central*	10	2%
South Central	64	12%
Mid-West	48	9%
Southeast	123	22%
Northeast	176	32%
Gender	**N**	
Female	154	28%
Male	399	72%
Non-binary	1	<1%
Did not specify	1	<1%
Level of Education	**N**	
High school diploma*	15	3%
Associate's degree	29	5%
Bachelor's degree in IT-related field	159	29%
Bachelor's degree in non-IT-related field	55	10%
Master's degree in IT-related field	192	35%
Master's degree in non-IT-related field	71	13%
Ph.D.	34	6%

**NOTE: Less than 25 responses in this category*

Table 1: Responses by Firm Size, Organization Type, Region of the US, Gender, and Level of Education

Participants came from a variety of firm sizes, organization types, and regions of the US (Table 1). The largest percentage (34%) of responses were from professionals in firms with 1000-9999 employees. Corporations gave us 56% of our responses. The largest responses by region came from the Northeast (32%), Southeast (22%), and Southwest (21%). Regardless of region, the top 5 states were New York (22%), California (16%), Florida (7%), Texas (6%), and Pennsylvania (4%). APPENDIX A depicts the states included in each economic region of the US (adapted from Rosen et al., 2014). Please note that due to a lack of representative sample from these areas, 2 responses from Alaska and 2 responses from Hawaii are not included in any of our analysis.

Of the participants, 28% identified as female and 72% identified as male. One respondent identified as non-binary, and one declined to specify a gender. The top 2 responses for highest level of education received were Master's degree in an IT-related field (35%) and Bachelor's degree in an IT-related field (29%), respectively. Additionally, the average number of years in IT was 11 years and average years at current position was 8.8 years.

Lastly, we look at the IS/IT industry-specific statistics from our data set. Table 2 details the breakdown by job category. The top 3 categories in our data set are Software Development (52%), Project Management (16%) and Networking (11%) with Security (4%) and Analytics (2%) having the lowest representations.

Job Category	N	
Business/Systems Analyst	34	6%
DB Admin	52	9%
Networking	60	11%
Project Management	89	16%
Software Development	289	52%
Security*	20	4%
Analytics*	11	2%

*NOTE: Less than 25 responses in this category

Table 2: Responses by Job Category

5. RESULTS AND DISCUSSION

The goal of this research was to investigate the relationships that certifications hold regarding (a) their value among IS/IT professionals and (b) their effect on salary range. This section investigates the insights we extract from an analysis of our data set.

5.1 Certifications

In the survey, participants were asked to identify any certifications that they had from a predefined list (see Section 3 above). Of the 555 responses, 531 (95.7%) indicated at least 1 certification while 24 (4.3%) identified 0 certifications. Of those with at least 1 certification, the average number of certifications is 3.48 (SD = 2.12) with 59.5% (N = 316) having between 1 and 3 certifications (the maximum number indicated was 10). The most frequent number of certifications was 3 (22.4%; N = 119). Figure 1 shows a histogram of the number of certifications for all responses that had at least 1 certification.

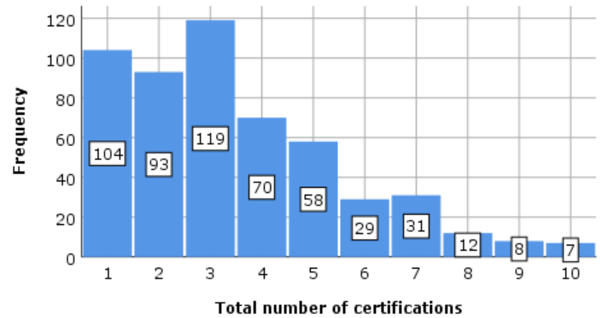


Figure 1: Number of Certifications and Their Frequencies (N = 531)

When asked, 36.2% of those with at least 1 certification stated a certification had been obtained prior to graduation while 63.8% did not have any certifications prior to graduating. This is an interesting finding that suggests many certifications are acquired while working. This can be investigated further by future researchers and intersects with some of the findings of other recent studies (e.g., Wierschem & Mendez Mediavilla, 2018) and with prior calls for employers to invest in certifications as a form of human capital enhancements.

One of the takeaways from our survey results are the value of specific certifications as indicated by the number of respondents that have acquired each certification. We present this information in Table 3. The certifications are ordered by their ranking for the whole of the US with corresponding rows showing the rankings by economic region of the US. The top 5 certifications in each region are shaded. The percentages shown are the percentage of respondents that indicated that they had the certification. Please note that each respondent was able to select multiple certifications.

From this information, we can see that Microsoft CSE is the dominant certification across the US and in most regions. The AWS certifications closely follow. This finding is not surprising given the large number of Software Development professionals that responded to our survey. Notable exceptions are South Central where an AWS certification takes the lead and Mid-West where a CompTIA certification takes the lead. The top 5 certifications (Microsoft CSE, AWS

	Ranking (% with Certification)							
Certification	**All**	**NW***	**SW**	**NC***	**SC**	**MW**	**SE**	**NE**
Microsoft CSE	1 (46%)	1 (53%)	1 (53%)	1 (67%)	4 (34%)	5 (29%)	1 (42%)	1 (80%)
AWS Cloud Practitioner	2 (38%)	3 (41%)	3 (45%)	6 (22%)	1 (44%)	7 (27%)	2 (34%)	2 (57%)
AWS Solutions Architect	3 (37%)	4 (41%)	2 (46%)	7 (22%)	3 (37%)	3 (33%)	5 (31%)	3 (57%)
CompTIA A+ Tech	4 (34%)	6 (35%)	4 (38%)	2 (33%)	6 (31%)	2 (38%)	4 (31%)	5 (54%)
CompTIA Network +	5 (32%)	2 (47%)	6 (30%)	5 (22%)	2 (37%)	1 (40%)	3 (32%)	6 (46%)
Security+	6 (30%)	5 (35%)	5 (31%)	8 (11%)	5 (31%)	6 (29%)	6 (29%)	7 (45%)
Professional Cloud Architect	7 (28%)	7 (29%)	9 (23%)	3 (22%)	8 (19%)	4 (29%)	7 (26%)	4 (55%)
CISCO CNA	8 (25%)	8 (29%)	7 (27%)	9 (11%)	7 (26%)	8 (24%)	8 (24%)	8 (40%)
CISM	9 (20%)	12 (6%)	8 (27%)	- (0%)	9 (18%)	12 (11%)	9 (18%)	9 (37%)
Certified Risk IS Control	10 (18%)	9 (18%)	10 (17%)	11 (11%)	10 (16%)	9 (24%)	10 (17%)	10 (29%)
PMP	11 (11%)	- (0%)	12 (9%)	4 (22%)	11 (13%)	10 (20%)	12 (11%)	13 (15%)
CISSP	12 (11%)	- (0%)	13 (9%)	10 (11%)	13 (11%)	11 (11%)	11 (13%)	11 (18%)
CIS Auditor	13 (10%)	11 (6%)	11 (13%)	- (0%)	12 (13%)	14 (4%)	13 (8%)	12 (16%)
Scrum Master	14 (6%)	10 (6%)	14 (5%)	- (0%)	14 (8%)	13 (4%)	14 (6%)	14 (9%)
Total Responses	*531*	*17*	*112*	*9*	*62*	*45*	*114*	*112*

**NOTE: Less than 25 responses in this category*
- 0% were not ranked

Table 3: Top Certifications Across the US

Cloud Practitioner, AWS Solutions Architect, CompTIA A+ Tech, and CompTIA Network+) remain steady across most regions in slightly varying orders. It is also interesting to note that some regions have 0 certifications indicated for CISSP, PMP, CISM, CIS Auditor, or Scrum Master. It's important to note that these 0% numbers are Northwest and North Central regions which have the smallest representation in our sample (see Table 1).

5.2 Salary Ranges

In our survey, participants were asked to identify the salary range that applied to their current position. The following presents the results of our analysis of this salary range data. In each situation, we utilized chi-square (χ^2) tests for independence to determine if the category of interest was related to the salary range data.

Table 4 shows the average salary ranges by job category. We should note that we did not give a sliding scale for specific numerical salaries, instead we offered 16 categories of $10,000 ranges that went from "less than $10,000" to "$150,000 or more". This data shows that the overall average salary ranged from $90,000 to $109,999 across all categories.

When looking at job categories, the highest salary range of $120,000 to $129,999 is found in Analytics; however, we note that this category also had fewer than 25 responses. The next two highest salary ranges were for Business/Systems Analyst and Software Development each at $100,000 to $119,999. The lowest salary range is found in Networking ($70,000 to $89,999). Our statistical comparison indicates job category is not independent from salary range (χ^2 (90, N = 555) = 127.5, p = 0.006); that is, the two are related in this sample.

When looking at region of the US, the highest salary ranges were found in the Southwest and Northeast ($100,000 to $119,999) and the lowest was in South Central ($80,000 to $99,999). Our statistical comparison indicates region of the US is not independent from salary range (χ^2 (90, N = 555) = 129.3, p = 0.004); that is, the two are related in this sample.

When looking at highest level of education completed, the highest salary range is at the Ph.D. level ($130,000 to $149,999) and the lowest is at the high school diploma level ($50,000 to $69,999). This finding is consistent with the long-held belief that obtaining degrees will result in higher salaries, an idea also supported by human capital theory. Our statistical comparison indicates level of education is not independent from salary range (χ^2 (90, N = 555) = 230.1, p < 0.001); that is, the two are related in this sample.

	Salary Range	
Job Category	**Lower**	**Upper**
Business/Systems Analyst	$100,000	$119,999
DB Admin	$90,000	$109,999
Networking	$70,000	$89,999
Project Management	$90,000	$109,999
Software Development	$100,000	$119,999
Security*	$90,000	$109,999
Analytics*	$120,000	$129,999
	Salary Range	
Region of US	**Lower**	**Upper**
Northwest*	$90,000	$109,999
Southwest	$100,000	$119,999
North Central*	$90,000	$109,999
South Central	$80,000	$99,999
Mid-West	$90,000	$109,999
Southeast	$90,000	$109,999
Northeast	$100,000	$119,999
	Salary Range	
Level of Education	**Lower**	**Upper**
High school diploma*	$50,000	$69,999
Associate’s degree	$70,000	$89,999
Bachelor’s degree in IT-related field	$80,000	$99,999
Bachelor’s degree in non-IT-related field	$90,000	$109,999
Master’s degree in IT-related field	$100,000	$119,999
Master’s degree in non-IT-related field	$100,000	$119,999
Ph.D.	$130,000	$149,999

**NOTE: Less than 25 responses in this category*

Table 4: Average Salary Ranges by Job Category, Region of the US, and Level of Education

5.3 Certification’s Effect on Salary Ranges

The last item of interest is how certifications relate to salary ranges. Since our goals are exploratory in nature, we are specifically looking at these questions:

- Does having one or more certifications affect salary range?
- Does the number of certifications affect salary range?
- Of those who have certifications, does having more increase salary range?

When investigating average salary ranges by number of certifications, we find that the average range remains the same ($90,000 to $109,999) for both the group of participants who indicated 0 certifications and those who have at least 1 certification (i.e., a dichotomous variable). Our statistical comparison indicates having a certification or not is independent from salary range (χ^2 (15, N = 555) = 16.7, p = 0.339); that is, the two have no relationship in this sample. However, we split the sample to investigate this further. Since our previous analysis revealed that there were possible relationships with salary range and the job category, region of the US, and level of education, these were used to split our sample. For consistency with prior salary studies, we also included gender and firm size as categories to split our sample.

Table 5 summarizes the results of our statistical comparisons when we split our sample by job category, region of the US, level of education, gender, and firm size. There was no significant relationship between the salary ranges of those without certifications and those with at least 1 certification by job category, highest level of education received, gender, or firm size. The only statistically significant relationships found are by region of the US; specifically, in the Southwest, South Central, and Northeast regions. This indicates that having 1 or more certifications may affect salary range in these 3 regions. The results from an independent samples t-test reveal a statistically significant difference in salary ranges between those having a certification or not in the Southwest (t(115) = 2.3; p = .02; mean difference = 4.32, C.I. = [.61, 8.04]), but not in the South Central (t(62)= -.08; p = .93) or Northeast (t(174)= .96; p = .34) regions.

Further dividing respondents with certifications reveals that those with more than 3 certifications (the average number of certifications in our sample) increased their salary range up to $100,000 to $119,999. When we introduce three categories of (1) having 0 certifications, (2) having 1-3 certifications, and (3) having more than 3 certifications, we do find a significant relationship with salary range in our overall sample (χ^2 (30, N = 555) = 48.5; p = .02); that is, there is a relationship between certification categorized in these 3 levels and salary range. The results from a one-way ANOVA reveal a statistically significant difference in salary ranges in these 3 levels of having a certification ($F(2, 552) = 8.4$; $p < .001$). Using a Tukey HSD post-hoc test for comparisons, we find that there is a statistically significant difference between people with more than 3 certifications and people with between 1 and 3 certifications ($p < .001$) such that the difference in salary range is 1.48 on our 16-point salary range scale (95% C.I. = [.62, 2.33]).

Job Category	χ^2	df, N	p
Business/Systems Analyst	15.5	10, 34	.11
DB Admin+	-	-	-
Networking	12.9	13, 60	.46
Project Management	10.7	14, 89	.71
Software Development	22.2	15, 289	.10
Security*+	-	-	-
Analytics*	11.0	6, 11	.09
Region of US	χ^2	**df, N**	**P**
Northwest*	-	-	-
Southwest	44.5	15, 117	**.000**
North Central*	10.0	7, 10	.19
South Central	25.5	13, 64	**.02**
Mid-West	22.4	15, 48	.10
Southeast	15.6	14, 123	.34
Northeast	25.6	15, 176	**.04**
Level of Education	χ^2	**df, N**	**p**
High school diploma*	10.8	8, 15	.21
Associate’s degree	10.8	9, 29	.29
Bachelor’s degree in IT-related field	18.5	14, 159	.19
Bachelor’s degree in non-IT-related field	20.7	12, 55	.06
Master’s degree in IT-related field	10.2	14, 192	.74
Master’s degree in non-IT-related field	9.2	13, 71	.75
Ph.D.	4.8	5, 34	.44
Gender	χ**2**	**df, N**	**p**
Female	13.4	14, 154	.50
Male	20.4	15, 399	.16
Non-binary	-	-	-
Did not specify	-	-	-
Firm Size	χ^2	**df, N**	**p**
< 11*	7.0	3, 7	.07
11-20*	5.5	8, 12	.71
21-100	9.8	11, 46	.55
101 - 499	18.8	15, 103	.22
500 - 999	7.3	15, 160	.95
1000 - 9999	13.9	15, 188	.53
10000+	5.8	8, 39	.67

**NOTE: Less than 25 responses in this category*
+all respondents in this category indicated they had 1 or more certification

Table 5: Having a Certification’s Effect on Salary Range (split samples)

There is no significant difference between people with no certification and people with between 1 and 3 ($p = .99$) or between people with more than 3 ($p = .28$).
+

Job Category	χ^2	df, N	p
Business/Systems Analyst	25.3	20, 34	.19
DB Admin	14.2	12, 52	.29
Networking	20.5	26, 60	.77
Project Management	26.4	28, 89	.55
Software Development	47.1	30, 289	**.02**
Security*	11.9	10, 20	.29
Analytics*	18.6	12, 11	.01
Region of US	χ^2	**df, N**	**p**
Northwest*	9.3	6, 17	.16
Southwest	64.4	30, 117	**.000**
North Central*	15.7	14, 10	.33
South Central	48.8	26, 64	**.004**
Mid-West	37.9	30, 48	.15
Southeast	30.3	28, 123	.35
Northeast	40.6	30, 176	.09
Level of Education	χ^2	**df, N**	**p**
High school diploma*	17.0	16, 15	.39
Associate’s degree	20.9	18, 29	.28
Bachelor’s degree in IT-related field	41.8	28, 159	**.05**
Bachelor’s degree in non-IT-related field	28.2	24, 55	.25
Master’s degree in IT-related field	22.6	28, 192	.75
Master’s degree in non-IT-related field	26.6	26, 71	.43
Ph.D.	11.6	10, 34	.31
Gender	χ**2**	**df, N**	**p**
Female	26.4	28, 154	.55
Male	47.7	30, 399	**.02**
Non-binary	-	-	-
Did not specify	-	-	-
Firm Size	χ^2	**df, N**	**p**
< 11*	9.8	6, 7	.13
11-20*	15.0	16, 12	.53
21-100	21.4	22, 46	.50
101 - 499	42.4	30, 103	.07
500 - 999	21.8	30, 160	.86
1000 - 9999	27.5	30, 188	.60
10000+	15.4	16, 39	.50

**NOTE: Less than 25 responses in this category*
+all respondents in this category indicated they had 1 or more certification

Table 6: Having a Certification’s Effect on Salary Range (3 categories, split samples)

Table 6 summarizes the results of our statistical comparisons using this new 3-category variable when we split our sample by job category, region of the US, level of education, gender, and firm size. While the relationship is significant

across the whole sample with this new categorization, there are still no significant relationships when split by firm size. Significance does, however, exist when looking at job category, region of the US, and gender here.
Only one job category indicated a relationship between certification and salary range – Software Development. The relationship between the 3-level categorization of certifications and salary in the Software Development job category was analyzed with a one-way ANOVA that revealed a statistically significant difference in salary ranges ($F(2, 286) = 4.8$; $p = .009$). Using a Tukey HSD post-hoc test for comparisons, we find that there is a statistically significant difference between people in Software Development with more than 3 certifications and people with between 1 and 3 certifications ($p = .006$) such that the difference in salary range is 1.53 on our 16-point salary range scale (95% C.I. = [.36, 2.70]). There is no significant difference in Software Development between people with no certification and people with between 1 and 3 ($p = .78$) or between people with more than 3 ($p = .93$).

Similar to the dichotomous variable, when split by region of the US, the Southwest and South Central regions show significance, but the Northeast does not. The results from a one-way ANOVA reveal a statistically significant difference in salary ranges in these 3 levels of having a certification in the Southwest region ($F(2,114)= 6.3$; $p = .003$), but not the South Central region ($F(2,61) = .2$; $p = .85$). Using a Tukey HSD post-hoc test for comparisons, we find that there is a statistically significant difference between people in the Southwest with more than 3 certifications and people with no certifications ($p = .01$) such that the difference in salary range is 5.36 on our 16-point salary range scale (95% C.I. = [-9.80, -.92]). Also, there is a statistically significant difference between people in the Southwest with more than 3 certifications and people with between 1 and 3 ($p = .03$) such that the difference in salary range is 2.00 on our 16-point salary range scale (95% C.I. = [.20, 3.80]. There is no significant difference between people in the Southwest with no certification and those having between 1 and 3 ($p = .17$).

Our sample also shows that having a Bachelor's degree in an IT-related field reveals a relationship with certifications and salary range using the 3-level categorization for number of certifications. The results from a one-way ANOVA reveal a statistically significant difference in salary ranges in these 3 levels of having a certification for people having a Bachelor's degree in an IT-related field ($F(2,156) = 5.2$; $p = .007$). Using a Tukey HSD post-hoc test for comparisons, we find that there is a statistically significant difference between people having a Bachelor's degree in an IT-related field with more than 3 certifications and people with between 1 and 3 certifications ($p = .02$) such that the difference in salary range is 1.64 on our 16-point salary range scale (95% C.I. = [.17, 3.11]). There is no significant difference among people having a Bachelor's degree in an IT-related field with no certification and people with between 1 and 3 ($p = .07$) or between people with more than 3 ($p = .55$).

Furthermore, our sample has a relationship between salary ranges and this 3-level categorization of number of certifications in the male sample. The results from a one-way ANOVA reveal a statistically significant difference in salary ranges in these 3 levels of having a certification for males ($F(2,396) = 5.3$; $p = .005$). Using a Tukey HSD post-hoc test for comparisons, we find that there is a statistically significant difference between males with more than 3 certifications and males with between 1 and 3 certifications ($p =.004$) such that the difference in salary range is 1.38 on our 16-point salary range scale (95% C.I. = [.37, 2.39]). There is no significant difference in males with no certification and those with between 1 and 3 ($p = .56$) or between those with more than 3 ($p = .98$).

These results suggest that, in general, having certifications may not begin to increase salary until 3 or more are acquired. Evidence from the Software Development job category, a Bachelor's in an IT-related field, and males support this notion. Even so, it may be that the Southwest region of the US does not follow this because there were significant effects between having no certifications and both having between 1 and 3 and having 3 or more.

6. FUTURE RESEARCH AND REMARKS

Using our analysis and results, we hope future researchers investigate the following questions:

- Why is the Southwest region of the US a consistent factor that changes how certification relates to salary?
- How does obtaining certifications during employment modify salary ranges?
- What other theoretical models help explain certification's effects on salary

and what other factors might these theories introduce?

Every research has limitations and ours is not exempt. While we acknowledge some limitations in our study, we hope future researchers can account for these and extend our work. First, we note that we chose to use salary ranges instead of a continuous salary measurement. This did modify how we reported some results and the statistical methods we employed.

Additionally, we acknowledge the fact that our sample did have some frequencies in certain categories that were low (e.g., the Security and Analytics job categories; the Northwest and North Central regions of the US; the High School Diploma level of education; and firm sizes below 20 employees). While there is an argument that some of these categories may play an important role in understanding certifications in the IS/IT field, we believe our sample served its purpose to give insight into the field, certifications, and salaries.

The same may be true for categories that had high frequencies of response. For example, the Software Development job category had the highest representation in our sample (59.5%). It's possible that this fact influenced some of the findings at the full sample level. We provided analysis by job category, region of the US, and other factors to attempt to view the results from multiple angles. These attempts allow us to better understand our data set and what valuable information we can extract from it.

7. CONCLUSIONS

Based on the findings, certifications continue to be prominent in the IT/IS field with over 95% of participants holding at least one which has also increased from 2020 at 89% and 62% in 2018 (Cummings & Janicki, 2021). While the number of participants not holding a certification was relatively low, there was no difference in salary based on whether they were holding a certification or not. While certifications do provide some validation to a person's skills level, the results bring into question the value of a certification for those looking for a higher salary. However, when evaluating education level, those with a Bachelor's degree in an IT-related field may find it beneficial to get certifications as a means of competitive advantage in the job market. This education level was found to have a significant difference in salary when analyzing people with at least 1 certification. Overall, certifications will continue to dominate the field and should be viewed as a means of expanding one's skills and not necessarily benefiting one's salary.

8. REFERENCES

Aasheim, C., Shropshire, J., Li, L., & Kadlec, C. (2012). Knowledge and Skill Requirements for Entry-Level IT Workers: A Longitudinal Study. *Journal of Information Systems Education*, *23*(2), 193–204.

Becker, G. S. (1992). Human capital and the economy. *Proceedings of the American Philosophical Society*, *136*, 85–92.

Becker, G. S. (1993). *Human capital: A theoretical and empirical analysis, with special reference to education* (3rd ed). The University of Chicago Press.

Burns, T. J., Gao, Y., Sherman, C., & Klein, S. (2018). Do the Knowledge and Skills Required By Employers of Recent Graduates of Undergraduate Information Systems Programs Match the Current ACM/AIS Information Systems Curriculum Guidelines? *Information Systems Education Journal*, *16*(5), 56–65.

Cummings, J., & Janicki, T. (2021). Survey of Technology and Skills in Demand: 2020 Update. *Journal of Information Systems Education*, *32*(2), 150–159.

Cummings, J., & Janicki, T. N. (2020). What Skills Do Students Need? A Multi-Year Study of IT/IS Knowledge and Skills in Demand by Employers. *Journal of Information Systems Education*, *31*(3), 208–217.

Dillon, T. W., & Kruck, S. E. (2008). Identifying Employer Needs from Accounting Information Systems Programs. *Journal of Information Systems Education*, *19*(4), 403–410.

Dong, T., & Triche, J. (2020). A Longitudinal Analysis of Job Skills for Entry-Level Data Analysts. Journal of Information Systems Education, 31(4), 312–326.

Janicki, T. N., Cummings, J., & Kline, D. (2014). Information Technology Job Skill Needs and Implications for Information Technology Course Content. *Information Systems Education Journal*, *12*(6), 59–70.

Knapp, K. J., Maurer, C., & Plachkinova, M. (2017). Maintaining a cybersecurity curriculum: Professional certifications as valuable guidance. *Journal of Information Systems Education*, *28*(2), 101–114.

Lee, C. K., & Han, H.-J. (2008). Analysis of Skills Requirement for Entry-Level Programmer/Analysts in Fortune 500 Corporations. *Journal of Information Systems Education*, *19*(1), 17–28.

Legier, J., Woodward, B., & Martin, N. (2013). Reassessing the Skills Required of Graduates of an Information Systems Program: An Updated Analysis. *Information Systems Education Journal*, *11*(3), 79–89.

Mills, R. J., Chudoba, K. M., & Olsen, D. H. (2016). IS Programs Responding to Industry Demands for Data Scientists: A Comparison Between 2011-2016. *Journal of Information Systems Education*, *27*(2), 131–140.

Prabhakar, B., Litecky, C. R., & Arnett, K. (2005). IT skills in a tough job market. *Communications of the ACM*, *48*(10), 91–94.

Quan, J. J., Dattero, R., & Galup, S. D. (2007). Information Technology Wages and the Value of Certifications: A Human Capital Perspective. *Communications of the Association for Information Systems*, *19*(6), 81–114.

Rosen, L. D., Janicki, T., & Gebauer, J. (2014). Operations Management Salary Report. *Production and Inventory Management Journal*, *49*(1), 90–101.

Saia, R. (2011). *9 Hot Skills for 2012*. https://www.computerworld.com/article/2550102/9-hot-skills-for-2012.html

Schultz, T. W. (1961). Investment in human capital. *American Economic Review*, *51*(1), 1–17.

Tastle, W. J., & Russell, J. (2003). Analysis and Design: Assessing Actual and Desired Course Content. *Journal of Information Systems Education*, *14*(1), 77–90.

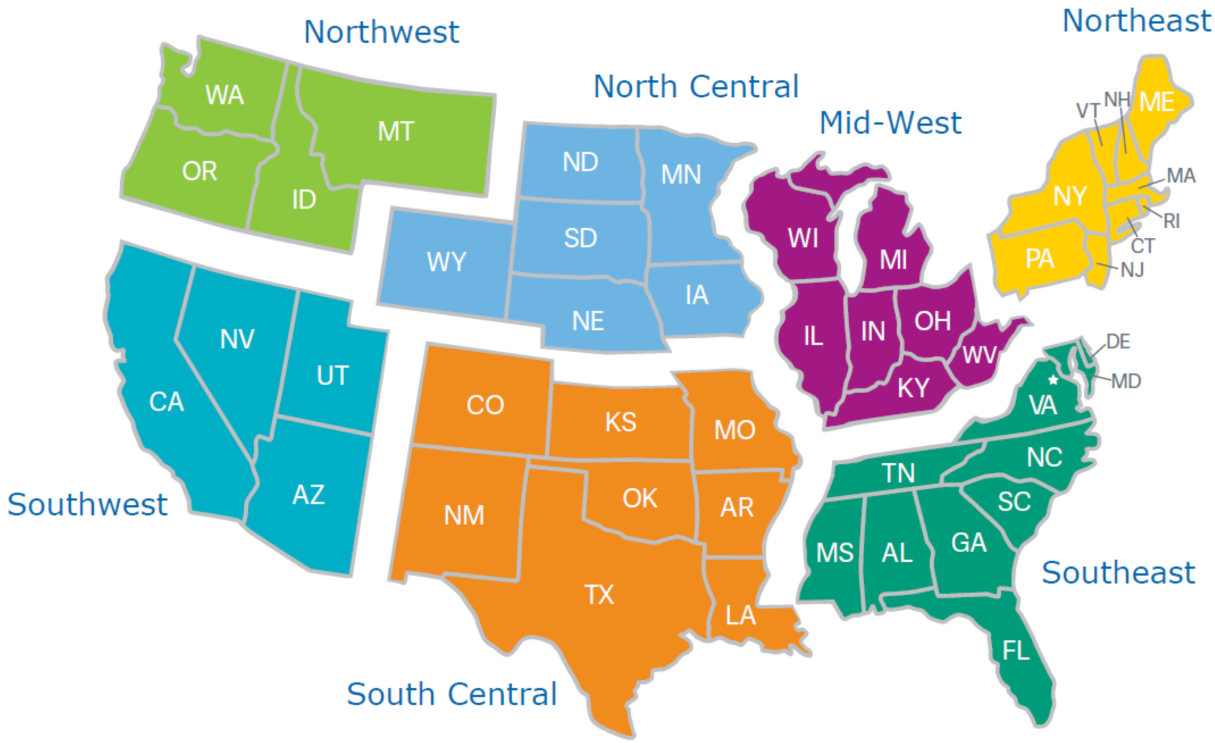
Todd, P. A., McKeen, J. D., & Gallupe, R. B. (1995). The Evolution of IS Job Skills: A Content Analysis of IS Job Advertisements from 1970 to 1990. *MIS Quarterly*, *19*(1), 1–27.

Wierschem, D., & Mendez Mediavilla, F. (2018). Entry level technology positions: No degree required. *Journal of Information Systems Education*, *29*(4), 253–268.

Wilkerson, J. W. (2012). An Alumni Assessment of MIS Related Job Skill Importance and Skill Gaps. *Journal of Information Systems Education*, *23*(1), 85–97.

Zaheer, M. I., Ajayi, S. O., Zulu, S. L., Oyegoke, A., & Kazemi, H. (2021). Understanding the key competencies of market-ready building surveying graduates from employers' perspectives. *Journal of Engineering, Design and Technology*, *19*(1), 291–314.

APPENDIX A
States Included in the Economic Regions of the US



* Adapted from (Rosen et al., 2014)

Northwest (NW)
WA, OR, ID, MT

Southwest (SW)
CA, NV, UT, AZ

North Central (NC)
WY, ND, SD, NE, MN, IA

South Central (SC)
CO, NM, KS, OK, TX, MO, AR, LA

Mid-West (MW)
WI, IL, MI, IN, OH, KY, WV

Northeast (NE)
PA, NY, NJ, CT, RI, MA, ME, NH, VT

Southeast (SE)
VA, DE, MD, DC, TN, NC, SC, GA, AL, MS, FL

A Serverless Real-Time Data Streaming Architecture for Synchronous Online Math Competition

Yu-Che Liu
liuyuche@city.edu

Sam Chung
chungsam@cityu.edu

School of Technology & Computing (STC)
City University of Seattle (CityU)
Seattle, WA

Abstract

During the synchronous online math competition, a high volume of data is continuously generated and collected from the user at every event. Those valuable data can be processed and analyzed to support many decisions in a second such as user state, cheating detection, online help desk, etc. In addition, enormous demands from different roles accessing those data, such as data analysts, data scientists, and executives, have increased recently. This paper presents a new architecture for synchronous real-time data streaming for online testing. Also, this paper identifies three challenges that need to be addressed: First, most online testing organizations rely on open-source frameworks for big data processing and streaming - Hadoop and Kafka. Second, we add the serverless architecture for synchronous real-time data streaming to an open-source learning management system, Moodle, to meet the synchronized online test requirement. Third, we discuss the benefit of the architecture in terms of high availability, cost, and technologies.

Keywords: Serverless, Real-time Data Streaming Architecture, Synchronous Online Math Competition, Open-Source Big Data Processing Framework, Open-Source Data Streaming Framework

1. INTRODUCTION

During the recent COVID-19 situation, education and industry were forced to go online. As a result, education and industry need to perform online tests for students and provide an online business platform for their employees (Wahid et al., 2015). However, building a synchronized system takes many resources for the company to hire experts and build a highly scalable infrastructure. Moreover, small to mid-size companies have a challenge with real-time data processing.

Processing real-time data is a complex job for a large-scale distribution system. Therefore, the market demand for real-time streaming has been highly increased (Liu et al., 2014). More companies are challenged to provide real-time processing for extensive data analysis due to the expensive cost, maintenance difficulty, and low performance.

The synchronous online system brings a huge benefit to companies. However, building and maintaining the system can cost some resources and energy (Kontaxakis et al., 2021). Small-size companies do not have enough resources to build and maintain this huge real-time data processing system.

The main difficulty of building a synchronized system is maintaining the users' state on the server. To maintain the users' state, we need to maintain the communication protocol between producers and consumers to stream users' data to the destination (Xu et al., 2021). The options to build the communication protocols are HTTP/s and WebSocket. However, building the communication protocols can be a massive project and sometimes take several months.

Another difficulty in building a synchronized system is processing the data in real-time. Processing the data in near real-time requires several computing resources running in parallel to support instant data processing. Traditionally, companies must purchase many physical servers and set up batch processing software with stream processing capability, including Hadoop, Apache Spark, Flink, etc. However, the system was built on the on-premise server using Kafka and Flink services (Thein, 2014).

In this paper, we challenge how to maintain users' data in the synchronous online exam with high availability (Chaffai et al., 2017). Keeping users' data synchronous can bring a lot of traffic pressure on the server and require more engineering work to provide availability and reliability. Thus, we propose an AWS Serverless cloud computing solution, including AWS Kinesis, to help companies build fast, easy-to-maintain, high-performance real-time data processing platforms.

This paper proposes a serverless real-time data streaming architecture and applies the architecture to synchronous online math competitions:

1. Serverless architecture reduces development workload and maintenance costs.
2. Serverless architecture optimizes and maintains the user state data.
3. Serverless architecture provides a solution to process data in near real-time.

First, most online testing organizations rely on open-source frameworks for big data processing and data streaming - Hadoop and Kafka. Second, we add the serverless architecture for synchronous real-time data streaming to an open-source learning management system, Moodle (Chaffai et al., 2017).

There is an open-source streaming software to help us build our synchronized application without reinventing the wheels, including Kafka, ActiveMQ, and RabbitMQ (Dobbelaere et al., 2017).

An open-source streaming software can build real-time data processing architecture within a limited time. However, cloud computing can help companies manage and reduce building system workloads and budgets. For instance, Amazon Kinesis is an AWS fully managed service that engineers can focus on ingesting and processing streaming data without managing any infrastructure. Amazon Kinesis utilizes massive computing resources from AWS to create a real-time data processing platform (Gannon et al., 2017).

AWS serverless framework provides a solution to let developers build serverless applications on AWS. The solution can integrate multiple AWS services to help the developers focus on building an application without dealing with all the infrastructure. Compared with the other real-time data processing frameworks, including Hadoop, MapReduce, Spark, and Kafka, using AWS serverless framework can provide lower upfront operational cost and low maintenance (Nguyen et al., 2021)

2. BACKGROUND

This section will provide some background knowledge of serverless and how serverless applies to data processing.

Nowadays, mobile users are growing exponentially, which generates a lot of data from user events. The more data stored in the database, the more possibilities to find the relationship between data. Relationships can be used to build a machine learning model to solve some issues. Storing and processing massive datasets requires tons of computing resources and storage space. To support its business needs, it costs many resources to purchase equipment and human resources upfront. Not every company is capable of handling this amount of cost. The emergence of Cloud computing opens another possibility to achieve the same features but less cost and management (Nicoletti, 2016).

The emergence of cloud computing allows developers to focus on building business logic instead of setting and maintaining the server configuration (Kemal et al., 2009). The Cloud provider takes care of physical hardware settings and maintenance tasks. As a result, the developers can spin up several servers based on business requirements within several clicks on the Cloud provider dashboard without going through all the processes of setting up a physical machine.

On top of the cloud computing concept, Serverless architecture provides optimized solutions for automatically scaling up and building client and server-side applications using the same software development kit.The serverless tool brings the companies two primary benefits: cost and scalability. When the product goes live, serverless tools charge based on the volume of the service usage. Therefore, it brings substantial cost savings compared to building on-premises infrastructure (Gannon et al., 2017).

The demand for big data analysis has become higher than before. Users generate different types of data, such as logs, application status, and transactions, and send them to data infrastructures, which transfer raw data into information stored in in-memory or persistency databases. Then, data consumers retrieve the data from the database, continuously processing it with complex use cases such as machine learning models, reports, and feature engineering.

In education, there are several learning management tools such as Google Classroom, Blackboard Learning Management System, Moodle, Etc. However, the current learning management systems lack real-time data processing support, which cannot provide synchronization of online exams. The synchronization feature can provide a better user experience closer to the on-site exam.

3. RELATED WORK

Processing real-time data streaming has become popular these days. Data plays a vital role in streaming. In the past, people tried to figure out how to manage the data by executing SQL commands in the database. When high-speed internet comes, it provides higher bandwidth for more data to transmit.

Many research approaches analyze and process the data in real time. For example, Akanbi (2020) monitors data generate from different IoT devices and proposes a four-stage architecture to collect, stream, analytics, and store the environmental sensor data. However, their approach does not give real-time feedback based on the result of the analysis. Furthermore, all four stages are not highly cohesive; some use cloud computing, and others use open-source software on on-premise servers. This combination greatly allows the developer to choose technologies, but it increases the response time between layers.

Javed et al. (2017) states that throughput plays an important role when the team process real-time data. High throughput can have lower latency.

Table 1 compares two research studies that conduct real-time data processing platforms. Both researchers collect the data in real time. However, they do not synchronize data with front-end implementations and use cloud computing completely.

Criteria	Adeyinka, 2018	Javed et al., 2017
Architecture	IOT, Kafka	Flink and Kafka
Collect real-time data	Yes	Yes
Synchronize data with front-end	No	No
Do real-time analysis	No	Yes
Performance	N/A	160 records/ms
Serverless	No	No
Cost	N/A	High

Table 1: A Summary of Related Work

4. APPROACH

The new architecture contains four parts: collecting users' data from the online test system, sending real-time data into the data pipeline, and processing and analyzing data into information stored in the database for the online testing system to synchronize data.

Figure 1 shows the serverless architecture to process, analyze, and store real-time data. To build this architecture, we build a customized online test plugin on Moodle learning management system to access users' test data and send those data back to AWS Kinesis firehose processing in real-time. After AWS Kinesis firehose receives data, data will enter a customized data pipeline to perform Amazon Kinesis Data Analytics SQL statement and query the result based on customizing data schema sending to AWS in-memory database, Amazon ElastiCache. In addition, the Moodle plugin can retrieve users' previous status and record when users disconnect from the synchronizing exam.

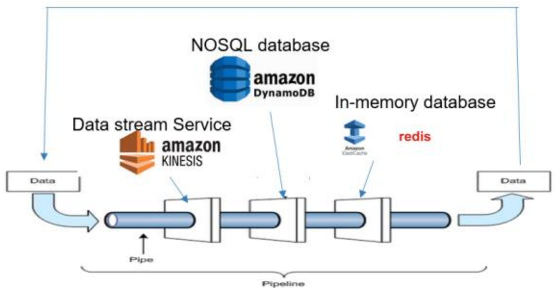


Figure 1: Serverless Architecture

Moodle, a learning management system, is an open-source learning tool, which provides e-learning, and flipped classroom features in various industries. We built a customized Moodle learning management system plugin that provides an online synchronization exam.

5. DATA COLLECTION

Based on the architecture in this paper, the data collection starts from the front-end interface, which is Moodle learning management system. When the user answers each question, the program will send the answer data to the server immediately instead of waiting for the user to answer all the questions.

A group of data is being transferred into AWS Kinesis. The different parameter is used for AWS Kinesis to compare the traditional architecture of data steaming.

Figure 2 shows how to collect synchronized user data created from the user behavior. Data is structured in a key-value pair format, allowing AWS Kinesis to retrieve and analyze the data from applications quickly.

Figure 3 shows when AWS Kinesis receives the data and will run SQL commands to analyze the user's status in real time.

Data latency is critical for the synchronized exam; this paper compares the data latency performance under different architectures. The start timestamp is recorded when the data is being sent out on the client-side, and the stopped timestamp is generated on the server-side to calculate the data latency between the front and backend.

```php
//sample data
$name = "online-test-student-data";
$content = '{"user":"spencer", "status":"000001"}';

try {
    $result = $firehoseClient->putRecord([
        'DeliveryStreamName' => $name,
        'Record' => [
            'Data' => $content,
        ],
    ]);
    var_dump($result);
} catch (AwsException $e) {
    echo $e->getMessage();
    echo "\n";
}
```

Figure 2: Synchronized User's Status in PHP

```sql
CREATE OR REPLACE STREAM "STUDENT_STATUS" ("student_id" INTEGER, "status" INTEGER);

CREATE OR REPLACE PUMP "STREAM_PUMP" as
INSERT INTO "STUDENT_STATUS"
SELECT STREAM "student_id", sum("status")
FROM SOURCE_SQL_STREAM_001
GROUP BY "student_id", STEP ("SOURCE_SQL_STREAM_001".ROWTIME BY INTERVAL '1' MINUTE)
ORDER BY STEP("SOURCE_SQL_STREAM_001".ROWTIME BY INTERVAL '1' MINUTE), avg("status");
```

Figure 3: Real Time Analysis in SQL

In order to lower data latency, our approach is to collect the information from the interface like Figure 4 and restructure the data into three columns, including status, student_id, and timestamp, as shown in Table 5.

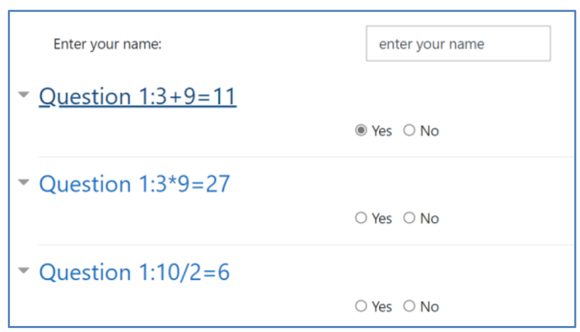


Figure 4: User's Data on Moodle System

Status (VARCHAR)	student_id (INTEGER)	TIMESTAMP
s1120	3	1.6312066665
s1100	701	1.6312909355
s0000	977	1.6312718455

Figure 5: Data Schema

6. Data Analysis

This section presents the data collection from AWS Kinesis and Moodle system. The data is generated from AWS Kinesis API with a message buffer size of 1 MB, and each Buffer interval is 60 seconds. The idea for deciding on the buffer size is based on the size of all the questions in the online exam. We evaluate the amount of data generated from the student; each tested student has around 50 questions, and each question takes 2 bytes, which equals 0.000002MB. Therefore, one student can send 0.0001MB of data to the AWS Kinesis server for an online test. Based on the architecture purpose of this paper, Figure 6 shows that our approach can handle around 500,000 students taking the online exam simultaneously. The X-axis is the time period, and the Y-axis is the number of records our architecture can support.

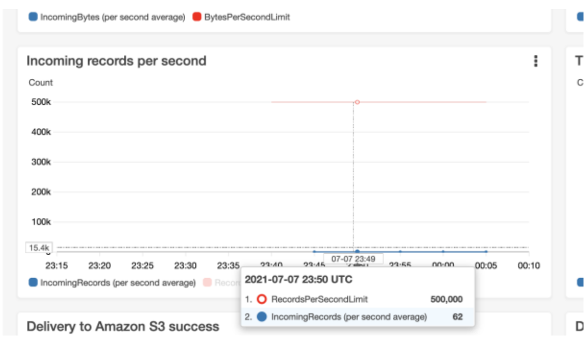


Figure 6: Data digest in a second.

7. FINDINGS

Synchronized online test system
Our solution provides a synchronization online test system on Moodle education system. Teacher can implement a synchronized test system without struggling with technical issues and reduce the time to develop software. Students can get a similar testing experience using their devices anywhere. They do not need to worry about connection and internet speed issues, and all the answers will be streamed in real-time to the server and synchronized with the online test system.

Cloud adoption
The traditional method to build the real-time data processing platform is made on the on-premises server. However, creating an on-premises server is a massive cost for the companies in the early stage, such as purchasing a physical server and hiring a real-time management expert. By using cloud solutions, it can have a more flexible budget and reduce the software development time.

High availability
Cloud computing provides high availability to guarantee real-time data platform availability. In addition, all the data will be synchronized in three availability zones and one cold storage in the AWS S3 bucket to increase data durability.

Our approach utilizes the AWS cloud computing environment to build a real-time data processing platform. Compared to previous work, our approach provides real-time analysis, high performance, and lower cost using serverless architecture, as shown in Table 2 below.

Criteria	Adeyinka, 2018	Javed et al., 2017	Our Approach
Architecture	IOT, Kafka	Flink, Kafka	Moodle, Amazon Kinesis, Amazon DynamoDB
Collect real-time data	Yes	Yes	Yes
Synchronize data with front-end	No	No	Yes
Do real-time analysis	No	Yes	Yes
Performance	N/A	160 records/ ms	500 records/ ms
Serverless	No	No	Yes
Cost	N/A	High	Low

Table 2: A Comparison of Previous Work and Our Approach

8. CONCLUSIONS

This paper presents the new architecture for real-time data streaming based on the existing education system and identifies three challenges that need to be addressed in our architecture. First, most organizations or education areas rely on open-source learning systems. Second, on top of the synchronized learning management system, we add customized solutions for improvement and customization to give real-time ability to meet the synchronized online test requirement.

Building and maintaining real-time features for online testing systems consume more resources for small to medium-sized companies. Adopting cloud computing saved many engineer costs and reduced the software development time. AWS cloud computing can achieve a high volume of real-time data processing and various data storage options. Connecting Moodle system with real-time data processing provides the online testing event with high availability.

The system can be improved to make the system become more cost-effectively practical and better performance. First, real-time data streaming processing brings massive data from the application. It will be a considerable cost for the companies or organizations to store the data in their system for a long time. AWS provides different layers of storage options. Data can be moved between cold and hot storage. Second, a real-time data platform will be deployed to serve multiple users from different locations. Different regions may want to track their data or query based on their time zone. The biggest challenge is to maintain the data in proper order and location.

9. REFERENCES

Akanbi, A. (2020, November). ESTemd: A Distributed Processing Framework for Environmental Monitoring based on Apache Kafka Streaming Engine. In 2020 the 4th International Conference on Big Data Research (ICBDR'20) (pp. 18-25).

Chaffai, A., Hassouni, L., & Anoun, H. (2017, March). E-learning real time analysis using large scale infrastructure. In Proceedings of the 2nd international conference on big data, cloud and applications (pp. 1-6).

Dobbelaere, P., & Esmaili, K. S. (2017, June). Kafka versus RabbitMQ: A comparative study of two industry reference publish/subscribe implementations: Industry Paper. In Proceedings of the 11th ACM international conference on distributed and event-based systems (pp. 227-238).

Gannon, D., Barga, R., & Sundaresan, N. (2017). Cloud-native applications. IEEE Cloud Computing, 4(5), 16-21.

Javed, M. H., Lu, X., & Panda, D. K. (2017, December). Characterization of big data stream processing pipeline: a case study using Flink and Kafka. In Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (pp. 1-10).

Delic, K. A., & Walker, M. A. (2008). Emergence of the academic computing clouds. Ubiquity, 2008 (August), 1-1Kontaxakis, A., Deligiannakis, A., Arndt, H., Burkard, S., Kettner, C. P., Pelikan, E., & Noack, K. (2021, June). Real-time processing of geo-distributed financial data. In Proceedings of the 15th ACM International Conference on Distributed and Event-based Systems (pp. 190-191).

Liu, X., Iftikhar, N., & Xie, X. (2014, July). Survey of real-time processing systems for big data. In Proceedings of the 18th International Database Engineering & Applications Symposium (pp. 356-361).

Nguyen, K., & Chung, S. (2021). Low Maintenance, Low Cost, Highly Secure, and Highly Manageable Serverless Solutions for Software Reverse Engineering. In Proceedings of the Conference on Information Systems Applied Research ISSN (Vol. 2167, p. 1508).

Nicoletti, B. (2016, March). Cloud computing and procurement. In Proceedings of the International Conference on Internet of things and Cloud Computing (pp. 1-7).

Thein, K. M. M. (2014). Apache kafka: Next generation distributed messaging system. International Journal of Scientific Engineering and Technology Research, 3(47), 9478-9483.

Wahid, A., Sengoku, Y., & Mambo, M. (2015, January). Toward constructing a secure online examination system. In Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication (pp. 1-8).

Xu, J., Yin, J., Zhu, H., & Xiao, L. (2021, May). Modeling and verifying producer-consumer communication in Kafka using CSP. In 7th Conference on the Engineering of Computer Based Systems (pp. 1-10).

Editor's Note:

This paper was selected for inclusion in the journal as an CONISAR 2022 Distinguished Paper. The acceptance rate is typically 7% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2022.

Identification of Stressed Wolf Populations Based on Hormone Levels Using Support Vector Machines (SVM)

John C. Stewart
stewartj@rmu.edu

G. Alan Davis
davis@rmu.edu

Department of Computer Information Systems
Robert Morris University
Moon Township, PA 15108

Abstract

In North America, wolf populations have been relentlessly hunted and persecuted since Europeans landed in the new world. In recent years, in an effort to restore the balance of flora and fauna in ecosystems, wolves have been reintroduced in some areas. In other areas, wolf populations are still hunted, based upon the premise of "managing them." Prior studies have suggested that physiological indicators, specifically elevated hormone levels, are symptomatic of higher stress levels in individual wolf subjects in heavily-hunted populations. This stress has far-reaching implications for reproduction, social structure and pack dynamics. The current study supports prior studies that used statistics to show elevated stress levels in hunted wolf populations and classification of individual wolf subjects as belonging to hunting-based stressed populations, based on physiological data, and by using machine learning.

Keywords: Machine Learning, data mining, Support Vector Machines, physiological indicators, Classification, Data Analytics

1. INTRODUCTION/LITERATURE REVIEW

Human caused mortality of predator populations (from hunting) has resulted in collateral negative effects on the impacted population (Coltman, 2003; Darimont 2009). The longstanding traditional objective of hunting has been the search for, and killing of, the strongest and fittest of the group or population, thereby reducing the reproduction of the healthiest members of the population. Evidence has shown that in the hunting of rams for trophies, there is a predominance of those of heavier weight and larger horn size (Festa-Bianchet et al., 2004, Coltman et al., 2003). These rams are of stronger breeding stock. Also, it was determined that they were of relative younger age, thereby having a negative impact on the reproduction rate of the populations (Coltman et al., 2002).

The intricate and complex social structure of wolf populations leads to an extreme vulnerability to additional increases in mortality and a derailment of pack behavior dynamics as a result of human intervention (Haber, 1996). Although wolf populations can recover from less severe and limited decreases in population, chronic pressures can negatively impact behavior, the foundation of social structure, and genetic capability. This combination of factors may reduce the possibility of group and pack recovery to sustainable and thriving levels. (Rausch, 1967; Haber, 1996; Jezdrzejewski et al. 2005; Sidorovich et al., 2007; Rutledge et al., 2010, 2012).

Heavily-hunted wolf populations produce more female offspring (Sidorovich et al., 2007). Researchers have determined that genetic diversity in wolf populations is impacted by intense hunting (Jezdrzejewski et al., 2005). The social dynamics of neighboring wolf populations can be affected by the harvesting of wolves outside these protected areas (Rutledge et al., 2010). The mortality of wolf pups can increase, leading to reductions in the rate of population growth (Rausch, 1967).

The impact of hunting on population numbers can be easily determined, but accompanying physiological effects have not been measured or documented. It has been concluded that levels of hormones, like cortisol, are an indicator of increased stress in hunted individuals (Bateson & Bradshaw 2007). Also, stress can negatively affect the social behavior of a species population (Gobish et al., 2008).

Testosterone is a required component to male reproduction capability, but also has an effect on behavior. If an imbalance exists in the social structure, it is possible that testosterone may increase (Oliveria, 2004).

A number of studies have concluded that elevated levels of the hormones cortisol, testosterone, and progesterone in pregnant females are a reflection of the reproductive activity in the population (Foley et al., 2001). A relationship between female testosterone levels and the social structure of the populations has also been determined (Albert et al., 1991 & Bryan, et al., 2013).

More recently, researchers have looked at the changes in hormone levels as an indication of physiological effects of hunting by humans. One study has evaluated hormone levels in wolf populations to determine how human-caused mortality may impact group behavior, reproduction, and social dynamics. (Bryan et al., 2015). The researchers in this study concluded that elevated hormone levels can be found in subjects found in heavily-hunted wolf populations. Another study determined that, using machine learning, individual wolf subjects could be classified as belonging to either heavy or low hunting populations based on the level of these hormones (Stewart et al., 2021).

2. RESEARCH METHODOLOGY

In the current research we are attempting to further previous work to determine if additional machine learning methods might improve the accuracy of classification of wolf subjects based on hormone levels. As in the prior work, we are classifying individual wolf subjects as belonging to a heavily-stressed population due to hunting, or as a member of a population with lower hunting pressure. The criteria for determining stress is the measurement of hormones and reproductive steroids in the wolf's fur. Similarly, we will evaluate the hormone levels of two separate wolf populations in Northern Canada studied previously (Bryan et al., 2015; Stewart et al., 2021).

The differences in these two wolf populations in this dataset is marked by the level of hunting. Wolves in the tundra-taiga area were heavily hunted using snowmobiles and firearms. Taiga is characterized by dense conifers, like spruce and pine. Conversely, tundra regions lack any tree cover. Wolves in the second area, boreal forests, had a lower level of mortality and were killed predominately by trapping. Boreal forests consist of deciduous and conifer trees, and experience wide-ranging temperatures over the course of the year (Musiani, M. & Paquet, P.C., 2004).

Bryan et al., (2015) concluded that elevated levels of stress and subsequent increased reproduction activity in the heavily hunted tundra-taiga wolves, were linked to high rates of hormone production (testosterone, progesterone, and cortisol). The researchers in the 2015 study compared the tundra-taiga wolves to wolves in areas of lower hunting pressure (i.e., the boreal forest), concluding statistical difference exists. (Bryan et al., 2015). In a prior work, we determined that classification of individual wolves in the same dataset was possible using machine learning, specifically k-nearest neighbor (Stewart et al., 2021). In this current work, our research questions are: 1) Can we determine the population that individual wolves belong to, based on the level of stress-related hormones using Support Vector Machines (SVM)?, and 2) Is this an improved method over k-Nearest neighbor in our prior work?

Sampling Method

The samples (n=148) were collected in a prior study in Nunavut, Northwest Territories and Alberta, Canada (Musiani et al., 2007). The samples (See Appendix, populations 1 and 2) consisted of wolf hair samples collected during the winter months. The process of extracting the hormones from the wolf hair, including quality control methodologies, is outlined in the Bryan et al., study (2015).

Bryan et al., (2015) used statistical analysis to differentiate the tundra-taiga wolves from the boreal forest wolves. The researchers used ANOVA and Welch's t-tests to compare the two wolf populations, concluding that wolves from the more heavily hunted populations had increased levels of reproduction and stress related hormones. They suggested that these physiological characteristics are in response to environmental factors, including human-induced mortality (Bryan et al., 2015).

The researchers proposed that confounding factors, specifically, ecological and genetic-based differences that could explain the gap in the level of hormones in the two populations. They concluded that the higher levels of cortisol in the tundra-taiga wolves could be attributed to long-term shortages in the food supply in summer, when wolves must travel farther to catch up with migrating caribou. Additionally, the massing of tundra-taiga wolf populations near caribou in summer causes interactions among wolves of different packs. (which could also explain the elevated levels of testosterone). The boreal wolves, conversely, have more traditional territories and stability, leading to fewer intergroup interactions (Walton, et al., 2001, Musiani et al., 2007).

To test the influence of these confounding factors, the researchers used a control group of wolves (n=30) from a heavily-hunted population in a boreal forest region (See Appendix, population 3). The hormone samples in the control group showed higher levels of cortisol than in the boreal forest populations. The wolves in the control group also had similar levels of cortisol as wolves in the heavily hunted northern tundra-taiga region. Therefore, the researchers concluded that higher cortisol levels are the result of increased mortality rates, possibly coupled with some habitat related factors (Bryan et al., 2015).

The current research seeks to determine and support the prior research on classification of the wolf subjects in this dataset. Prior work has answered the question as to whether human-exploited wolf populations are more heavily impacted physiologically (Bryan et al., 2015). And, an individual wolf can be classified to a population of hunted wolves by the level of specific hormones in the fur (Stewart et al., 2021). The current work expands upon this previous study and attempts to classify wolf subjects using Support Vector Machines (SVM).

Research Question

The research question to be evaluated in this study is as follows:

Can Individual wolves be classified into one of two populations, those belonging to a heavily-exploited population, or a member of a less-hunted group, based on hormone levels using SVM? SVM are a machine-learning methodology for classifying data points using an n-dimensional feature space (Cortes & Vapnik, 1995). SVM can be used to support the results obtained by two previous studies (Bryan et al., 2015; Stewart et al., 2021). That is, wolves can accurately be classified into one of two groups: 1) those with high levels of hunting-induced stress, and 2) those with less stress using SVM.

The objective of the current study is to determine whether the physiological consequences of hunting (as determined by levels of stress and reproductive hormones in hair, an indicator of elevated endocrine activity), can be used to classify wolves as belonging to a highly-stressed group or a less-stressed group.

To test our research question, we used data previously analyzed by Bryan et al., (2015) and SVM as the classification methodology to determine wolf membership in heavily stressed versus low stressed populations, based on hormone levels. The 2015 dataset included subject wolves from two separate areas and environments. The dataset contained 45 wolves from a lightly-hunted group in a northern boreal forest, and 103 wolves from a heavily-hunted Tundra-taiga forest area.

All samples were taken as part of a prior study (Musiani et al., 2007). The samples consisted of hair from the wolf subjects. Cortisol, testosterone, and progesterone (females) levels were measured in each hair sample. The data, listing population, gender, hair color, and levels of the three hormones can be found in the Appendix. In this work, unlike the 2021 study, we included fur color and gender variables in the analysis.

Support Vector Machines

SVM was used to compare cortisol and testosterone levels in the two different populations, and to determine the accuracy in classifying each subject into one of the populations, based on its hormone levels. Bryan et al., (2015) determined that higher levels of cortisol and testosterone were found in the tundra-taiga wolves and concluded that this higher level may be an indicator of social instability.

Due to the lower numbers of northern boreal forest wolves, stratified sampling was used. In addition, the data were partitioned into 70% for training and 30% for testing. A SVM linear model was developed using the Caret package in R, and 10-fold cross validation was used to improve accuracy and reduce over-fitting.

SVM algorithms are designed to find the optimum hyperplane in an N-dimensional space (N = the number of features) that separates, and most distinctly classifies, the data points (Figure 1).

There are a number of possible hyperplanes that can separate the two classes of the target variable. The goal is to determine the plane that has the maximum margin, (i.e., the maximum distance between data points of the different classes. Maximizing the margin distance elicits a higher level of confidence in the classification of new data points (Boser et al., 1992). The goal of classification using SVM then, is to determine the maximum separation between the possible outcomes of the classification of a target variable (i.e., two possible outcomes in this case).

SVM is a powerful classification method with the potential for high accuracy compared to other classification methods. It has wide application in classification, including cancer genomes (Huang et al., 2018), chemoinformatics and drug research (Rodríguez-Pérez, 2022), and even the classification of running technique between experienced and novice runners (Carter et al., 2022).

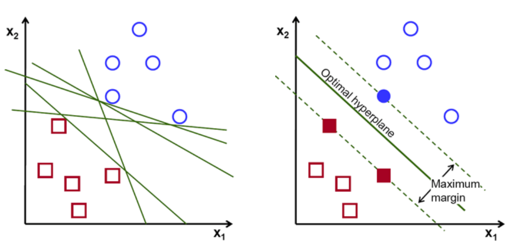


Figure 1: SVM determines the optimal hyperplane separating data points

3. RESULTS AND DISCUSSION

The results of our model yielded an accuracy of nearly 83% (as shown in Table 1). Additionally, the accuracy value is substantially above the No Information Rate, which indicates that the model is superior to simply choosing the dominant class from the target variable. Additionally, the Kappa (a measure of agreement between actual and predicted values, taking chance into account) is in a range that indicates good agreement. Considering all of the components in the aggregate, this suggests a markedly significant model with the capability of predicting the group an individual wolf subject belongs to, based upon the levels of cortisol and testosterone in the fur.

Measurement	Value
Accuracy	0.8276
Sensitivity	0.6667
Specificity	0.9412
Kappa	0.631
No Information Rate	0.5862

Table 1: Results of Classification of Wolf Subjects based on Cortisol and Testosterone Levels Using SVM

4. CONCLUSIONS

Past research on this topic has proposed that elevated levels of the hormones cortisol, testosterone, and progesterone in taiga-tundra wolves are explained by the synergistic effects of hunting pressures, the habitat, or sampling (Bryan et al., 2015). In the Bryan et al., study, the researchers compared cortisol levels in the taiga-tundra wolves to those of a control group of 30 wolf subjects (i.e., Little Smokey wolves) in a heavily-hunted boreal forest area in an effort to explain the differences in habitat and ecosystem characteristics. The results of this study showed statistically higher cortisol levels in both the Little Smokey and taiga-tundra wolves, compared to the northern boreal forest wolves.

Another study used the k-NN classification algorithm to show that individual wolves can be classified as belonging to heavily hunting-pressured groups based on cortisol and testosterone levels (Stewart et al., 2021). This classification was also shown to be at a highly-accurate level. That study also concluded that classification of female wolves (using the k-NN classifier) is possible with a favorable accuracy, based on the females' levels of progesterone.

Our results in the current work support the findings of Bryan et al., (2015) that showed statistically-significant differences in hormone levels between taiga-tundra and boreal forest wolf populations (i.e., heavily hunted vs. lightly trapped populations). Our results also support the results of Stewart et al., (2021) showing classification with high accuracy is possible in classifying hunting-stressed wolf subjects based on hormone levels.

Our findings support our suggestion that individual wolves can be classified as belonging to a heavily exploited population based on hormone levels using SVM. The similar results to the prior study using k-NN supports the use of machine learning models to classify the wolf subjects in this small dataset despite the relative imbalance in the target variable.

Prior studies have concluded that the potential implications of heavy human-caused mortality in wolves are substantive chronic stress, and diminished reproduction and breeding. The negative effects on breeding, compared with non-distressed populations are unclear. However, predictable genetic outcomes, as in the case with in-breeding, lack of diversity, increased disease, along with an elevated danger of population extinction are potential long-term impacts of heavy hunting (Leonard et al., 2005).

There are several implications revealed by the differences in hormone levels as determined by Bryan et al., 2015; Stewart et al., 2021), and this current work study. First, reproduction rates may be altered (and the social structure, along with the reproduction rates) when there is no longer a dominant pair (or pack hierarchy), and additional pack members are also breeding. The stability of the social group, characterized by a single litter per pack each year, is unbalanced (Haber, 1996). High levels of testosterone aid in any challenges an individual wolf may have within the social structure, where strength and dominance of the situation are necessary (Wingfield et al., 2001).

With a link between stress levels in wolf populations and human-based hunting, aside from the impact on wolf populations, the effects on entire ecosystems could be impacted. Wolves are recognized as a keystone species in their natural habitat (Boyce, 2018; Ripple & Beschta, 2012). Therefore, their absence or minimization can have far reaching impacts on entire ecosystems.

Limitations of Study

It should be noted that the sample size in this study was relatively small, particularly with the northern boreal forest wolves (i.e., n = 45). However, the research was unfortunately limited by the amount of available data. Additional machine-learning techniques and models could be employed in future studies to improve the results using methods to address unbalanced small datasets. These additional techniques might be used to determine whether we can improve the classification accuracy of wolf subjects, based on hormone levels as indicators of human-caused stress.

5. REFERENCES

Albert, D., Jonik, R. & Walsh, M. (1992). Hormone-dependent aggression in male and female rats: experiential, hormonal, and neural foundations. Neuroscience & Biobehavioral Reviews, 16(2), 177–192.

Bateson, P. & Bradshaw, E.L. (2007). Physiological effects of hunting red deer (Cervus elaphus). Proceedings of the Royal Society of London. Series B: Biological Sciences, 264(1389), 1707–1714.

Boser, B.E., Guyon, I.M. & Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. 5th Annual ACM Workshop on COLT (ed. Haussler, D.) 144–152. ACM Press, Pittsburgh, PA, USA.

Boyce, Mark S. (2018). Wolves for Yellowstone: dynamics in time and space. Journal of Mammalogy, 99(5), 1021–1031.

Bryan, Heather M. et al. (2015). Heavily hunted wolves have higher stress and reproductive steroids than wolves with lower hunting pressure. Functional Ecology, 29(3), 327-356. https://doi.org/10.1111/1365-2435.12354

Bryan, H.M., Darimont, C.T., Paquet, P.C., Wynne-Edwards, K.E. & Smits, J.E. (2013). Stress and reproductive hormones in grizzly bears reflect nutritional benefits and social consequences of a salmon foraging niche. PLoS One, 8(11), e80537.

Joshua Autton Carter, Adrian Rodriguez Rivadulla & Ezio Preatoni (2022). A support vector machine algorithm can successfully classify running ability when trained with wearable sensor data from anatomical locations typical of consumer technology, Sports Biomechanics, 1-18. DOI: 10.1080/14763141.2022.2027509

Coltman, D.W., O'Donoghue, P., Jorgenson, J.T., Hogg, J.T., Strobeck, C. & Festa-Bianchet, M. (2003). Undesirable evolutionary consequences of trophy hunting. Nature, 426(6967), 655–658.

Coltman, D. W., Festa-Bianchet, M., Jorgenson, J. T. & Strobeck, C. (2002). Age-dependent sexual selection in bighorn rams. Proceedings of the Royal Society of London. Series B: Biological Sciences, 269(1487), 165-172.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273–297.

Darimont, C.T., Carlson, S.M., Kinnison, M.T., Paquet, P.C., Reimchen, T.E. & Wilmers, C.C. (2009). Human predators outpace other agents of trait change in the wild. Proceedings of the National Academy of Sciences, 106(3), 952–954.

Festa-Bianchet, M., Coltman, D. W., Turelli, L. & Jorgenson, J. T., (2004). Relative allocation to horn and body growth in bighorn rams varies with resource availability. Behavioral Ecology, 15(2), 305-312.

Foley, C.A.H., Papageorge, S. & Wasser, S.K. (2001). Noninvasive stress and reproductive measures of social and ecological pressures in free ranging African elephants. Conservation Biology, 15(4), 1134-1142.

Gobush, K.S., Mutayoba, B.M. & Wasser, S.K. (2008). Long term impacts of poaching on relatedness, stress physiology, and reproductive output of adult female African elephants. Conservation Biology, 22(6), 1590–1599.

Haber, G.C. (1996). Biological, conservation, and ethical implications of exploiting and controlling wolves. Conservation Biology, 10(4), 1068–1081.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics & Proteomics, 15(1), 41–51. https://doi.org/10.21873/cgp.20063

Jezdrzejewski, W., Branicki, W., Veit, C., Me€ Augorac, I., Pilot, M.g., Bunevich, A. et al. (2005). Genetic diversity and relatedness within packs in an intensely hunted population of wolves Canis lupus. Acta Theriologica, 50(1), 3–22.

Leonard, J.A., Vilà, C. & Wayne, R.K. (2005). Legacy lost: genetic variability and population size of extirpated US grey wolves (Canis lupus). Molecular Ecology, 14(1), 9–17.

Musiani, M. & Paquet, P.C. (2004) The practices of wolf persecution, protection, and restoration in Canada and the United States. BioScience, 54(1), 50–60.

Musiani, M., Leonard, J.A., Cluff, H.D., Gates, C.C., Mariani, S., Paquet, P.C. et al. (2007) Differentiation of tundra/taiga and boreal coniferous forest wolves: genetics, coat colour and association with migratory caribou. Molecular Ecology, 16, 4149–4170.

Oliveira, R.F. (2004). Social modulation of androgens in vertebrates: mechanisms and function. Advances in the Study of Behavior, 34, 165–239.

Packard, J.M. & Mech, L.D. (1980). Population regulation in wolves. Biosocial Mechanisms of Population Regulation (eds M.N. Cohen, R.S. Malpass & H.G. Klein), pp. 135–150. Yale University Press, New Haven, Connecticut, USA.

Packard, J.M. & Mech, L.D. (1983). Population regulation in wolves. Symposium on Natural Regulation of Wildlife Populations. Proceedings 14 (eds F.L. Bunnell, D.S. Eastman & J.M. Peek), 151–174. University of Idaho, Moscow, Idaho, USA.

Packard, J.M., Mech, L.D. & Seal, U.S. (1983). Social influences on reproduction in wolves. Wolves of Canada and Alaska (ed. L.N. Carbyn), pp. 78–85. Canadian Wildlife Service Report Series, 45, Ottawa, Ontario, Canada.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rausch, R.A. (1967). Some aspects of the population ecology of wolves, Alaska. American Zoologist, 7(2), 253–265.

Ripple, William & Beschta, Robert. (2012). Trophic cascades in Yellowstone: The first 15 years after wolf reintroduction. Biological Conservation, 145(1), 205–213.

Rodríguez-Pérez, R., Bajorath,J. (2022). Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery. J Comput Aided Mol Des 36, 355–362. https://doi.org/10.1007/s10822-022-00442-9

Russell, E., Koren, G., Rieder, M. & Van Uum, S.H. (2014). The detection of cortisol in human sweat: implications for measurement of cortisol in hair. Therapeutic Drug Monitoring, 36(1), 30–34.

Rutledge, L.Y., Patterson, B.R., Mills, K.J., Loveless, K.M., Murray, D.L. & White, B.N. (2010). Protection from harvesting restores the natural social structure of eastern wolf packs. Biological Conservation, 143(2), 332–339.

Rutledge, L.Y., White, B.N., Row, J.R. & Patterson, B.R. (2012). Intense harvesting of eastern wolves facilitated hybridization with coyotes. Ecology and Evolution, 2(1), 19–33.

Sapolsky, R.M. (1993). The physiology of dominance in stable versus unstable social hierarchies. Primate Social Conflict (eds W.A. Mason & S.P. Mendoza), 171– 204. State University of New York Press, Albany, New York, USA.

Sidorovich, V., Stolyarov, V., Vorobei, N., Ivanova, N. & Jedrzejewska, B. (2007) Litter size, sex ratio, and age structure of gray wolves, Canis lupus, in relation to population fluctuations in northern Belarus. Canadian Journal of Zoology, 85(2), 295–300.

Stewart, J.C., Davis, G. A. & Iogche, D. (2021). Using physiological effects and k-nearest neighbor to identify hunting-stressed wolf populations. Proceedings of the Conference on Information Systems Applied Research, 14 (5529), Washington DC, USA.

Walton, L.R., Cluff, H.D., Paquet, P.C. & Ramsay, M.A. (2001). Movement patterns of barren-ground wolves in the central Canadian Arctic. Journal of Mammalogy, 82(3), 867–876.

Wingfield, J.C., Lynn, S. & Soma, K.K. (2001). Avoiding the 'costs' of testosterone: ecological bases of hormone-behavior interactions. Brain Behavior and Evolution, 57(5), 239– 251.

APPENDIX A
Wolf Hair Data Collected during Musiani, et al. Study (2007)

Individual	Sex	Population	Colour	Cpgmg	Tpgmg	Ppgmg
1	M	2	W	15.86	5.32	NA
2	F	1	D	20.02	3.71	14.37622
3	F	2	W	9.95	5.3	21.65902
4	F	1	D	25.22	3.71	13.42507
5	M	2	D	21.13	5.34	NA
6	M	2	W	12.48	4.6	NA
7	M	1	W	26.78	4.58	NA
8	M	1	D	15.41	9.27	NA
9	F	1	D	33.87	4.81	19.9127
10	F	2	W	17.29	5.07	34.59806
11	F	1	W	9.43	4.47	25.88548
12	F	1	W	8.84	3.75	15.86882
13	F	1	D	34	4.76	33.08362
14	F	1	D	14.3	6.06	24.82876
15	M	1	D	12.16	5.75	NA
16	M	1	D	22.43	6.15	NA
17	F	2	W	26.26	4.93	25.00037
18	M	2	W	15.8	5.24	NA
19	M	1	W	7.93	4.14	NA
20	M	1	D	4.75	3.34	NA
21	M	2	W	9.17	4.02	NA
22	M	2	W	21.52	4.91	NA
23	M	1	W	10.79	3.91	NA
24	F	2	W	22.69	6.47	21.50033
25	F	2	W	22.17	4.28	31.8274
26	F	2	W	15.34	5.53	34.0765
27	F	1	W	20.48	5.06	20.21606
28	F	1	W	16.19	4.79	18.29115
29	F	1	W	24.05	3.7	21.29735
30	M	2	W	16.45	6.09	NA
31	F	2	W	21.91	4.19	36.40797
32	F	2	W	32.24	6.94	40.92793
33	F	2	W	23.99	5.97	45.9136
34	F	2	W	27.82	7.76	47.2674
35	F	2	W	19.83	6.55	40.93838
36	F	2	W	12.16	4.34	26.65583
37	F	2	W	19.05	6.34	23.90413
38	F	2	D	13.91	4.72	26.36326
39	F	2	D	17.16	9.25	34.64966
40	F	1	W	30.16	6.8	19.61885
41	F	2	W	24.38	5.49	28.12497
42	F	2	D	10.14	3.81	NA

43	M	2	W	18.4	4.98	NA
44	M	2	W	15.21	7.17	NA
45	M	2	W	24.64	15.13	NA
46	M	2	W	22.49	14.45	NA
47	M	2	W	17.42	5.36	NA
48	M	2	W	29.51	9.12	NA
49	M	2	W	27.3	10.75	NA
50	M	2	W	14.04	7.19	NA
51	M	2	W	11.77	5.17	NA
52	M	2	W	23.6	6.97	NA
53	M	2	W	18.14	5.7	NA
54	M	2	W	11.25	4.4	NA
55	F	1	W	14.82	10.81	NA
56	F	2	W	26.39	6.47	24.46521
57	M	2	W	15.15	4.52	NA
58	M	2	W	14.04	6.01	NA
59	M	2	W	21.39	7.36	NA
60	F	2	W	20.02	5.19	31.40929
61	M	2	W	24.64	14.08	NA
62	M	2	W	13.46	4.09	NA
63	M	2	W	18.79	9.74	NA
64	F	2	W	11.77	4.95	21.01472
65	F	2	W	19.96	7.62	28.06955
66	F	2	W	12.68	3.82	27.90797
67	F	2	W	19.76	5.26	27.37918
68	M	2	D	20.35	14.98	NA
69	F	2	W	17.68	5.97	53.28191
70	F	2	W	23.66	6.13	48.53432
71	F	2	W	17.23	7.24	NA
72	F	2	W	25.74	4.88	37.65696
73	F	2	W	19.89	6.35	31.90467
74	F	1	D	14.24	3.95	28.87637
75	M	2	W	17.55	5.02	NA
76	M	2	W	16.32	5.86	NA
77	M	2	W	15.34	5.78	NA
78	F	2	W	11.64	4.87	22.87393
79	M	2	W	13.65	5.04	NA
80	M	2	W	11.57	5.24	NA
81	M	2	W	20.35	5.98	NA
82	M	2	W	8.91	4.58	NA
83	M	2	W	9.1	4.4	NA
84	M	2	D	21.65	7.81	NA
85	M	1	D	10.6	3.65	NA
86	M	1	D	12.35	9.57	NA
87	F	1	D	7.93	3.83	16.77475
88	F	1	D	8	4.26	19.49892

89	F	1	D	7.61	4.24	22.56011
90	M	1	W	11.96	5.62	NA
91	M	1	D	14.82	5.35	NA
92	F	1	W	14.43	5.08	34.81566
93	F	1	D	19.57	6.81	16.67624
94	F	1	W	12.55	3.25	13.19328
95	F	1	D	12.61	3.54	13.62372
96	F	1	D	10.21	4.49	18.52082
97	M	1	D	15.99	5.82	NA
98	F	1	D	32.24	4.8	25.20981
99	M	1	D	15.41	5.68	NA
100	M	1	D	13.98	5.45	NA
101	M	1	D	16.32	6.65	NA
102	M	1	D	6.37	3.31	NA
103	M	1	W	8.19	3.81	NA
104	M	1	W	12.29	3.95	NA
105	F	2	W	12.16	4.37	13.17322
106	F	2	W	16.19	4.43	26.32807
107	F	2	W	11.83	3.48	16.40101
108	F	2	W	10.47	3.9	17.56024
109	F	2	W	21.13	5.09	29.29508
110	F	2	W	18.59	4.49	21.51784
111	F	2	W	12.09	3.96	28.49073
112	F	2	W	13	3.83	30.98607
113	F	2	W	12.09	4.65	28.62749
114	F	2	W	13.26	4.48	25.66584
115	F	2	W	12.03	4.32	19.28812
116	F	2	W	17.36	5.01	30.00925
117	F	2	W	18.14	3.56	12.7591
118	F	2	W	15.93	4.65	22.72246
119	F	2	W	12.29	5.01	23.24402
120	F	2	W	17.42	4.38	18.35924
121	F	2	W	13.2	5.3	18.88097
122	F	2	W	14.5	5.01	21.06504
123	F	2	D	11.44	4.04	16.154
124	M	2	D	11.57	5.68	NA
125	M	2	W	15.28	3.9	NA
126	M	2	W	13.46	5.1	NA
127	M	2	W	13.2	4.76	NA
128	M	2	W	11.25	4.89	NA
129	M	2	W	16.58	7.54	NA
130	M	2	W	13.2	5.07	NA
131	M	2	W	14.04	5.65	NA
132	M	2	W	17.03	5.81	NA
133	M	2	W	17.81	4.88	NA
134	M	2	W	12.48	4.86	NA

135	M	2	W	11.44	4.34	NA
136	M	2	W	40.43	9.13	NA
137	M	2	D	14.3	4.53	NA
138	M	2	W	14.89	4.32	NA
139	M	2	W	16.77	4.4	NA
140	M	2	D	9.95	4.31	NA
141	M	2	W	10.34	4.36	NA
142	M	2	W	20.54	8.06	NA
143	F	1	W	12.81	6.25	26.73429
144	F	1	W	16.51	4.62	28.10653
145	M	1	D	11.12	6.71	NA
146	M	1	D	11.64	4.51	NA
147	M	1	W	18.92	7.57	NA
148	M	2	W	19.89	5.35	NA
149	U	3		9.69	4.23	NA
150	U	3		19.37	4.26	NA
151	U	3		19.76	4.56	NA
152	U	3		11.31	7.73	NA
153	U	3		11.25	3.81	NA
154	U	3		13.85	4.28	NA
155	U	3		17.62	4.54	NA
156	U	3		22.82	4.34	NA
157	U	3		18.14	10.33	NA
158	U	3		13.52	8.12	NA
159	U	3		21.58	5.79	NA
160	U	3		8.91	29.74	NA
161	U	3		9.17	3.14	NA
162	U	3		14.17	10.32	NA
163	U	3		12.09	6.7	NA
164	U	3		54.47	61.79	NA
165	U	3		10.4	4.2	NA
166	U	3		50.31	5.48	NA
167	U	3		33.74	9.61	NA
168	U	3		14.76	8.94	NA
169	U	3		22.3	6.16	NA
170	U	3		23.21	10.59	NA
171	U	3		19.24	5.66	NA
172	U	3		13.07	4.4	NA
173	U	3		49.14	6.21	NA
174	U	3		73.19	6.41	NA
175	U	3		37.05	4.75	NA
176	U	3		16.45	7.29	NA
177	U	3		43.81	6.09	NA
178	U	3		14.89	3.53	NA