IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 15, 2021, accepted April 29, 2021, date of publication May 3, 2021, date of current version May 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3077073

Combining Readability Formulas and Machine Learning for Reader-oriented Evaluation of Online Health Resources

YANMENG LIU[1], MENG JI[1], SHANNON SHANSHAN LIN[1,2], MENGDAN ZHAO[1], AND ZIQING LYV[1,3]

[1]School of Languages and Cultures, The University of Sydney, Sydney, NSW 2006, Australia
[2]Faculty of Health, University of Technology Sydney, Sydney, NSW 2006, Australia
[3]School of Foreign Languages, Jiangsu University of Science and Technology, Zhenjiang 212003, China

Corresponding author: Meng Ji (christine.ji@sydney.edu.au)

ABSTRACT Websites are rich resources for the public to access health information, and readability ensures whether the information can be comprehended. Apart from the linguistic features originated in traditional readability formulas, the reading ability of an individual is also influenced by other factors such as age, morbidities, cultural and linguistic background. This paper presents a reader-oriented readability assessment by combining readability formula scores with machine learning techniques, while considering reader background. Machine learning algorithms are trained by a dataset of 7 readability formula scores for 160 health articles in official health websites. Results show that the proposed assessment tool can provide a reader-oriented assessment to be more effective in proxy the health information readability. The key significance of the study includes its reader centeredness, which incorporates the diverse backgrounds of readers, and its clarification of the relative effectiveness and compatibility of different medical readability tools via machine learning.

INDEX TERMS Readability, health information, readability formula, machine learning, reader-oriented.

I. INTRODUCTION

The general population rely heavily on the internet as a main resource to search for health information [1]–[2]. And the number of internet user seeking health information has been increasing dramatically, for example, in U.S., over 74%of the population search for health information online [3]. And in a survey by British primary care physicians, 75% of their patient come with health information retrieved from the Internet [4]. Studies demonstrate the popularity of online health information seeking and the trend is continuous growing [5]–[7], and many professional organizations and governments are now providing electronic information for online access. Even for people who are in the middle of physical care, over half of them would use internet to learn more about treatment options, medications, and their medical conditions [7]–[9].

Although along with the advance of internet technology, searching online for health information is becoming more and more popular for the general population, the utility of health information online may largely depend on its readability. When seeking for health information online, the general public often find it difficult to understand the searching results [10], [11]. Concerns about the readability of online material have been demonstrated in various studies [12]–[15]. Hard-to-read health information online can cause confusion, misunderstanding, medical errors, let alone to guide its readers to do the right things according to their symptoms [16], [17]. The need for easy-to-read health information specifically on the topic of infectious diseases is even more urgent in recent years. The outbreaks of infectious diseases such as Ebola, Zika, Dengue fever and COVID-19 around the world require the advancement of effective communication strategies [18]. Typically, the occurrence of infectious diseases features rapid spreading, threatening people's lives in large quantity, and thus requires urgent response from the public, so the health information on this point is vital to public health and safety [19]. For example, during recent COVID-19 outbreak, internet has been serving as an important vehicle for the public to obtain disease-related information. Since the information is intended to the public, it would be very

The associate editor coordinating the review of this manuscript and approving it for publication was Shen Yin.

supportive for collectively break the chain of infection if it is easy to understand.

II. RESEARCH GAPS

The assessment of health and medical information understandability among the public has been mostly based on the use of formula-based readability tools (varieties and functionalities to be elaborated in the next section). Readability tools provide fast, convenient measures of key linguistic features considered as potential barriers to the effective comprehension of English medical materials. These linguistic assessment tools were originally developed for, clinically tested with readers who are native English speakers. With the increasing use of English as the main language in global health education and health promotion, the review and assessment of health educational resources written in English for readers and patients from non-native English backgrounds requires research-based evidence to inform and support global health education practice and policy making. This highlights the issue of the suitability of using existing readability tools to evaluate English materials for different, diverse reader groups. For example, the definition of difficult versus easy words (part of readability tools such as Gunning Fog Index, and Linsear Write Formula can be very different from readers with different English proficiency, health education levels and familiarity with health education traditions in major English-speaking countries. The interaction of these external factors may also have an impact on the accessibility of English health materials. In this study, we employed human raters from similar language and cultural backgrounds to review and evaluate the readability of online English health materials published by international and country-specific health authorities and not-for-profit health organizations. They were international students in English tertiary education. Their evaluation (interrater reliability reached over 0.7) reflected the actual level of understandability of the collected resources for evaluation. We then tested the validity and effectiveness of using a variety of existing readability tools to predict the readability of English health materials based on the evaluation results from the target readership, that is, international students enrolled as research students in Australian universities. Our study aimed to fill in the gap in health readability research, whereby the effectiveness of using existing readability tools to assess English health materials for non-native speakers with very limited if any exposure to English-based health education environments and traditions. Our study also evaluated the effectiveness of using a variety of machine learning algorithms for the automated calculation of health information readability to aid in the rapid, user-adaptive evaluation of health education materials, which has become increasing relevant, significant in global health education and health promotion.

III. READABILITY MEASUREMENT

Measuring the readability is one important step toward making the health information understandable by the public. Thanks to continuous efforts by pioneering scholars in the field, several readability formulas were developed and widely used, including Flesch Reading Ease Score [20], Gunning Fog Index [21], Fleasch-Kincaid Grade Level [20], Coleman-Liau Index [22], Simple Measure of Gobbledygook (SMOG) Index [23], Automated Readability Index [24], Lensear Write Formula [25], etc. Readability formulas are easy to calculate and provide quantitative evidence to support the research, as the formulas are built based on linguistic features, like word length, sentence length, etc. shown in Table 1. These formulas have been supporting researchers' studies for assessing and subsequently improving readability of assorted types of reading materials [26], [27].

As its definition goes "... the level of ease or difficulty with which text material can be understood by a particular reader who is reading that text ..." [28], readability is a relative concept which largely depends on the readers of the text. While in the calculation of readability scores by mentioned formulas, diverse features of readers are hardly considered. The idea of "one score fits all" underlying mentioned readability formulas obviously have large space for improvement on this point.

Meanwhile, several scholars employ human assessors to rate the readability difficulty of medical materials [29]. Manual practice is both time-consuming and costly [30]. And for the massive health information on the internet, human rating is far from practical [31]. Another concern of human rating is the individual bias. Inconsistency of assessment is unable to provide a reliable reference for readability estimation and suggestions for easy-to-read improvement. But it cannot be denied that human assessment takes readers' characteristics into consideration in a satisfying way, as assessors make the judgement out of their backgrounds, which represents a group of people sharing similar social and cultural features. Therefore, when using human assessment results as criteria for readability estimation, inter-rater and intra-rater consistency is very important to ensure the positive use of human assessment.

IV. MACHINE LEARNING FOR READABILITY ESTIMATION

Machine learning (ML) provides an alternative way to estimate readability of written texts. Applying ML method into readability assessment is not new, and numerous scholars have come to the consensus that ML methods can help to improve current readability estimation. In ML, readability estimation is treated as a process of classification, and ML techniques work as a classifier to predict which readability level the text belongs to, with the guidance of multiple statistical features. This approach enjoys the advantage that the whole process is data-driven, and the data are obtained automatically, with less manual labor and human bias involved.

Several studies introduce ML methods into readability estimation, in a hope to include more linguistic features in the evaluation process than traditional readability formulas do. Variables processed by ML methods in previous studies were enlarged from grammatical features to semantic

TABLE 1. Outline of seven readability formulas.

Readability tools	Formulas	Interpretations of scores
Flesch Reading Ease Score	Score=206.835-(1.015*ASL[a]) - (84.6*ASNW[b])	The score is a number from 0 to 100 - a higher score indicates easier reading.
Gunning Fog	Score =0.4*(ASL[a]+PHW[c])	The index estimates the years of formal education needed to understand the text on a first reading.
Fleasch-Kincaid Grade Level Readability	Score= (0.39 * ASL[a]) + (11.8*ASNW[b])-15.59	The output approximates the U.S. grade level that readers need to comprehend the text.
Coleman-Liau Index	Score=5.89*ACW[d]-0.3*sentence/(100*words)-15.8	The output approximates the U.S. grade level that readers need to comprehend the text.
SMOG Index	Score = 3 + Square Root of Polysyllable Count	The output approximates the U.S. grade level that readers need to comprehend the text.
Automated Readability Index	Score= (ANLW[e]*4.71) +(ANWS[f]*0.5)-21.43	The output approximates the U.S. grade level that readers need to comprehend the text.
Linsear Write Formula	Score= (Easy words*1+Hard words*3)/Sentence count If >20, divided by '2'; If <20 or =20, subtract '2', and then divide by '2'.	The output approximates the U.S. grade level that readers need to comprehend the text.

[a]ASL: Average Sentence Length.
[b]ASNW: Average Syllables Number per Word.
[c]PHW: Percentage of Hard Words.
[d]ACW: Average Character per Word.
[e]ANLW: Average Number of Letters per Word
[f]ANWS: Average Number of Words in Sentences

features, and the readability accuracy was improved to some extent [31]–[33]. Si and Callan conducted preliminary work of using unigram language model to predict the difficulty of reading science web pages [32]. They argued that readability formulas ignore the content information in the evaluation, so they used unigram to represent the content in their experiment. The result showed that the proposed method achieved more accurate assessment for readability of science web pages. With the help of a variation of the multinomial naïve Bayes classifier, Collins-Thompson and Callan captured semantic difficulty across grade levels [31]. It concluded that reasonably accurate readability measures can be built by using simple statistical language modelling techniques. Similarly, another study deployed Support Vector Machine to classify queries from different grade categories [33]. Both syntactic and semantic features were derived from queries, and the model worked better than readability formulas.

However, when applying the ML techniques, researchers grouped readability levels according to the grade or the age of the author in advance. Then, they tested the model accuracy through comparing the prediction results with the grouping. Unfortunately, the grade or the age of authors cannot represent the readability level because the readability is assessed from the perspective of readers rather than the authors. In other words, the grade or age of the author does not necessarily equal to the grade or age of readers who can understand the writing of the very author. This would directly influence the statistical modelling and prediction accuracy.

Some other researchers also apply ML into readability evaluation. They labelled the readability levels of texts by human, rather than grade or years of education. Of course, human labelling results is more reader oriented. What's more, it gives the ML techniques a full play as this approach is more sophisticated at learning the regularities at training phase, so as to be more accurate at approximating the human results at testing phase [34]. In many studies, when employing human to label, researchers would disregard features of human labelers in ML approach [33]–[36]. Even though they have described the demographic features or education background of the human labelers in their studies, these features were not taken into consideration in the analysis or discussion. This also supports the studies concluding that human labelers do less well in some studies, because different groups may have different understandings of reading level for the same text [37]. These features are worthy being studied, because the human labelers share similar backgrounds, and the use of ML can be a means of tuning the readability models to the needs of a particular group of readers.

Another neglect in previous ML approach in readability assessment studies is the value of traditional readability formulas. Literatures mostly extract linguistic features at lexical, syntactical and textual level as ML features to study [38]–[43]. Undoubtedly, these features influence the readability level of texts, but the studies tend to deny the value of traditional formulas as they were developed based on 'superficial' features of word and sentence length. This study argues that the traditional readability formulas may have their values in judging readability levels, therefore, the study deploys statistics by traditional readability formulas as features of texts to train and test ML models, so as to combine traditional readability formulas together with ML approach to better serve the readability studies.

Given this context, this study presents automatic readability assessment models based on a case of 160 infectious diseases articles on websites in Australia. The models are developed based on mentioned readability formulas

and human evaluation for the purpose of supporting the readability evaluation with concrete statistical evidence and considering reader characteristics in the evaluation. And ML methods are adopted to approximate human readability assessment results, and the proposed reader-oriented readability models can adjust themselves according to specific reader features. To our best of knowledge, this study is the first co-designed empirical study that examines the reading difficulties of infectious disease health information online in English language and uses a mix methodology to build up the readability assessment models, combining existing readability formulas and the machine-learning algorism in a reader-oriented manner.

V. METHODOLOGY

A. STUDY DESIGN

This study of medical readability evaluation consisted of three steps. First, the materials on the topic of infectious diseases were rated manually by human assessors, and human rating results are treated as a golden standard for readability assessment. This was also the labelling procedure for ML training and testing later. Second, seven popular readability formulas were deployed to get readability scores of infectious diseases in a traditional manner. Multiple rating scores provided statistics for ML to build up models for readability evaluation. Third, nine ML tools were trained with mentioned data and tested with reference to human rating result to verify the proposed models.

The design of this study emphasized that readability is assessed from the perspective of readers, which requires that the assessment model must be adaptive enough to reflect different groups of readers' competences to understand the medical materials, because in reality, readers are so diverse in cultural, social and educational backgrounds. In other words, a reader-oriented readability evaluation model is realized by ML method, which is actually approximating itself to a human rating process, ending up with similar human judgement of readability.

B. DATA COLLECTION

In this study, we investigated methods to estimate the readability level of online medical information, specifically, infectious diseases education texts for the public. The source of materials is credible international health websites, i.e., Health on the Net [44], Australian Government Department of Health [45], State Ministry of Health [46], [47]. A total of 1,200 individual articles of varying length were collected, and 160 of them were selected randomly as samples to study in the present study, covering different readability difficulties.

Human assessors were asked to read and rate all the sample materials according to their instincts on the difficulty of readability. The assessors were screened to have English as their second language, and were aged from 18 to 35 years old, from non-English speaking families, with advanced bilingual skills and high education level, yet limited medical knowledge about infectious diseases. Each assessor was presented with 1,200 materials individually. They were requested to rate the readability of the materials on a 10-point Likert scale with their subjectively instant understanding of a text. The one-point indicated no effort would require from the readers to understand the text, and the text appeared to be extremely easy to understand. The five-point would require some effort in reading to understand with the overall reasonable understandability, whereas the 10-point indicated the highest level of understandability despite the enormous effort from the readers, they could still feel confused or lost after reading. To minimize the fatigue from the assessors, each assessor was given four days to complete the task. On average, they finished 300 materials per day. The inter-rater reliability and intra-rater consistency were checked to eliminate influence of individual bias. The rating results of assessors are standardized with z score as the final readability score by human.

Seven methods have been used to estimate the readability of infectious disease materials, including Flesch Reading Ease Score [20], Gunning Fog Index [21], Fleasch-Kincaid Grade Level [20], Coleman-Liau Index [22], Simple Measure of Gobbledygook (SMOG) Index [23], Automated Readability Index [24], Lensear Write Formula [25]. That is to say, each online medical information sample would have seven readability scores as its statistical features to further train and test ML models. Here, we outline seven readability formulas used in this study (Table 1), whose scores would be further studied for ML training and testing later.

C. DEVELOPING ML MODELS

Using the seven medical resource readability evaluation formulas, we constructed and compared the performance of nine ML models: XGBoost Tree, Random Trees, Bayes Net, Random Forest, C&R Tree, C5.0, CHAID, Quest and Neural Net. Figure 1 (data presented in Tables 3, 4) shows the gains of the different ML models at the model training and testing stages. In ML, information gain is often exchangeable with Kullback–Leibler divergence, or the amount of information gained about a random variable or signal from observing another random variable. In our study, we used the seven well-established medical resource readability evaluation tools to predict the likely outcomes of the target variable under investigation, i.e. the human judgement of medical material readability.

The development of ML models is known as decision tree training, which identifies predictor variables and their associated value ranges in the classification process of the categorical target variable values. Specifically, in our study, the ML modelling explores the value ranges of the formula-based readability tools which can be used as effective variables to explain the variations in the readability scores by independent human assessors. The symbol $ indicates the estimated value/state of the target variable. In order to train the ML models, we first converted the standardized scores of the human evaluation into binary data (Table 2). Human

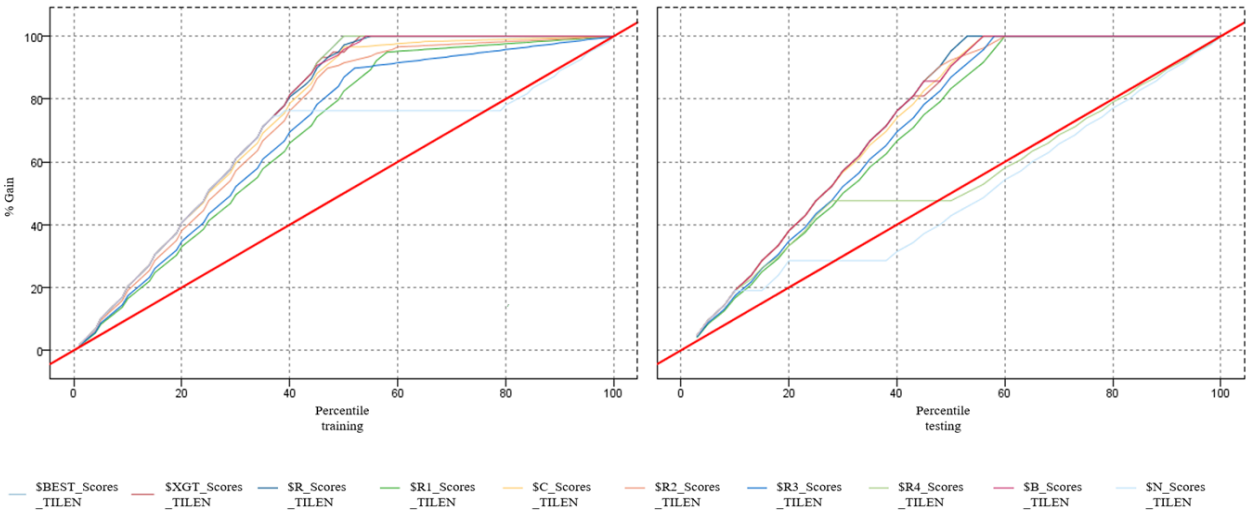


FIGURE 1. Comparison of Gains of different ML model.

TABLE 2. Binarization of human-based evaluation.

Band	Lower Limit Value	Upper Limit Value
1 (higher readability)	>=5.1802944	< 6.37875764
2 (lower readability)	>=6.37875764	< 7.57722052

evaluation scores between 5.180 and 6.379 were classified as Band 1 (TILEN – higher readability or lower difficulty group of texts); and human evaluation scores between 6.379 and 7.577 were classified as Band 2 (TILEN-lower readability or higher difficulty group of texts). The ML models predicted the probability of the group affiliation (high readability 1 versus low readability 2) of each text based on the automatic evaluation results of the seven medical readability instruments (Flesch Reading Ease, Gunning-Fog Index; Flesch-Kincaid US Grade Level; Coleman-Liau Index; SMOG Index; Automated readability index; Lensear Write Formula).

VI. RESULTS

Tables 3 and 4 show the incremental gains of the ML models. The BEST-Scores refers to the ensembled model based on the optimization of the 9 discrete ML models. In the model training stage, at the initial 10% level, the ensembled model ($BEST scores_ TILEN) explained as much as one fifth (20.339%) of the total variation of the binarized scores of the target variable, i.e. the readability of infectious disease educational resources based on the human assessment. The model gains then increased steadily by an average 20% in the next four stages before completing the iteration after analyzing 50% of the total corpus texts used for model training. In the testing stage, the ensembled model ($BEST scores_TILEN) first explained as much as 19.047% of the total variation of the testing corpus texts. The testing process completed after the model evaluated 60% of the total corpus texts used for model testing. Tables 3 and 4 compare the performance of the 9 ML models at both the model training and testing stages. Random Forest ($R4) shows a similar performance at the ensembled model at the training stage, yet its efficiency at the model testing stage drops significantly. By contrast, C&R Tree, C5.0, CHAID and Quest had better performance at the model testing stage, but less effective performance at the model training stage. The three best ML models were XGBoost Tree ($XGT-Scores), Random Trees ($R-Scores) and Bayes Net ($B-Scores) which completed the classification after analyzing 60% of the infectious disease corpus texts in both the model training and testing stages.

Figures 2 shows the classification of the target variable (human evaluation of the readability of infectious disease educational resources) using Flesch Reading Ease and Lensear Write Formula scores as the two large predictor variables (predicator importance: 25, 18, respectively). Medical texts about infectious diseases were classified into two contrastive groups: Y-axis shows the predicted group affiliation of medical: Band 1 included texts of higher readability (human-based evaluation of difficulty scores ranging between 5.18-6.38); Band 2 clustered texts of lower readability (human evaluation of difficulty scores ranging 6.38 and 7.58). X-axis shows the Flesch Reading Ease (FRE) scores generated automatically by the formula. XGBoost Tree identified two value ranges of Flesch Reading Ease Scores: 0-50 and 50-100. In Figure 2, a large number of medical texts of higher readability (Band 1 on Y-axis: $XGT-Scores_Bin) have FRE values between 38-85; by contrast, the majority of medical texts of lower readability (Band 2 on Y-axis) have FRE scores between 0-38. These value ranges were identified by XGBoost as key grammatical and lexical features (averaged sentence lengths in words and averaged word lengths in syllables) of medical texts on infectious diseases. Lensear Write Formula was used as another important predicator variable in the XGBoost Tree modelling process. In Figure 2, within the cluster of medical texts of higher readability ($XGT: Band 1), the majority of texts had Lensear Write Formula scores between 0-15; whereas within medical

TABLE 3. Comparison of information gains of ml models.

	Best Fit Line		XGBoost Tree		Random Trees		Bayes Net		Random Forest	
	$BEST-Scores		$XGT-Scores		$R-Scores		$B-Scores		$R4-Scores	
Percent	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
%					Gains %					
10	20.339	19.0476	20.339	19.0476	20.339	19.0476	20.339	19.0476	20.339	19.0476
20	40.678	38.0952	40.678	38.0952	40.678	38.0952	40.678	38.0952	40.678	33.3333
30	61.017	57.1429	61.017	57.1429	61.017	57.1429	61.017	57.1429	61.017	47.619
40	81.3559	76.1905	81.3559	76.1905	80.678	76.1905	81.3559	76.1905	81.3559	47.619
50	100	95.2381	94.9153	90.4762	97.1751	95.2381	95.7627	90.4762	100	47.619
60	100	100	100	100	100	100	100	100	100	58.0952
70	100	100	100	100	100	100	100	100	100	68.5714
80	100	100	100	100	100	100	100	100	100	79.0476
90	100	100	100	100	100	100	100	100	100	89.5238
100	100	100	100	100	100	100	100	100	100	100

TABLE 4. Comparison of information gains of ml models.

	C&R Tree		C5.0		CHAID		Quest		Neural Net	
	$R1-Scores		$C-Scores		$R2-Scores		$R3-Scores		$N-Scores	
Percent	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
%					Gains %					
10	16.507	16.6667	20.339	19.0476	19.0678	19.0476	17.3866	17.3913	20.339	19.0476
20	33.014	33.3333	40.678	38.0952	38.1356	38.0952	34.7731	34.7826	40.678	28.5714
30	49.521	50	59.661	56.71	57.2034	57.1429	52.1596	52.1739	61.017	28.5714
40	66.028	66.6667	78.6441	74.026	76.2712	76.1905	69.5462	69.5652	76.2712	31.4286
50	82.535	83.3333	96.6102	91.342	91.5254	92.381	86.9328	86.9565	76.2712	42.8571
60	95.2144	100	97.6271	100	96.6102	100	91.5839	100	76.2712	54.2857
70	96.4108	100	98.5472	100	97.4576	100	93.6879	100	76.2712	65.7143
80	97.6072	100	99.0315	100	98.3051	100	95.7919	100	78.0965	77.1429
90	98.8036	100	99.5157	100	99.1525	100	97.896	100	89.0482	88.5714
100	100	100	100	100	100	100	100	100	100	100

texts of lower readability ($XGT: Band 2), a large number of texts had Lensear Write Formula scores between 15-20. Findings in Figure 2 suggest that higher readability (5.18-6.38 on 1-10 difficulty scale) tends to be associated with higher Flesch Reading Ease scores (38-85) and lower Lensear Write Formula scores (0-15); by contrast, lower readability of medical texts on infectious diseases (6.38 and 7.58 on 1-10 difficulty scale) tends to be associated with lower Flesch Reading Ease Scores (0-38) and higher Lensear Write Formula scores (15-20). The effectiveness of the other predicator variables (the other 5 automatic medical readability formula) was largely reduced compared with the FRE scores and Lensear Write Formula. For example, in Figure 3, most texts of both higher and lower readability clustered within the value range of 0 and 20 of Coleman-Liau Index scores, reflecting the limited effectiveness of this predictor variable in separating medical resources of varying reading difficulties.

With the Random Tree Model, the predictor importance of Flesch Reading Ease (FRE) scores remains the largest predicator (0.35). This was followed by Lensear Write Formula (0.17) and Coleman-Liau Index (0.11). Figures 4 and 5 show the decision tree modelling outcomes of the Random Tree Model. Similar to the findings presented in Figures 2 and 3 of the XGBoost Tree Model, Flesch Reading Ease (FRE) scores and Lensear Write Formula scores facilitated the automatic classification process. The value range between 40 and 85 on the Flesch Reading Ease scale characterized the higher readability of medical resources on infectious diseases (data points marked by their group affiliation of 1 along the Y-axis of $R-Scores). By contrast, the lower value range of 0 to 40 of the FRE scores proved to be indicative of medical texts of lower readability (illustrated by their affiliation to $R-scores 2 along the Y-axis). On the other hand, medical resources of higher readability (clustered at the bottom of the graph) tended to be marked by cold colors (dark to bright blue

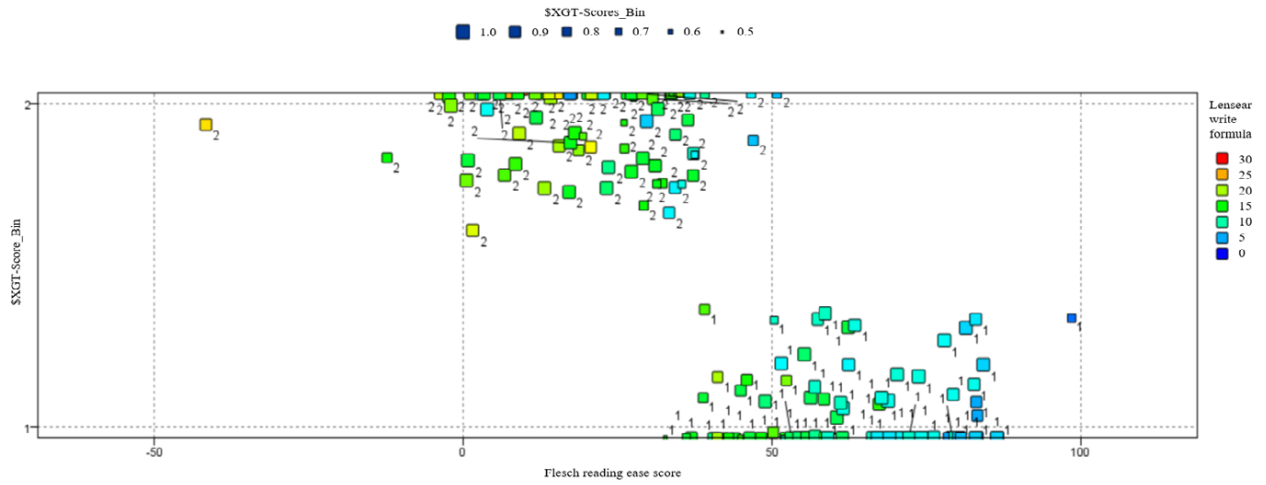


FIGURE 2. XGBoost Tree: Flesch Reading Ease Score and Lensear Write Formula as predictor variables.

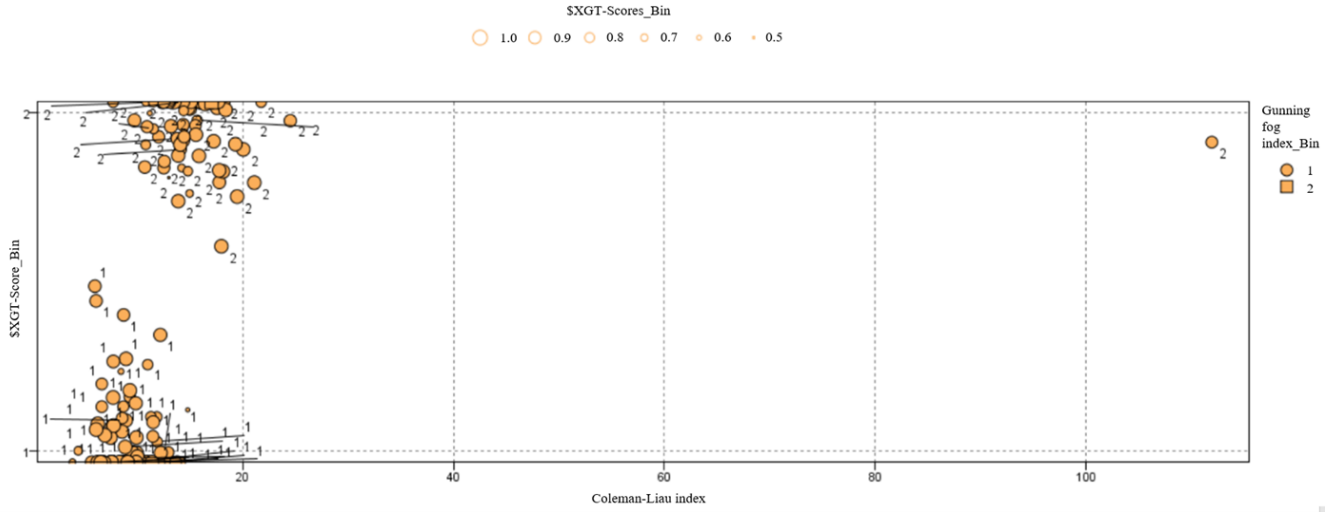


FIGURE 3. XGBoost Tree – Coleman-Liau Index and Gunning Fog Index as predicator variables.

and bright green), indicating lower Lensear Write Formula scores between 0 to 15. Medical resources of higher difficulty (clustered at the top of the graph) were distinguished by relatively warmer colors in Figure 4 (bright green to orange), linked with relatively higher Lensear Write Formula scores of between 15 and 25. Figure 5 shows the classification result of the Random Tree Model using Gunning Fog Index and Coleman-Liau Index as the predicator variables. Echoing the findings presented in Figure 3, neither of these two readability tools proved effective in separating medical texts of lower versus higher readability scores based on the human evaluation.

VII. DISCUSSIONS

A. SUMMARY OF FINDINGS

The distinguishing value ranges between the Flesch Reading Ease Scores and Lensear Write Formula scores underpin the validity of the XGBoost Tree Model and the Random Tree Model. Important similarities were identified: morphological complexity (measured by average syllables per word) and sentence length (measured by average words per sentence). These were the key surface linguistic features in evaluating the reading difficulties of health information on infectious disease as the focus of our study. Distinct mathematical formulas were developed for these readability evaluation tools over the years which are widely used in health education. There has been however controversy around the inconsistency of the results generated by these medical readability instruments. In our study, we used machine learning to reassess the relative effectiveness and compatibility of different medical readability tools when compared with the human evaluation results which were closest to the actual reading experience of the target readership, that is, young adults (18-25) from non-English speaking families, advanced bilingual skills, high education level, yet limited medical knowledge of infectious diseases. Our study suggests that existing controversy around the inconsistency among readability evaluation tools can be effectively clarified by

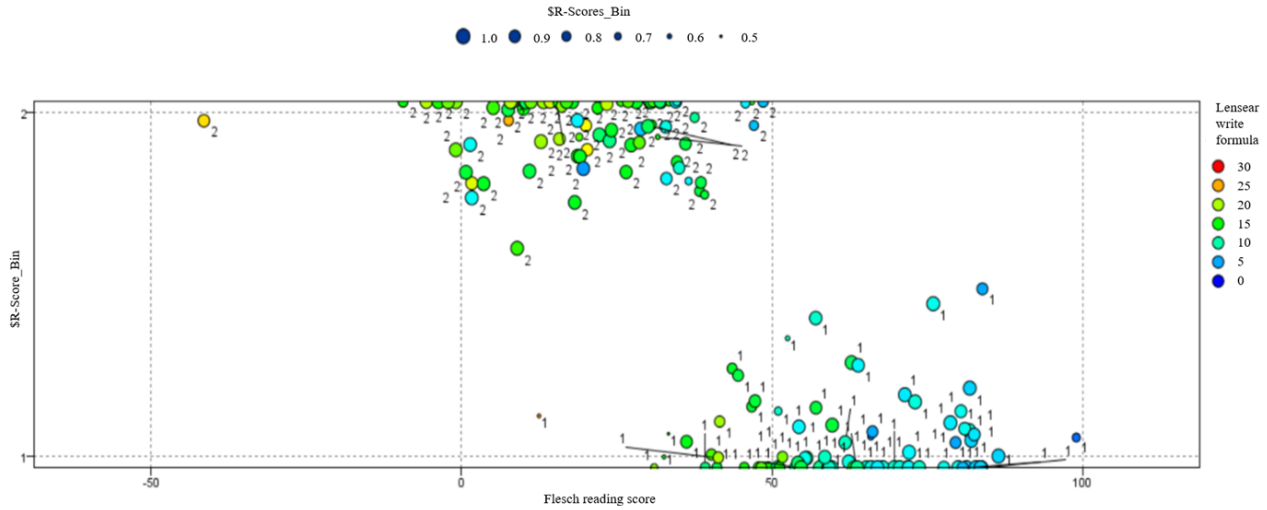


FIGURE 4. Random Tree Model – Flesch Reading Ease and Linsear Write Formula as predicator variables.

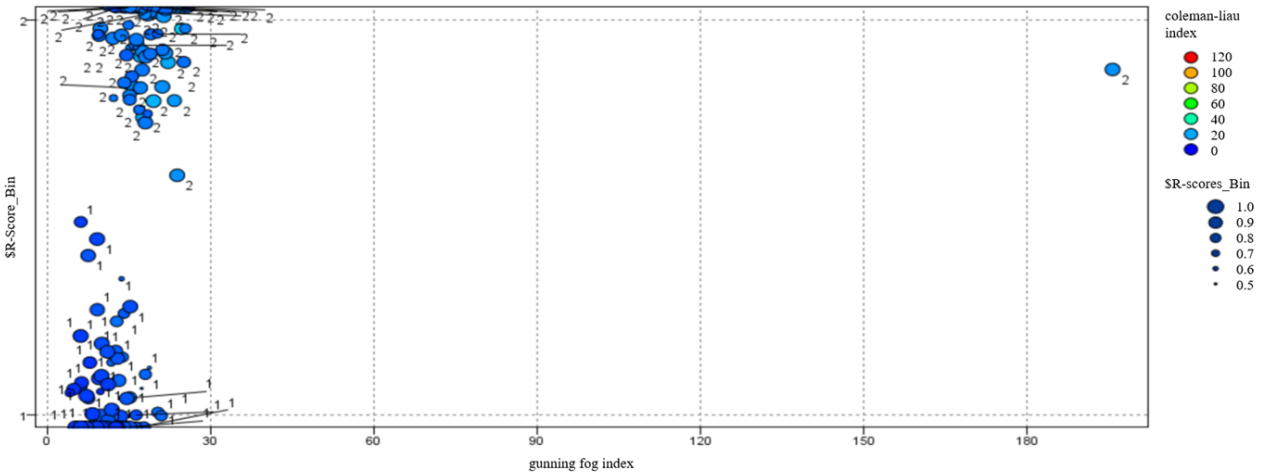


FIGURE 5. Random Tree Model – Gunning Fog Index and Coleman-Liau Index.

weighting the relative predicator importance of the formula-based evaluation results in the iterative process of ML training and testing using the human evaluation outcome as the target variable.

B. LIMITATION AND FUTURE WORK

Health information readability assessment represents an increasingly complex issue in health education research and health resource development, given the use of English as the dominant language in global health education and promotion. Our study explored the effectiveness of using machine learning to predict the readability of health materials for non-English speaking people, with a particular focus on international students enrolled in tertiary institutes of English-speaking countries (Australia, in this case). The selection of human raters from very similar language and cultural backgrounds, the same age group (18-25), and educational levels (university graduates) has ensured the consistency of the evaluation results, and subsequently, the precision, effectiveness of the machine learning algorithms developed to predict the readability of health materials for this reader group. In future research, more experiments are needed to develop models to predict health information suitability for different people, as it is suspected that for diverse readerships, the accuracy of the models may well vary. Further, readability tools rely on a very small number of linguistic and textual features such as average word length and average sentence length. This has largely simplified the complex research topic of health information accessibility. More linguistically rich health information analysis is required to allow the discovery of new linguistic interventions and methods to improve the readability of health information for diverse readers. This can be achieved through introducing more linguistically rich text annotation, analysis using natural language processing technologies.

VIII. CONCLUSION

For infectious diseases medical texts available on the Internet, the study findings reveal that a reader-oriented readability assessment can be achieved through a combination of

traditional readability formulas and ML techniques. The proposed methods approximate the human rating results well, as empirically demonstrated by high training and testing accuracy of the ML. More importantly, methods from the field of ML can be used to reassess and complement existing readability assessment methodology with a reference to human evaluation results. And the existing controversy around the inconsistency among readability evaluation formulas can be effectively clarified by weighting the relative predicator importance of the formula-based evaluation results in the iterative process of ML training and testing using the human evaluation outcome as the target variable.

Therefore, this study can shed some light on readability studies to provide better customized medical services by combining traditional readability formulas with ML approach as it can approximate human judgement results at a high accuracy, and this can be a time and cost-efficient way to curb the current public health communication crisis.

REFERENCES

[1] T. Szmuda, C. Özdemir, S. Ali, A. Singh, M. T. Syed, and P. Słoniewski, "Readability of online patient education material for the novel coronavirus disease (COVID-19): A cross-sectional health literacy study," *Public Health*, vol. 185, pp. 21–25, Aug. 2020, doi: 10.1016/j.puhe.2020.05.041.

[2] H. S. Wald, C. E. Dube, and D. C. Anthony, "Untangling the Web—The impact of Internet use on health care and the physician–patient relationship," *Patient Educ. Counseling*, vol. 68, no. 3, pp. 218–224, Nov. 2007, doi: 10.1016/j.pec.2007.05.016.

[3] P. Joseph, N. A. Silva, A. Nanda, and G. Gupta, "Evaluating the readability of online patient education materials for trigeminal neuralgia," *World Neurosurg.*, vol. 144, pp. e934–e938, Dec. 2020, doi: 10.1016/j.wneu.2020.09.123.

[4] M. Malone, "Health and the Internet-changing boundaries in primary care," *Family Pract.*, vol. 21, no. 2, pp. 189–191, Apr. 2004, doi: 10.1093/fampra/cmh215.

[5] R. W. Hsieh, L. Chen, T.-F. Chen, J.-C. Liang, T.-B. Lin, Y.-Y. Chen, and C.-C. Tsai, "The association between Internet use and ambulatory care-seeking behaviors in Taiwan: A cross-sectional study," *J. Med. Internet Res.*, vol. 18, no. 12, p. e319, Dec. 2016, doi: 10.2196/jmir.5498.

[6] D. Boughtwood, C. Shanley, J. Adams, Y. Santalucia, H. Kyriazopoulos, D. Pond, and J. Rowland, "Dementia information for culturally and linguistically diverse communities: Sources, access and considerations for effective practice," *Austral. J. Primary Health*, vol. 18, no. 3, pp. 190–196, 2012, doi: 10.1071/PY11014.

[7] F. Meric, "Breast cancer on the world wide Web: Cross sectional survey of quality of information and popularity of websites," *BMJ*, vol. 324, no. 7337, pp. 577–581, Mar. 2002, doi: 10.1136/bmj.324.7337.577.

[8] J. A. Diaz, R. A. Griffith, J. J. Ng, S. E. Reinert, P. D. Friedmann, and A. W. Moulton, "Patients' use of the Internet for medical information," *J. Gen. Internal Med.*, vol. 17, no. 3, pp. 180–185, Mar. 2002, doi: 10.1046/j.1525-1497.2002.10603.x.

[9] Y.-Y. Chen, L. Chen, T.-S. Huang, W.-J. Ko, T.-S. Chu, Y.-H. Ni, and S.-C. Chang, "Significant social events and increasing use of life-sustaining treatment: Trend analysis using extracorporeal membrane oxygenation as an example," *BMC Med. Ethics*, vol. 15, no. 1, p. 21, Dec. 2014, doi: 10.1186/1472-6939-15-21.

[10] M. R. A. Hamid, M. M. Isamudin, F. N. B. Allam, and S. S. Buhari, "Understandability and actionability of Web-based education materials on hypertension management," *Environ.-Behav. J.*, vol. 5, no. 14, pp. 127–133, Jul. 2020, doi: 10.21834/ebpj.v5i14.2230.

[11] J. Palotti, G. Zuccon, and A. Hanbury, "Consumer health search on the Web: Study of Web page understandability and its integration in ranking algorithms," *J. Med. Internet Res.*, vol. 21, no. 1, Jan. 2019, Art. no. e10986, doi: 10.2196/10986.

[12] M. R. M. S. Giorgi, O. S. D. de Groot, and F. G. Dikkers, "Quality and readability assessment of websites related to recurrent respiratory papillomatosis," *Laryngoscope*, vol. 127, no. 10, pp. 2293–2297, Oct. 2017, doi: 10.1002/lary.26521.

[13] V. Narwani, K. Nalamada, M. Lee, P. Kothari, and R. Lakhani, "Readability and quality assessment of Internet-based patient education materials related to laryngeal cancer," *Head Neck*, vol. 38, no. 4, pp. 601–605, Apr. 2016, doi: 10.1002/hed.23939.

[14] S. A. MacLean, C. H. Basch, D. Ethan, and P. Garcia, "Readability of online information about HPV Immunization," *Hum. Vaccines Immunotherapeutics*, vol. 15, nos. 7–8, pp. 1505–1507, Aug. 2019, doi: 10.1080/21645515.2018.1502518.

[15] D. Nutbeam, "Health literacy as a public health goal: A challenge for contemporary health education and communication strategies into the 21st century," *Health Promotion Int.*, vol. 15, no. 3, pp. 259–267, Sep. 2000, doi: 10.1093/heapro/15.3.259.

[16] R. W. White and E. Horvitz, "Cyberchondria: Studies of the escalation of medical concerns in Web search," *ACM Trans. Inf. Syst.*, vol. 27, no. 4, pp. 1–37, Nov. 2009, doi: 10.1145/1629096.1629101.

[17] R. White, "Beliefs and biases in Web search," presented at the 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Dublin, Ireland, 2013, doi: 10.1145/2484028.2484053.

[18] J. Parmer, C. Baur, D. Eroglu, K. Lubell, C. Prue, B. Reynolds, and J. Weaver, "Crisis and emergency risk messaging in mass media news stories: Is the public getting the information they need to protect their health?" *Health Commun.*, vol. 31, no. 10, pp. 1215–1222, Oct. 2016, doi: 10.1080/10410236.2015.1049728.

[19] D. Toppenberg-Pejcic, J. Noyes, T. Allen, N. Alexander, M. Vanderford, and G. Gamhewage, "Emergency risk communication: Lessons learned from a rapid review of recent gray literature on Ebola, Zika, and yellow fever," *Health Commun.*, vol. 34, no. 4, pp. 437–455, Mar. 2019, doi: 10.1080/10410236.2017.1405488.

[20] R. Flesch, "A new readability yardstick," *J. Appl. Psychol.*, vol. 32, no. 3, p. 221, 1948.

[21] R. Gunning, *Technique of Clear Writing*. New York, NY, USA: McGraw-Hill, 1952.

[22] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring," *J. Appl. Psychol.*, vol. 60, no. 2, pp. 283–284, 1975, doi: 10.1037/h0076540.

[23] G. H. M. Laughlin, "SMOG grading—A new readability formula," *J. Reading*, vol. 12, no. 8, pp. 639–646, 1969.

[24] R. Senter and E. A. Smith, "Automated readability index," Aerosp. Med. Res. Lab., Wright-Patterson Air Force Base, OH, USA, Tech. Rep. AMRL-TR-66-220, 1967.

[25] J. O'Hayre, *Gobbledygook Has Gotta Go*. Washington, DC, USA: U.S. Department of the Interior, Bureau of Land Management, 1966.

[26] G. S. Yi and A. Hu, "Quality and readability of online information on in-office vocal fold injections," *Ann. Otol., Rhinol. Laryngol.*, vol. 129, no. 3, pp. 294–300, Mar. 2020, doi: 10.1177/0003489419887406.

[27] A. P. O. Ferster and A. Hu, "Evaluating the quality and readability of Internet information sources regarding the treatment of swallowing disorders," *Ear, Nose Throat J.*, vol. 96, no. 3, pp. 128–138, Mar. 2017, doi: 10.1177/014556131709600312.

[28] H. Kim, Q. Zeng-Treitler, S. Goryachev, A. Keselman, L. Slaughter, and C. A. Smith, "Text characteristics of clinical reports and their implications for the readability of personal health records," in *Proc. 12th World Congr. Health Medical Inform., Building Sustain. Health Syst. (Medinfo)*, 2007, vol. 129, no. 2, pp. 1117–1121.

[29] J. Zheng and H. Yu, "Readability formulas and user perceptions of electronic health records difficulty: A corpus study," *J. Med. Internet Res.*, vol. 19, no. 3, p. e59, Mar. 2017, doi: 10.2196/jmir.6962.

[30] R. Nagata, T. Iguchi, F. Masui, and A. Kawai, "A method for rating English texts by reading level for Japanese learners of English," *Syst. Comput. Jpn.*, vol. 36, no. 6, pp. 1–13, Jun. 2005, doi: 10.1002/scj.20326.

[31] L. Feng, "Automatic readability assessment," Ph.D. dissertation, City Univ. New York, 2010.

[32] K. Collins-Thompson, "Computational assessment of text readability: A survey of current and future research," *Recent Adv. Automat. Readability Assessment Text Simplification*, vol. 165, no. 2, pp. 97–135, Dec. 2014, doi: 10.1075/itl.165.2.01col.

[33] L. Si and J. Callan, "A statistical model for scientific readability," presented at the 10th Int. Conf. Inf. Knowl. Manage., Atlanta, GA, USA, 2001, doi: 10.1145/502585.502695.

[34] X. Liu, W. B. Croft, P. Oh, and D. Hart, "Automatic recognition of reading levels from user queries," presented at the 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Sheffield, U.K., 2004, doi: 10.1145/1008992.1009114.

[35] T. François and E. Miltsakaki, "Do NLP and machine learning improve traditional readability formulas?" presented at the 1st Workshop Predicting Improving Text Readability Target Reader Populations, Montreal, QC, Canada, 2012.

[36] S. Alotaibi, M. Alyahya, H. Al-Khalifa, S. Alageel, and N. Abanmy, "Readability of arabic medicine information leaflets: A machine learning approach," *Procedia Comput. Sci.*, vol. 82, pp. 122–126, Jan. 2016, doi: 10.1016/j.procs.2016.04.017.
[37] E. Pitler and A. Nenkova, "Revisiting readability: A unified framework for predicting text quality," presented at the Conf. Empirical Methods Natural Lang. Process., Honolulu, HI, USA, 2008.
[38] S. E. Petersen and M. Ostendorf, "A machine learning approach to reading level assessment," *Comput. Speech Lang.*, vol. 23, no. 1, pp. 89–106, Jan. 2009.
[39] C. H. Björnsson, "Readability of newspapers in 11 languages," *Reading Res. Quart.*, vol. 18, no. 4, pp. 480–497, 1983, doi: 10.2307/747382.
[40] R. Zowalla, M. Wiesner, and D. Pfeifer, "Automatically assessing the expert degree of online health content using SVMs," *Stud. Health Technol. Inf.*, vol. 202, pp. 48–51, Jan. 2014. [Online]. Available: http://europepmc.org/abstract/MED/25000012
[41] M. Terblanche and L. Burgess, "Examining the readability of patient-informed consent forms," *Open Access J. Clin. Trials*, vol. 2, pp. 157–162, Oct. 2010.
[42] M. Adnan, J. Warren, and M. Orr, "Assessing text characteristics of electronic discharge summaries and their implications for patient readability," presented at the 4th Australas. Workshop Health Inform. Knowl. Manage., Brisbane, QLD, Australia, vol. 108, 2010.
[43] M. Wiesner, R. Zowalla, and M. Pobiruchin, "The difficulty of German information booklets on psoriasis and psoriatic arthritis: Automated readability and vocabulary analysis," *JMIR Dermatol.*, vol. 3, no. 1, Feb. 2020, Art. no. e16095, doi: 10.2196/16095.
[44] *Health on the Net Foundation*. Accessed: Jan. 18, 2021. [Online]. Available: https://www.hon.ch/en/search.html
[45] *Australian Government Department of Health*. Accessed: Jan. 18, 2021. [Online]. Available: https://www.health.gov.au/
[46] *Government of Western Australia Department of Health*. Accessed: Jan. 18, 2021. [Online]. Available: https://healthywa.wa.gov.au/
[47] *NSW Health*. Accessed: Jan. 18, 2021. [Online]. Available: https://www.health.nsw.gov.au/Pages/default.a

YANMENG LIU received the M.A. degree from Xi'an Jiaotong University, Xi'an, China. She is currently pursuing the Ph.D. degree with The University of Sydney, Australia.

In the past five years, she has successively contributed to major research grants funded by national research councils in Australia and China. Her Ph.D. was supported by The University of Sydney (USYD) China Studies Centre Graduate Fast Grant, USYD FASS Research Bursary Scholarship, USYD Raymond Hsu Scholarship, and AR Davis Postgraduate Research Scholarship. She has authored or coauthored 12 articles in international journals and conferences and translated two books. Her research interests include empirical language studies, data-driven multilingual corpus analyses, language quality evaluation, and machine learning in language studies.

MENG JI received the M.A. degree from University College London and the Ph.D. degree from Imperial College London.

She specializes in empirical translation studies, especially data-driven multilingual corpus analyses. She has published on environmental translation, healthcare translation, statistical translation stylistics/authorship attribution, and international multilingual education (statistical translation quality evaluation). She is the author/editor of more than two dozen research books (with Cambridge University Press, Oxford University Press, Routledge, Palgrave, Springer, John Benjamins, Waseda University Press in Tokyo, and University of Montréal Press), the editor of two special journal issues published by the MIT Press (Leonardo), USA, and the University of Montreal Press in Canada (Meta: Journal des traducteurs), and more than 50 journal articles and book chapters on empirical translation studies. Her research has been supported by the British Academy, the Japanese Society for the Promotion of Sciences, the Australian Research Council, Economic and Social Research Council of the U.K., Toshiba International Foundation, Worldwide University Networks Research Development Fund, and a number of leading universities in Europe, North America, Japan, South Korea, and Brazil. She is a qualified professional translator between English, Spanish, and Chinese having previously worked for major international organizations before teaching at universities.

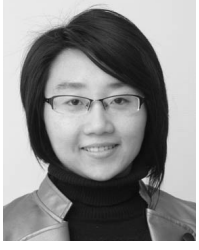
Dr. Ji is the Editor of *The Oxford Handbook of Translation and Social Practices*, (New York: Oxford University Press (with Sara Laviosa), 2020) and *Advances in Empirical Translation Studies* (Cambridge University Press (with Michael Oakes), 2019). She is a Guest Special Section Editor of *Leonardo: TransCreation: Creativity and Innovation in Translation* (Cambridge: The MIT Press, 2020). She is a Founding Series Editor of the *Cambridge Studies in Language Practices and Social Development* (Cambridge University Press). She is a Founding Editorial Board Member of the series of Cambridge Elements of Translation and Interpreting, Cambridge University Press. She is the Founding Editor of *Routledge Studies of Empirical Translation and Multilingual Communication* (New York/Oxon: Routledge).

SHANNON SHANSHAN LIN received the B.S. degree (Hons.) in nutrition, the Master of Nutrition Management degree, and the Graduate Certificate in Diabetes Management and Education. She is currently pursuing the Ph.D. degree with the School of Languages and Cultures, The University of Sydney, Australia.

She is an Accredited Practising Dietitian, Credentialled Diabetes Educator, and with a particular research interest in culturally and linguistically (CALD) and indigenous populations. She has been actively involved in the various committees both national and internationally, including the Australian Diabetes Educators Association, Diabetes Australia, International Diabetes Federation, and the Nursing Association of China. Her Ph.D. was supported by the Commonwealth Government Research Training Scholarship.

MENGDAN ZHAO was born in Hubei, China, in 1992. She received the B.A. degree in teaching Chinese as a second language, the B.S. degree in applied psychology from Southwest Jiaotong University, China, in 2014, and the M.A. degree in linguistics and applied linguistics from East China Normal University, China, in 2017. She is currently pursuing the Ph.D. degree with the School of Languages and Cultures, The University of Sydney, Australia.

From 2010 to 2020, she has authored or coauthored seven English/Chinese journal articles, contributed to major research grants funded by national research councils in Australia. She has presented at several international conferences in Australia, New Zealand, and Switzerland. Her research interests include public health translation and communication, and corpus research methodologies.



ZIQING LYV received the B.A. degree in English and the M.A. degree in English literature from Southeast University, Nanjing, China, in 2007 and 2010, respectively. She is currently pursuing the Ph.D. degree in Chinese studies with The University of Sydney, Sydney, NSW, Australia.

From 2010 to 2018, she was a Lecturer with the School of Foreign Languages, Jiangsu University of Science and Technology, Jiangsu, China. Her research interest includes big-data driven reception studies of Chinese science fiction, health translation, and evaluation.