Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

Employing Computational Linguistics to Improve Patient-Provider Secure Email Exchange: The ECLIPPSE Study

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Balyan, Renu	State University of New York at Old Westbury	
McNamara, Danielle S.	Arizona State University	
Crossley, Scott A.	Georgia State University	
Brown III, William	University of California	
Karter, Andrew J.	UCSF Division of General internal Medicine at	
Schillinger, Dean	UCSF Division of General internal Medicine at	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) 2022

Check type of content being submitted and complete one of the following in the box below:

- ☐ If article: Name of journal, volume, and issue number if available
- ☐ If paper: Name of conference, date of conference, and place of conference
- ☑ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

Natural Language Processing in Healthcare,211-241,CRC Press

DOI or URL to published work (if available) https://www.taylorfrancis.com/chapters/edit/10.1201/978100313

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** National Center for Educat through **[Grant number]** Grants R305 to **Institution]** Arizona State U. The opinions expressed are those of the authors and do not represent views of the **[Office name]** Arizona State University or the U.S. Department of Education.

10 Employing Computational Linguistics to Improve Patient-Provider Secure Email Exchange: The ECLIPPSE Study

Renu Balyan
Department of Mathematics, Computer & Information Sciences, State University of New York at Old Westbury, New York, USA

Danielle S. McNamara
Department of Psychology, Arizona State University, Tempe, Arizona, USA

Scott A. Crossley
Department of Applied Linguistics/ESL, College of Arts and Sciences, Georgia State University, Atlanta, GA, USA

William Brown III
Center for AIDS Prevention Studies, University of California, San Francisco, San Francisco, CA, USA

UCSF Health Communications Research Program, Center for Vulnerable Populations, Zuckerberg San Francisco General Hospital and Trauma Center, San Francisco, California, USA

Andrew J. Karter
UCSF Division of General Internal Medicine at Zuckerberg San Francisco General Hospital and Trauma Center, San Francisco, California, USA

Division of Research, Kaiser Permanente Northern California, Oakland, California, USA

DOI: 10.1201/9781003138013-10

Dean Schillinger
UCSF Division of General Internal Medicine at Zuckerberg San Francisco General Hospital and Trauma Center, San Francisco, California, USA

Division of Research, Kaiser Permanente Northern California, Oakland, California, USA

UCSF Health Communications Research Program, Center for Vulnerable Populations, Zuckerberg San Francisco General Hospital and Trauma Center, San Francisco, California, USA

CONTENTS

10.1 INTRODUCTION

As per the Centers for Disease Control and Prevention (2020), an estimated 34.2 million people in the United States (10.5% of the U.S. population) in 2018 were suffering from diabetes mellitus (DM). Self-managing DM, like with most chronic conditions, can be complex and requires frequent communication in between visits with healthcare providers. "*Health literacy (HL) –defined as a patient's ability to obtain, process, comprehend, communicate and act on basic health information*" (Protection and Act 2010; Schillinger et al. 2017) – plays a critical role for DM patients. Limited HL results in poor health outcomes (Schillinger et al. 2002; Schillinger et al. 2004; Sarkar et al. 2010), and contributes to preventable suffering, excess healthcare costs, and rapid physical functions decline (Smith et al. 2015). Online patient portals (e.g., Kaiser Permanente Northern California: KPNC patient portal, kp.org) are widely being used to provide communication support to patients and providers via SMs. Patients accessing such portals exhibit better (a) medication adherence (Sarkar et al. 2014; Lyles et al. 2016), (b) healthcare utilization (Reed et al. 2013), and (c) glycemic (blood sugar) control (Reed et al. 2012; Harris et al. 2013). However, the effectiveness of such web-based communication measures can be influenced by the HL of a patient. The overarching goal of the ECLIPPSE project was to identify ways to harness such online communication to reduce health disparities related to HL.

10.2 THE ECLIPPSE PROJECT

The ECLIPPSE (Employing Computational Linguistics to Improve Patient-Provider Secure Emails exchange) is a National Library of Medicine (NLM) funded project that aimed to: a) develop and validate an automated LP by assessing the linguistic features of secure messages (SMs) generated by patients and sent to their primary care physicians (Balyan et al. 2019; Crossley et al. 2021; Schillinger et al. 2020); b) demonstrate the validity of a patients' LP by examining associations with measures of patients' HL, their reports of their providers' communication, and several diabetes outcomes (Balyan et al. 2019; Crossley et al. 2021; Schillinger et al. 2020); c) develop and validate an automated CP using the linguistic features of the SMs written by primary care physicians to their DM patients (Crossley et al. 2020); d) determine the prevalence of LP-CP discordance (if any exists) across physician-patient dyads and explore the outcomes; and e) create and evaluate an automated LP and CP-based patient portal prototype that provides providers feedback in real-time to reduce the their SMs complexity and better accommodate patients' HL while they are writing SMs.

Two recent studies carried out as part of the ECLIPPSE project assessed the NLP tools' capability to classify HL of a patient (Balyan et al. 2019; Crossley et al. 2021). Both the studies used NLP tools on the corpus created by combining SMs sent by patients to their clinicians for developing patients' self-reported HL (obtained via survey) and expert- rated HL ML models. Their models reported results similar to previous HL research: patients classified by the model as having limited HL had higher probability of being older, belonging to a minority group, and

attained limited education. In addition, these patients had higher rates of hypoglycemia (an adverse drug event), comorbidities, healthcare utilization, worse medication adherence, and poor blood sugar control. Crossley et al. (2020) additionally leveraged expert ratings of the SMs to examine the linguistic complexity of the SMs sent by physicians to their patients so as to develop a physician CP. A depiction of the workflow for the overall process is shown in Figure 10.1.

10.3 IMPORTANCE OF CHARACTERIZING PATIENT HEALTH LITERACY

HL is representative of procactive, interactive and effective set of skills and communication behaviors (Sudore et al. 2009) between patients and clinicians including skills related to literacy, verbal (speaking and listening), numeracy, and health and healthcare-related digital skills. Research shows that poor communication is an important arbitrator in health outcomes and limited HL relationships, and that improving communication can mitigate health disparities associated with HL (Schillinger et al. 2004; Schillinger et al. 2009).

In the United States, older populations and socio-economically disadvantaged, including those of low income or education, some minority groups, and English-speaking immigrants are more likely to have limited HL (Kirsch et al. 2002; Institute of Medicine 2004; Smith et al. 2015). Limited HL has been found to be associated with poor health and mortality across populations and medical conditions (Sudore et al. 2006). In the diabetes context, limited HL is associated with a higher prevalence of DM, worse glycemic control, severe hypoglycemia (Sarkar et al. 2010), worse medication adherence (Karter et al. 2009), and higher rates of complications (Schillinger et al. 2002). Therefore, limited HL represents a costly and critical public and clinical health problem that is potentially remediable (Institute of Medicine 2004; Bailey et al. 2014).

One solution to this problem is to increase the quantity and quality of patient-physician electronic communications. Patients accessing such measures are more likely to adhere to prescribed regimens, achieve better outcomes, and have favorable patterns for healthcare utilization (Zhou et al. 2010). Technology platforms when developed in collaboration with and for the limited HL patients, and are accessible to them can indeed disproportionately improve health outcomes and support those patients that have the greatest communication needs (Schillinger 2007; Schillinger et al. 2008).

Patient portals as a source of communication via SMs are gaining greater importance, and enabling asynchronous and between-visit communications. Therefore, patients must achieve some level of digital communicative HL skills to take advantage of these portals. SM exchange is more relevant for chronic illnesses such as DM patients, because of their frequent inter-visit needs for communication. However, patients having limited HL may find it difficult to message their provider or understand their instructions or responses (Sarkar et al. 2010). Therefore, a patients' ability to effectively process (understand and act on) and write SMs is an inherent skill set to patient HL.

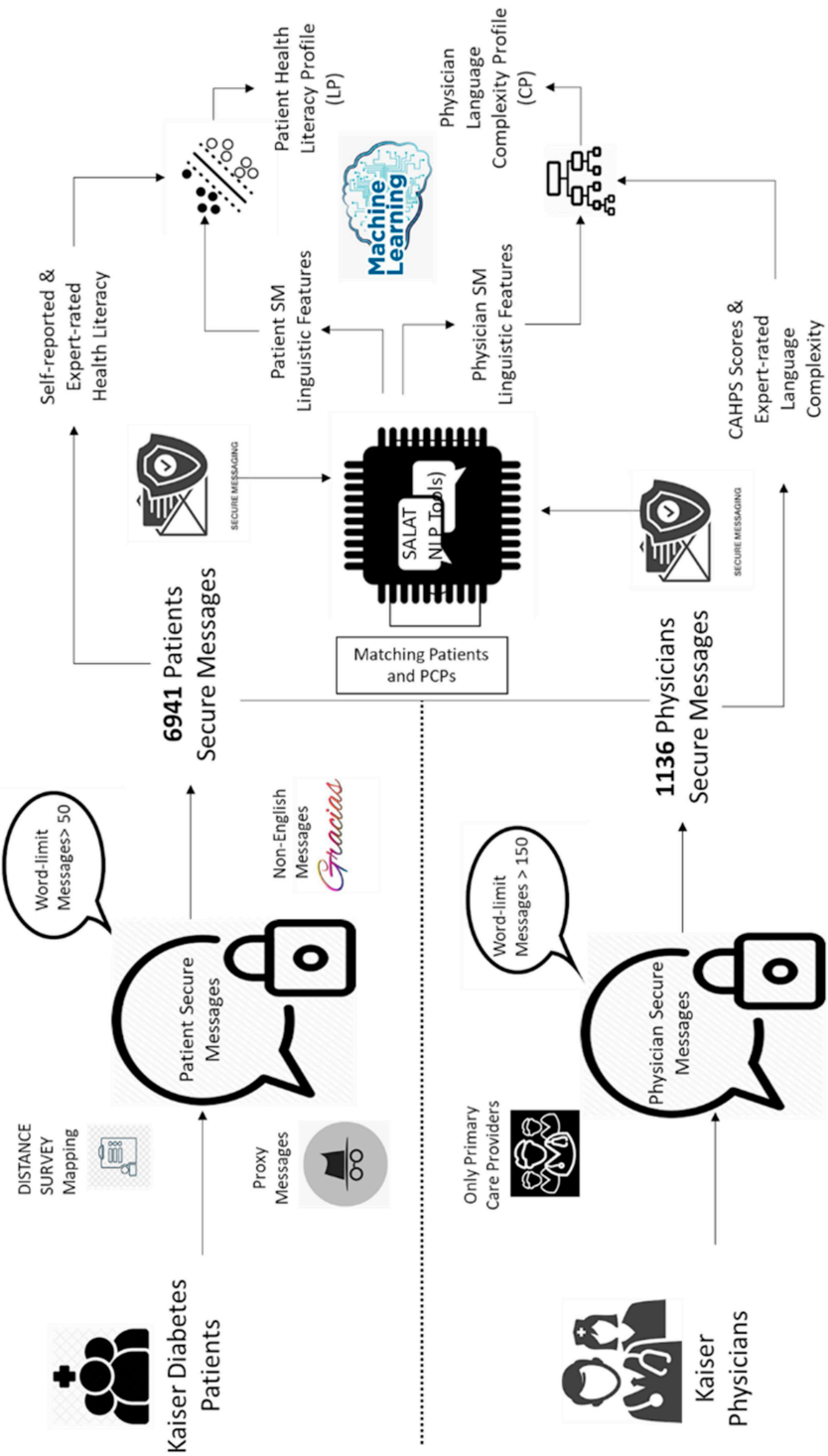

FIGURE 10.1 ECLIPPSE project overall workflow.

Current measurement tools for HL present numerous challenges for administrating and scalability because they are time-consuming and need in-person administration. A more efficient approach to identify limited HL using NLP and ML techniques could improve quality, inform care management initiatives (DeWalt et al. 2011; Karter et al. 2015), reduce disparities related to communication (Seligman et al. 2005), and target and tailor strategies related to population management (Brach et al. 2012).

10.4 IMPORTANCE OF CHARACTERIZING PHYSICIAN WRITING COMPLEXITY

In parallel, physicians need to develop communication skills and adopt the behaviors so as to tailor their written communications to match the needs of their patients. Providers must engage with patients in a way that provides actionable and meaningful information, and promotes shared meaning by providing comprehensible support (Brach, Dreyer, and Schillinger 2014). Physicians rarely employ recommended communication strategies for patients with limited HL (Schillinger et al. 2003), and frequently use clinical jargon (Castro et al. 2007), which can be incomprehensible in particular to patients having limited HL. Most physicians are unaware of their patients' HL, yet appear to be responsive after receiving this information, and utilize more recommended strategies (Seligman et al. 2005). Furthermore, research has demonstrated that HL-related disparities may be diminished by reducing the health communications literacy demands directed towards patients (DeWalt et al. 2009; Schillinger et al. 2009; Baker et al. 2011).

Not much is known about the readability of physicians' SMs. Use of classic readability formulas for assessing medical and health-related text has frequently indicated that many adults find the texts difficult to comprehend as these are written at higher levels (Berland et al. 2001; Kusec, Brborovic, and Schillinger 2003; Boulos 2005; Kandula and Zeng-Treitler 2008; Walsh and Volsko 2008; Hill-Briggs, Schumann, and Dike, 2012; McAndie, Gilchrist, and Ahamat 2016; Kugar et al. 2017; Schumaier et al. 2018). Approximately 89 million people in the US have been estimated to not understand and process most of the available health-related materials (Kindig, Panzer, and Nielsen-Bohlman 2004). The American Medical Association and the National Institutes of Health suggest that patient-oriented health texts should be written at lower grade levels (6th–8th grades; Badarudeen and Sabharwal 2010; Kugar et al. 2017).

Several problems have been reported in terms of measuring the readability, predictability, and reliability of physicians' SMs, when using existing classic readability methods (Meade, Byrd, and Lee 1989; Friedman and Hoffman-Goetz 2006; Wang et al. 2013; Wu et al. 2013; Barton et al. 2016; Zheng and Yu 2018). There are no efficient tools available for assessing the complexity of SMs written by the physician(s) to their patient(s) (Grabeel et al. 2018; Crossley et al. 2020). Therefore, development of a robust and reliable measure for physicians' SM linguistic complexity when used with a patient HL measure, could enable identification and determination of physician-patient linguistic discordance, ascertain its proximal communication consequences, and clinical outcomes. Furthermore,

developing such a measure would assist and enable health systems in identifying physicians who might be in need of and benefit from additional communication-related training and support (Kim et al. 2007; Oliffe et al. 2019).

10.5 NLP APPROACHES: HEALTH LITERACY AND TEXT READABILITY

While NLP has long been used in the medical domain for numerous applications including semantic lexicon development, clinical narrative representation, and text quality assessment (Johnson 1999), only a few studies have examined the use of NLP in HL for investigating the readability of medical texts and use of NLP features to predict readability. To our knowledge, no NLP research has assessed physicians' oral or written communications to their patients.

10.5.1 NLP and Health Literacy

A recent HL research measures review has found around 200 HL measures, out of these 52% measures required paper and pencil, and 12% required more than 15 minutes to administer. None of these studies has assessed the SMs written by patients' to their physicians to measure a patients' communicative HL. In general literacy studies have shown that writing skill (i.e., linguistic production) is strongly correlated with reading skill (i.e., linguistic comprehension), thus providing a strong reason to harness patients' SMs for assessing communicative HL.

As a result, employing NLP tools could be one approach to identify patients' HL for assessing the writing skill of a patient. NLP approaches can efficiently analyze data in huge amounts that is tedious and time-intensive for humans to accomplish. An automated HL measure using SMs may help overcome obstacles and challenges (e.g., scaling for larger patient populations) reported previously while conducting interviews or questionnaires for measuring HL. An automated NLP-based HL classification measure would provide an efficient means to automatically identify at-risk patients having limited HL, and requiring interventions.

10.5.2 NLP and Text Readability

Most of the research assessing medical text readability has used readability formulas such as the Dale-Chall – created for students in grade 4 or above, the scores are based on number of unfamiliar words in the text (Chall and Dale 1995); Flesch-Kincaid Grade Level (FKGL) – uses sentence length and syllables in a word, weighing longer sentences more than words (Kincaid et al. 1975); Flesch Reading Ease (FRE) computes a score between 1 and 100 but weights longer words more as compared to long sentences (Flesch 1948); Simple Measure of Gobbledygook (SMOG) – used in healthcare for medical writing and considers the percentage of words having three or more than three syllables (McLaughlin 1969), Frequency of Gobbledygook – takes into consideration the percentage of long words and the sentence length (Gunning 1952), and Fry – used across a range of sectors and also uses sentence and syllables for computing the scores (Fry 1968). However,

performance, validity, and effectiveness of these formulas in medical and other domains (Wang et al. 2013; Wu et al. 2013; Zheng and Yu 2017) have been strongly questioned. One possible explanation for the ineffectiveness of these readability formulas is absence of a strong mapping between their linguistic features (i.e., word and sentence length) and the linguistic constructs that are predictive of reading comprehension (Bruce, Rubin, and Starr 1981; Davison and Kantor 1982; Rubin 1985; Bruce and Rubin 1988; Graesser, McNamara, and Kulikowich 2011; Smith 2012; Balyan, McCarthy, and McNamara 2018; McNamara et al. 2019; Balyan, McCarthy, and McNamara 2020). In response, studies by readability researchers across a variety of domains have investigated and successfully shown NLP features capability to predict readability for medical texts (Kim et al. 2007; Zeng–Treitler et al. 2012; Wu et al. 2013; Zheng and Yu 2018).

In the medical domain, researchers have augmented the basic features with advanced NLP measures including lexical sophistication/diversity (Gemoets et al. 2004), word familiarity (Zheng and Yu 2018), syntactic, semantic, and cohesion features (Gemoets et al. 2004; Kim et al. 2007; Zeng–Treitler et al. 2012), and part of speech (PoS) tags (Kim et al. 2007; Zeng–Treitler et al. 2012). These new readability models that use NLP features perform better than the classic readability formulas in the medical domain (Kim et al. 2007; Zeng–Treitler et al. 2012) as well as outside of the medical domain (Crossley, Greenfield, and McNamara 2008; Pitler and Nenkova 2008; De Clercq et al. 2014; Crossley et al. 2017).

No research has focused on physicians' CP. However, automatically assessing CP is an important component of understanding how physicians interact with patients and whether or not patients can understand and act on suggestions made by the physician. Adjusting messages written to patients is a cumbersome task, especially when existing readability formulas are not efficient and effective, there is no certainty about why medical texts are difficult to comprehend, and physician training is lacking. Even then, there is little research that has examined physicians' written communication readability for messages written to patients.

10.6 TOOLS FOR LINGUISTIC FEATURES EXTRACTION

A Suite of Automatic Linguistic Analysis Tools (SALAT[1]) was used to extract linguistic features from the patient-physician SMs that measure different aspects of language, including text level information, lexical sophistication, syntactic complexity, and text cohesion. These NLP tools in turn use several other resources including Stanford Parser (De Marneffe, MacCartney, and Manning 2006), Wordnet (Miller 1995), CELEX word frequency database (Baayen, Piepenbrock, and Gulikers 1996), a psycholinguistic database (Coltheart 1981), a corpus (BNC; BNC Consortium 2007), and medical corpora including HIMERA (Thompson et al. 2016), i2b2[2] (Uzuner, Luo, and Szolovits 2007; Uzuner et al. 2008; Uzuner 2009; Uzuner, Solti, and Cadag 2010).

The SALAT consists of tools including *TAACO* (Crossley, Kyle, and McNamara 2016) – incorporates indices for local and global text cohesion analysis at both word and sentence levels; *TAALES* (Kyle and Crossley 2015) – constitutes tools for automatically assessing lexical sophistication, measures of concreteness of words,

their familiarity and meaningfulness (Kyle, Crossley, and Berger 2018); *TAASSC* (Kyle 2016; Crossley, Kyle, and McNamara 2017) – measures clausal (31) and phrasal (132) syntactic complexity indices, 190 indices of syntax sophistication, and 14 indices from the syntactic complexity analyzer by Lu (Lu 2010); and *SEANCE* (Crossley et al. 2017) – a sentiment analysis tool using a PoS tagger, and a number of dictionaries for sentiment, cognition, and social positioning, along with a negation feature. In addition to the SALAT, a tool for writing assessment (*WAT*; Crossley, Roscoe, and McNamara 2013) including indices specific to writing (text structure, cohesion, lexical sophistication, syntactic complexity, and rhetorical features) was also used (McNamara, Crossley, and Roscoe 2013).

The tools discussed in this section including TAALES and TAACO have been used for essays as well as unstructured data sets including, but not limited to, first and second language speech samples, children's e-mails, forum posts, early childhood writing, and beginning level L2 writing. Therefore, we decided to use these tools for SMs exchanged between physicians and patients.

10.7 MACHINE LEARNING APPROACHES

Several supervised machine learning classification models were used while building the patient LPs and physician CPs. A brief description of the algorithms is provided in Table 10.1 (Hastie, Tibshirani, and Friedman 2009; James et al. 2013; Balyan, McCarthy, and McNamara 2017). The use of linguistic and semantic indices was motivated by theoretical models of literacy and text complexity.

Our objective is to develop models of literacy and complexity that are linked to theoretical models of discourse and capable of driving feedback. While neural methods are gaining more focus these days and are indeed powerful, such methods result in less interpretable (explainable) and oftentimes highly-resource intensive models. As such, the decision to not use neural methods was driven by both the theoretical and applied requirements of this project. Likewise, the continued use of linguistic and semantic features in NLP models is crucial to the advancement of this field, as well as our understanding of language.

10.8 DEVELOPING AND VALIDATING PATIENTS' LITERACY PROFILES

Current direct measures of HL make widespread classification of patient HL challenging because they are time demanding and can be invasive. Hitherto, limited HL patients identification has proven laborious and infeasible to scale (DeWalt et al. 2011). Therefore, the "Literacy Profiles" were developed and validated as patients' HL automated measures for facilitating an economical, non-invasive, and scalable characterization of HL.

10.8.1 Data Source and Participants

The study used data from the KPNC (a fully integrated health system with ~4.4 million patients) Diabetes Registry consisting more than a million SMs written

TABLE 10.1
Machine learning models' brief description

Method Type	ML Algorithm	Description
Single Methods	Naïve Bayes	Naïve Bayes is a probabilistic method based on the posterior probability of Bayes' theorem. It makes an assumption about the features of independence (McCallum and Nigam 1998).
	Decision Trees	The Decision Trees represents the learned function as a set of if-then rules. The feature for a node is determined by information gain or gini index statistical properties (Mitchell 1997).
	Artificial Neural Networks	Artificial Neural Networks (ANN) model the human nervous system processing capabilities for information. These models use algorithms such as back-propagation for updating weights based on the feedback and are self-learning (Zhang 2000; Rojas 2013).
	Linear Discriminant Analysis	Linear Discriminant Analysis (LDA) is a "supervised" linear transformation technique that computes "linear discriminants" representing the axes that maximize the separation between multiple classes and is commonly used for reducing data dimensions (Fisher 1936).
	Support Vector Machine	Support Vector Machine (SVM) classify data into different classes by constructing a hyperplane. SVMs are among the best supervised learning algorithms that are efficient for high-dimensional data (Dumais et al. 1998; Joachims 1998).
Ensemble methods	Random Forests	Random Forests overcome the decision trees "overfitting" problem and construct multiple decision trees during the training phase implementing majority voting for classification (Smola and Schölkopf 1998; Schapire and Singer 2000).
	Bagging	Bootstrap Aggregation (Bagging) considers multiple classifiers, is a meta-algorithm creating training set bootstrap samples (sampling with replacement), and uses majority voting for classification (Breiman 1996).
	Boosting	Boosting like Bagging is a meta-algorithm. It iteratively trains multiple weak learners to build an ensemble and uses misclassified instances from the previous models. The final result is based on weighted sum of all classifier results (Krogh and Vedelsby 1995).
	Stacking	Stacked generalization (Stacking) creates an ensemble by combining multiple classifiers generated on a single data set. It creates a set of base classifiers and combines their outputs to train a meta-level classifier (Wolpert 1992).

by >150, 000 diabetes patients and >9,000 primary care physicians. The patients from KPNC registry who participated and completed a 2005–2006 DISTANCE survey (Diabetes Study of Northern California) and responded to the items related to self-reported HL (N = 14,357; Chew et al. 2008; Moffet et al. 2009; Ratanawongsa et al. 2013) were used. The survey details can be seen in Figure 10.2. Further details of the DISTANCE Study have been reported previously (Moffet et al. 2009). The data was collected using questionnaires completed online, via paper and pencil or on the telephone with a response rate of 62%. DISTANCE surveyed diabetes patients with average age 56.8 years (±10); and who were male (54.3%); Latino (18.4%), African American (16.9%), Caucasian (22.8%), Filipino (11.9), Asian (Chinese/Japanese; 11.4%), South Asian/ Native American/Pacific Islander/ /Eskimo (7.5%), and multi-racial (11.0%).

The original corpus included all (N = 1,050,577) diabetes patients-clinicians exchanged SMs from KPNC's patient portal from January 1, 2006 through December 31, 2015. However, SMs that patients exchanged with only their primary care physicians (PCPs) were included in the analyses described in the study. The SMs were filtered for the patients whose DISTANCE survey data was missing. In addition, non-English SMs (a less than 1% messages were removed) and those SMs not written by the patient but the proxy caregivers (determined by a proxy check-box in the KP.org or validated by an NLP algorithm; Semere et al. 2019) were also excluded. The final ECLIPPSE dataset for the patient HL measure constitutes 283,216 SMs written by 6,941 patients to their PCPs. All of the SMs for each patient were aggregated into a single file. The patients having less than 50 words in the aggregated SMs file were excluded. Previous research in NLP text in learning analytics was the determining factor for the 50-word exclusion threshold (Crossley et al. 2016; Crossley and Kostyuk 2017). All of the survey and SM data were confidential and stored and analyzed on secure servers behind KPNC firewalls that precluded copying and downloading, and maintained data security. Therefore, we are not in a position to provide any examples because the data used for this study contains protected health information (PHI) and access is protected by the KPNC Institutional Review Board (IRB).

10.8.2 Health Literacy Measures

Two LP measures based on NLP and ML were developed. In the first LP, a validated HL scale that contained three items measuring self-efficacy in specific HL competencies from the DISTANCE survey was included. The items were self-reporetd by patients using a 5-point Likert scale (1: "Always" and 5: "Never"; Sarkar et al. 2011). The items measured patients' confidence in filling medical forms, and patients' understanding of written medical information, and frequency of help needed while reading health documents. The new self-reported HL variable used in the study was created by computing average scores across these HL items. These scores were dichotomized to indicate limited and adequate HL (Balyan et al. 2019). The threshold considered for dichotomization was consistent with scale used in previously employed studies (Chew et al. 2008; Sarkar et al. 2008; Sarkar et al. 2011).

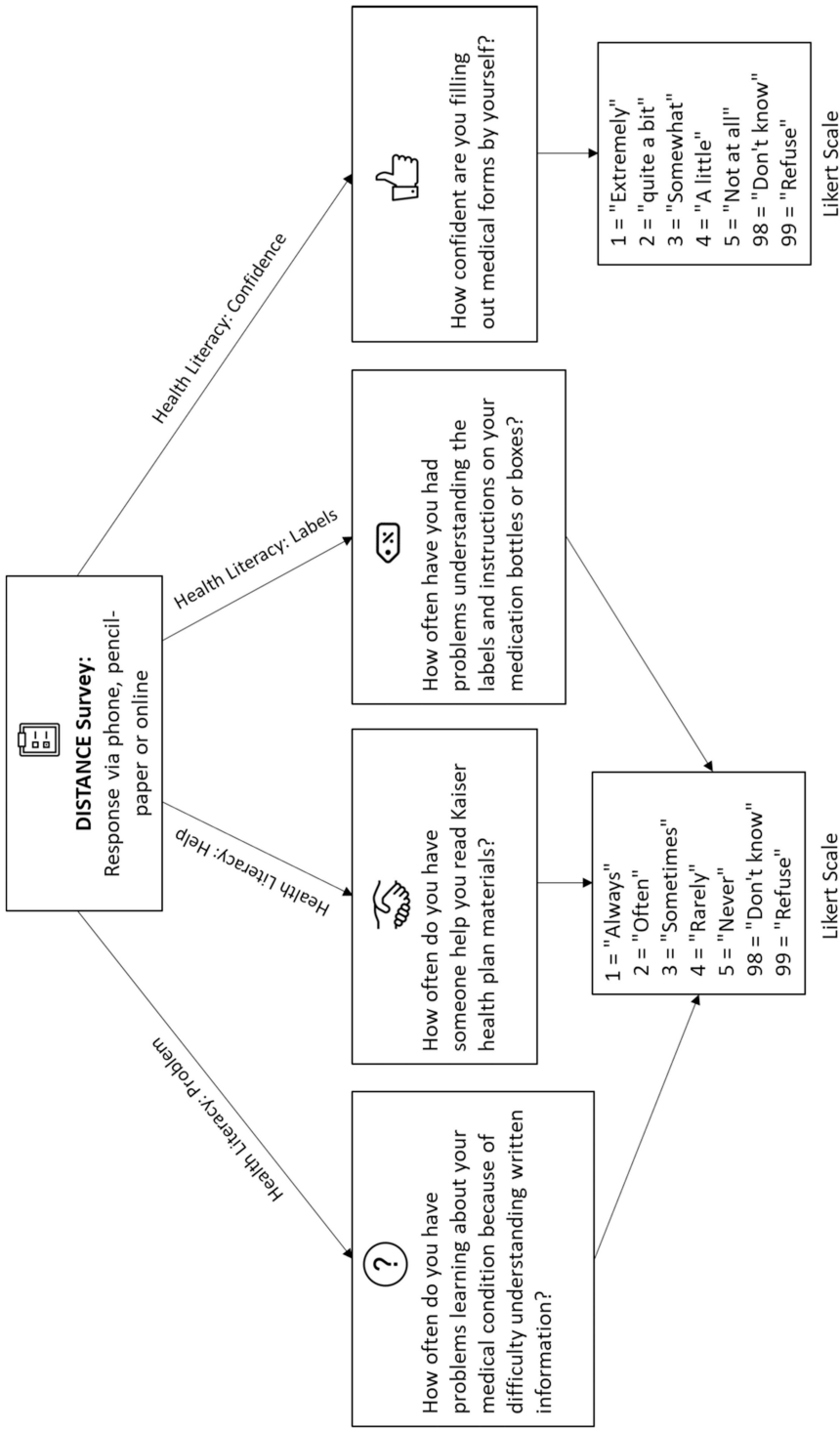

FIGURE 10.2 The DISTANCE survey health literacy-related items.

Because of concerns related to possible mismeasurement using self-report, a second LP was developed, in which HL was measured based on the quality of patients' SMs, using ratings by experts. These ratings used a subset of aggregated SMs written by 512 patients from the DISTANCE survey (Crossley et al. 2021). A scoring rubric to assess a patients' perceived HL based on their SMs was adapted from a rubric used for testing high school seniors writing abilities (Crossley, Kyle, and McNamara 2015). The patients' written English proficiency, organization, health vocabulary accuracy, and patients' intent towards their physician were assessed using a scale ranging from 1: low HL to 6: high HL (Crossley et al. 2021).

10.8.3 Creating Literacy Profiles

Using NLP and ML techniques, five separate LP prototypes (Schillinger et al. 2020), each varying in their sophistication and linguistic features, were developed for classifying patients' HL (self-reported; Balyan et al. 2019, and expert-rated; Crossley et al. 2021). Table 10.2 summarizes the LPs and the linguistic indices used while creating each LP.

10.8.4 Evaluating Literacy Profiles

Associations between these LPs' classifications of HL and patients' education, race/ethnicity, and age were then explored. Because associations are also known to exist between HL and communication between the patient and the provider (Schillinger et al. 2003, 2004; Castro et al. 2007; Sarkar et al. 2008), relationships between the LPs and patients' reports of physician communication from the Consumer Assessment of Healthcare Provider and Systems (CAHPS) survey (Schillinger et al. 2004), reported via the DISTANCE survey, were also examined. Associations between the LPs and diabetes-related outcomes, including adherence to medications based on continuous medication gaps (CMG; Steiner et al. 1988; Steiner and Prochazka 1997), hemoglobin A1c (HbA1c), hypoglycemia (Sarkar et al. 2010), and comorbidities were also examined. Finally, relations between each LP and healthcare utilization (outpatient, emergency room, and hospitalization utilization) were explored.

The findings are summarized in Table 10.3. The results in Table 10.3 are not a result of correlations but associations between the predictions made by the five literacy profiles and the patient health data available in their EHRs. Since the focus of the chapter was to show that NLP is capable of predicting the health literacy of a patient based on the SMs written by the patient and the LP followed the patterns similar to prior research in health domain, we did not include the health outcomes of the patients but only showed which LP was able to significantly predict or followed patterns known in the health domain. These results have been previously published (Schillinger et al. 2020).

In addition, the impact of race and ethnicity on the accuracy of models is indeed an extremely important topic and concern. We have also demonstrated that the LP models were not impacted by race (Schillinger et al. 2021).

TABLE 10.2

Linguistic indices and the literacy profiles* (Schillinger et al. 2020)

Literacy Profile	Linguistic Indices	Literacy Profiles Description
Flesch Kincaid (LP_FK)	*Readability:* length of words and sentences	Flesch-Kincaid, being the simplest, most widely available, and most commonly used readability formulas in the medical domain, was used as a baseline measure (Paasche-Orlow, Taylor and Brancati 2003; Wilson 2009; Piñero-López et al. 2016; Jindal and MacDermid 2017; Munsour et al. 2017; Zheng and Yu 2018).
Lexical Diversity (LP_LD)	*Lexical Diversity*: D-based words variety	Lexical diversity (LD), another commonly used method in the linguistics domain for assessing writing proficiency (Malvern et al. 2004; McCarthy 2005) captures cohesion and lexical richness of text. Both these features are consistent predictors of writing quality and text sophistication (McNamara et al. 2014; McNamara et al. 2015). As a result, lexical diversity model in addition to the Flesch-Kincaid model was used as a second baseline.
Writing Quality (LP_WQ)	*Word Frequency*: reference corpus words frequency *Syntactic Complexity*: words count before the main verb within a sentence *Lexical Diversity*: words variety based on MTLD.	A model (McNamara, Crossley, and McCarthy 2010) previously validated, derived from word frequency and syntactic complexity in addition to the LD (McCarthy 2005; McNamara et al. 2014) was used to classify patients' SMs based on the writing quality as low or high HL. The model was developed to test the significance of adding linguistic features on top of the existing LD feature.
Self-Reported (LP_SR)	*Concreteness*: word concreteness degree *Lexical diversity*: MTLD and D-based words variety. *Present tense*: incidence *Determiners*: incidence *Adjectives*: incidence *Function words*: incidence	The self-reported HL profile was created using 185 linguistic features extracted from the patients' written SMs. Some of the key features used in this LP are explained in this table. The motivation, rationale, development, and experimental details for LP_SR have been reported previously (Balyan et al. 2019).
Expert-Rated (LP_Exp)	*Age of Exposure*: estimated age of a word appears in a child's vocabulary *Lexical decision response time*: time taken by humans in judging characters *Attested lemmas*: count per verb argument construction *Determiner per nominal phrase*: determiner count in a noun phrase (NP) *Dependents per nominal subject*: subject structural dependent count in an NP *Number of associations*: with each word	The expert-rated LP was developed using eight linguistic indices to predict expert human ratings of HL. The LP was developed using a sampled subset of 512 expert-rated SMs. Additional details related to the LP_Exp development and experimental design have been previously reported (Crossley et al. 2021).

TABLE 10.3

Results for health literacy correlates vs literacy profiles

Literacy Profile / Health Literacy Correlate		FK	LD	WQ	SR	Exp	Summary
Sociodemographic	Race	√	√	√	√	√	All the LPs classified limited HL to be associated with non-white race. Three LPs (LP_LD, LP_SR, and LP_Exp) were associated with lower education, with LP_SR and LP_Exp observing the strongest effects. Only one LP (LP_SR) was associated with older patient age.
	Education	X	√	X	√	√	
	Age	X	X	X	√	X	
Provider Communication		X	X	X	√	√	The patients predicted as having limited HL by two LPs (LP_SR and LP_Exp) were more likely to rate communications with their health care providers as "poor" based on the CAHPS item. The findings for LP_SR were found to be somewhat more robust.
Health Outcomes	CMG	√	X	X	√	√	Limited HL classified by LP_FK, LP_SR, and LP_Exp was associated with poor medication adherence, greater comorbidity, and serious hypoglycemia. Poor medication adherence, and optimal and poor diabetes control was most significantly associated with LP_FK and LP_Exp.
	Hypoglycemia	√	X	X	√	√	
	Diabetes control	√	X	X	X	√	
Healthcare utilization	Outpatient visits	X	X	X	√	X	The only model capable of associating limited HL with higher rates of hospitalizations and outpatient visits was LP_SR. Both LP_SR and LP_Exp observed higher annual emergency room utilization rates for limited HL, but the differences related to HL were more robust for LP_SR.
	Hospitalization	X	X	X	√	X	
	Emergency room utilization	X	X	X	√	√	

Notes: FK: Flesch Kincaid; LD: Lexical Diversity; WQ: Writing Quality; SR: Self-Reported; Exp: Expert-Rated.

10.9 DEVELOPING AND VALIDATING PHYSICIAN COMPLEXITY PROFILES

Diabetes patients with ongoing self-management needs require counseling and guidance, and often use SMs to discuss issues such as test results, medication concerns, emerging symptoms, requests for appointment and referral, and responses to treatment (Friedman and Hoffman-Goetz 2006). Physician, in turn, must respond to these concerns. Insofar as shared understanding is a critical objective of communication exchange, identifying the linguistic features that make physicians' SMs more or less difficult to understand needs to be pursued. Hence, physicians' SMs to their patients were harnessed as a means to develop an automated readability formula based on advanced linguistic features associated with text complexity of the SMs, which was called the physician CP. To do so, linguistic features and the expert ratings of physician text complexity relationships were examined. Here again, as a reference, the newly developed text complexity profile performance was compared to the Flesch-Kincaid (Kincaid et al. 1975).

10.9.1 Data Source and Participants

The data used in this study was derived from SMs drawn from the ECLIPPSE Study (already discussed in section 10.8.1) and exchanged between January 1, 2006 – December 31, 2015. A total of 1,136 primary care physicians were found to have sent SMs to patients. Based on prior research (Crossley 2018), the SMs that included at least 150 words were aggregated to create a single file for each physician. A total of 2.7% physicians from the overall set were filtered out because their aggregated SMs were less than 150 words. A sufficient number of words are necessary to have a stable measure of text complexity and hence the decision to use 150-word threshold. Future work may indeed lead to algorithms that require fewer words; however, using a minimum number of words follows standard practice in the readability literature. To be consistent and to be able to examine concordance between the patient HL and the physician CP in the future, only those SMs written by physicians to patients were included, whose SMs were used in an earlier analysis (Crossley et al. 2020). It was observed that while using the electronic health record portal some physicians made elaborate use of "automated text" that is available to them while they draft their responses. These automatically generated texts in the SMs were retained for linguistic analyses because it proved difficult to create an automatic and generic NLP/ML model to reliably exclude these automated text segments, and because these texts represented the language used by the physicians when replying to their patients.

10.9.2 Expert Rating

The readability of the physicians' SMs sent to the patients were rated by two experts based on the difficulty faced by "struggling readers" in processing and comprehending the SMs for (Protection and Act 2010; Schillinger et al. 2017). Both of the raters were experts in medical discourse, had taught classes in literacy, and had

experience evaluating/rating medical texts. The raters were instructed to use a 5-point Likert scale scoring rubric. The raters were required to judge "the ease with which a struggling reader could understand the message?" The 5-point Likert scale ranged from: 1-"very easy" to 5-"very difficult". This approach used was similar to Kandula and Zeng-Treitler's (2008) 7-point scale rating.

10.9.3 Creating Physician Complexity Profile: MoTeR-P

The Model of Text Readability in Physicians (MoTeR-P) was developed from the SMs written by the physicians to their patients and the expert scores. NLP tools similar to those discussed in Section 10.6 were employed to compute each SM linguistic features. The linguistic features used to develop the CP were similar to text complexity measures that had been previously validated. TAALES and SÉANCE were used to compute the features related to lexical sophistication and semantics. Syntactic complexity-related features were calculated using TAASSC (Kyle 2016), and Coh-Metrix (Graesser et al. 2004). TAACO was used for text cohesion features. Finally, Coh-Metrix was used to return the traditional readability scores (FKGL and FRE).

SMs scored by experts in the range of 1–3 were classified as "low CP" (i.e., easy to understand), while those scored 4–5 were classified as "high CP" (i.e., difficult). The linguistic features were checked and pruned for non-normality based on skewness and kurtosis values (George 2011), effect size with the dependent variable, and multicollinearity prior to the statistical analyses. The remaining 85 features were used to develop MoTeR-P for predicting text complexity of expert ratings, by using these features as predictors in training a linear discriminant analysis (LDA) model. To account for the unbalanced data (more texts scored rated as 3 i.e., neither difficult nor easy), the LDA model probability weights were adjusted such that instead of using the default 0.50 probability, any weights greater than 0.55 were classified as "easy". To control for overfitting, a threshold of one predictor per 20 instances (i.e., 24 predictors) was used to train the LDA model. For comparison, another separate LDA model was trained using the FKGL for predicting the expert ratings of text complexity.

10.9.4 Evaluating Physician Complexity Profile

The MoTeR-P used a set of 24 linguistic features for training (see Table 10.4) and achieved an accuracy of 0.749, sensitivity and specificity of 0.674 and 0.788 respectively for the test data, $X^2 = 50.977$, $p < 0.001$, and 0.455 Cohen's Kappa indicating a moderate agreement. It was observed that the SMs predicted as difficult contained words that were less familiar, more abstract, high age of acquisition, occurred in fewer texts, more frequent academic function words, fewer actions and objects words, and were more diverse (lexically). Conversely, the SMs containing more frequent tri-grams that were also present in more texts were predicted as more readable. Syntactically, SMs having fewer syntactic overlap and more dependencies across sentences were classified as harder to understand. In addition, more difficult texts contained greater number of function words, across paragraphs verb overlaps

TABLE 10.4
Descriptive statistics of the features used in MoTeR-P (Crossley et al. 2020)

Feature	Mean (SD): Difficult	Mean (SD): Easy	Feature	Mean (SD): Difficult	Mean (SD): Easy
Word range score, CW, SUBTLEXus	3.196 (0.172)	3.303 (0.123)	Average direct object dependencies	1.253 (0.368)	1.148 (0.355)
Word age of acquisition scores, AW, Kuperman	5.496 (0.39)	5.284 (0.26)	Friends and family related words	0.191 (0.081)	0.217 (0.095)
Lexical diversity (D)	89.827 (23.659)	77.259 (23.543)	Argument overlap (binary), paragraphs	0.579 (0.197)	0.524 (0.203)
FW overlap, paragraphs	8.225 (4.016)	6.379 (3.936)	Words related to joy	0.617 (0.501)	0.788 (0.68)
FW frequency, COCA academic	16831.399 (3494.824)	15496.32 (3257.931)	Argument overlap, sentences	0.219 (0.078)	0.241 (0.09)
Lexical diversity (Maas)	0.021 (0.003)	0.023 (0.004)	Trigram range, COCA fiction	0.028 (0.008)	0.026 (0.007)
Bigram association strength (delta p), COCA academic	0.045 (0.012)	0.041 (0.01)	Trigram frequency, COCA newspaper	0.429 (0.118)	0.401 (0.111)
Terms related to action	0.609 (0.133)	0.655 (0.142)	Pronoun overlap, paragraphs	1.354 (0.66)	1.199 (0.648)
Syntactic similarity score	0.104 (0.031)	0.116 (0.039)	Construction frequency, SD, COCA fiction	650503.988 (135651.74)	620837.079 (139117.53)
Word familiarity, AW, MRC	593.239 (4.894)	594.622 (3.825)	Word concreteness, AW, MRC	2.642 (0.136)	2.669 (0.126)
Fear and disgust words	0.118 (0.067)	0.097 (0.07)	Incidence of words related to objects	0.134 (0.053)	0.145 (0.061)
Verb overlap (binary), paragraphs	0.764 (0.257)	0.674 (0.311)	Argument overlap, paragraphs	0.383 (0.159)	0.413 (0.164)

AW: all words; CW: content words; FW: function words; SD: standard deviation.

and fewer semantic overlaps, when cohesion was considered. From sentiment perspective, messages containing more words expressing fear and disgust, and fewer words associated with joy, friends, and family were classified as more difficult to understand.

The MoTeR-P model's performance, when compared with the FKGL model, was found to be superior. The FKGL-LDA model for the same dataset reported an accuracy of 0.65, X^2 6.56, p = 0.01, and demonstrated weak agreement with 0.154 Cohen's Kappa. Sensitivity and specificity for the FKGL model were found to be 0.302 and 0.838 respectively, which were much lower than those of the MoTeR-P.

10.10 CHALLENGES AND SOLUTIONS

Like any other project, the ECLIPPSE project team during implementation encountered and overcame several challenges. This section describes the challenges encountered (while employing NLP/ML techniques for developing/validating the LPs and CPs already discussed in the previous sections of this chapter), and the solutions devised to address these challenges (see also Brown et al. 2021). The challenges and the solutions are broadly categorized into a) data mining-related issues, b) linguistic indices analyses-based problems, and c) interdisciplinary collaboration-related complications.

10.10.1 SMs Corpus Mining

The first challenge was related to data mining of the SMs. The SMs for both the patient and physician in isolation and the patient-physician interactions had to be extracted. The data had to be extracted from multiple locations and mapped according to their identification numbers (IDs). For this, the patients' unique IDs referred to as the medical record numbers (MRNs) from their EHRs had to be matched to their IDs assigned by the KP patient portal. In turn, the message IDs of the KP patient portal had to be mapped to their message IDs in the EHR data to extract the SMs from their EHR notes. Subsequent to the data extraction, there were challenges related to missing or incorrect paragraph separators, sentence boundary markers, and punctuations (often referred to as structural markers). The absence of such markers influenced computation of several linguistic features across paragraphs, syntactic indices, potentially leading to imprecise computations and incorrect parsing.

The problems related to parsing were further compounded by the presence of some unstructured and ungrammatical contents in the SMs including test reports from labs, website links, automated signatures, office addresses along with their office hours, etc. Another issue that necessitated additional data security measures was the presence of patient/physician information such as their names and phone numbers in some SMs. It was not feasible to correctly de-identify all the data using automatic measures, therefore the data needed to be stored and could only be analyzed on secure servers using KPNC firewalls. While storing the data and analyzing it on the secure KPNC server did solve the issue related to the confidentiality of data and security, but this also resulted in challenges related to server accessibility, and delays in data processing.

In addition, physicians' SMs often included "smart texts" or "smart phrases". These texts/phrases refer to pre-defined automated content made available to the physicians by the online KP portal while the physicians respond to their patients. As described in section 10.9.1, these automated smart texts were not removed from the corpus because of large number of variations in the smart texts and phrases, it was infeasible to create an NLP algorithm that could generalize well and automatically identify all these varying texts/phrases with high accuracy. Another reason for retaining these automated texts was that these texts represented the language that was actually used by physicians while writing to their patients and would form an important and integral part of subsequent linguistic analyses.

Beyond the data mining problems, there were also issues on the patient side regarding who was the author of the SMs being written to the physician. It was observed sometimes that the SMs instead of being written by the patient were written by the patient proxies. The "ProxyID" algorithm was developed and used to identify hidden proxy messages (Semere et al. 2019). In addition, some SMs contained text that was in language other than English. Due to limited availability of NLP tools for non-English texts, scripts were created to identify such texts that successfully removed non-English (Spanish, in specific) texts if more than 50% of the text was in a language other than English. Due to this threshold for removing texts, some residual non-English text may have been left in the SMs.

10.10.2 Linguistic Indices Analyses

It is not always feasible to perform robust linguistic analysis if the SMs to be analyzed are short. We encountered some SMs in the corpus that were too short, leading to structural challenges in analyzing the data. In order to address this issue of content in an SM being insufficient to be linguistically analyzed correctly, a minimum word requirement was applied to the SMs. Those SMs written by patient containing fewer than 50 words were removed from analysis, and only SMs with more than 150 words written by physicians were considered for future analysis.

Selecting best features from a set of linguistic indices for training the machine learning models is always a challenge. We needed to select a set of linguistic indices for the LP and CP algorithms from a set of hundreds returned by different linguistic tools. As a result, commonly used filter methods such as multicollinearity, zero, and nearly zero variance and non-normal distributions were used to filter out some linguistic features before the model was trained. The model was trained using the indices obtained after filtering to identify some of the topmost important indices within the trained model to further reduce the linguistic indices. Imbalanced samples in HL estimations were another matter of concern while developing the LP models. Some traditional ML algorithms are not designed to handle skewed or imbalanced data and hence do not perform well with such data. Several methods (e.g., under-sampling, oversampling, or SMOTE) were examined to account for the imbalance in the datasets and different measures (such as updating thresholds, refining expert ratings) were explored to handle such data. The computational processes and the validity for these measures are detailed in papers

describing the LP (Balyan et al. 2019; Crossley et al. 2021; Schillinger et al. 2020) and CP (Crossley et al. 2020) development.

Another critical challenge faced by the team was assessing the LP and CP performance in the absence of "gold standards" for HL and linguistic complexity for the patient and physician respectively. One of the solutions included expert ratings of SMs for a small subset of the overall data. A challenge related to the expert ratings was the need to refine the scoring rubrics and reliably train the raters for assessing the HL and linguistic complexity of both the patient and the physician. The gold standard problem for the LP was overcome by applying two proxy measures: DISTANCE survey self-reported HL (Sarkar et al. 2010) and an expert-rated HL measure. As a result, two versions of the LPs were generated (LP-SR: Balyan et al. 2019; Schillinger et al. 2020; and LP-Exp: Crossley et al. 2021). For the physician CP, because of non-availability of a true gold standard, a measure was developed based on the expert ratings of the physician SM linguistic complexity (Crossley et al. 2020).

10.10.3 Interdisciplinary Collaboration

Experts working across several scientific disciplines and geographies encountered various challenges, such as similar terms across different domains (health services research, linguistics, and cognitive science) had different meanings, definitional differences, and lack of understanding related to tasks and methods at times resulted in confusion and inefficiency. A few more critical transdisciplinary challenges were related to methods transportability, research integrity or rigor, different interpretations of certain findings in terms of their real-world significance, and agreeing on priorities related to research and publications.

Challenges inherent to interdisciplinary collaboration were addressed by employing real-time and eventual clarification, proper documentation of ambiguous terms and tasks. Annual in-person meetings, biweekly video conferences, and regular email exchanges speed up decision making, resolve discrepancies related to terminologies, and ensured consistency and consensus-building. Providing background and context to align objectives and clarifying discipline-specific methodologies and terminologies also proved helpful. Even though more discussions were needed for some tasks, frequent delineation and re-visiting the grant aims helped mitigate tensions between different aspects of the project (theoretical vs. applied).

10.11 CONCLUSION

This chapter discussed the processes for development, evaluation, validation, and various challenges encountered and solutions devised for patients' LPs and physicians' CPs, the scientific products of the first two aims of the ECLIPPSE project. While a variety of healthcare research applications have employed NLP and ML approaches, the research discussed in this Chapter is the first of its kind to classify patients' HL and physician language complexity using the SMs exchanged between patients and physicians. Nonetheless, the studies have several limitations that should be noted.

First, in the context of patients' LP, only patients who wrote SMs were analyzed, which may have excluded patients with severe HL limitations. Second, the study was limited to SMs written in English only, excluding other patients with limited HL. Third, the LPs were modeled against HL that was self-reported by patients and rated by experts, due to non-availability of a comprehensive gold standard for HL; this may have limited the classification of HL. Fourth, this study analyzed data from a large and integrated healthcare system; hence, how well the models work in other systems and for patients other than DM2 patients needs to be assessed. Finally, adding patient demographics and clinical characteristics to the current linguistic features from SMs could increase the accuracy of our models.

Limitations of the study dealing with physicians' CP also exist. First, the current linguistic features did not include indicators of interactivity, shared meaning, and empathy, and hence may be insufficient in predicting communication-sensitive outcomes. Second, only linguistic features were used for the study; features beyond linguistics such as age, reading ability, background knowledge, and socio-cultural background were not incorporated. As such, future research needs to assess the extent to which physician CP is associated with patient comprehension, as directly measured. Third, only text content in the SMs was analyzed; other aspects, such as figures and charts were not considered. Finally, the CP work was limited to PCPs' SMs; other healthcare team members such as nurses, medical assistants, and sub-specialists were not considered.

We conclude that applying innovative NLP and ML approaches to generate a patient-physician LP and CP from their SMs is a feasible, and scalable strategy for identifying patients with limited HL, and to identify those physicians who write complex messages to their patients. This work can provide a tool that has potential to reduce disparities related to HL. Additionally, patients classified as limited HL can be provided feedback for promoting adherence (Sudore et al. 2009). It may prove useful to identify limited HL patients which may be further used to alert clinicians about their patients' potential difficulties in written and verbal instructions comprehension. Finally, identifying complex messages on the part of physicians, and providing them with feedback could enable these physicians to tailor their messages, making them more readable and easier to comprehend and act upon.

From our perspective, the NLP methods are an important contribution of this research. Indeed, the use of NLP combined with machine learning was the primary purpose of this project. The Literacy profiles derived using linguistic and semantic indices based on patients' written secure messages is an entirely novel contribution. Likewise, identifying linguistic complexity profiles based on the secure messages composed by the physicians to their patients is equally novel and a major contribution to the literature. The validity of these algorithms was assessed based on the health outcome patterns in the health domains. The work carried out in this project is innovative from an NLP application perspective. From the perspective of NLP, this is the first attempt (that we know of) to use NLP to develop literacy profiles of patients' health literacy or language complexity profiles based on physician written communication. Equally novel is the alignment of the LP and CP profiles with demographic and patient outcomes. The contribution of this chapter is that it is the

only documented work that provides an overview of the project, including both profiles and information on the work involving natural language processing.

10.12 FUTURE DIRECTIONS

Effective communication between a patient and a physician is innate to clinical practice, and important for delivering high-quality healthcare. In patient-physician communication, the *manner* in which information is communicated by a physician to a patient is as *important* as the information being communicated (Travaline, Ruchinskas, and D'Alonzo 2005). A large number of patient-physician relationships breakdown due to patients' dissatisfaction because many physicians overestimate their communication abilities (Ha and Longnecker 2010). While linguistic complexity matching ("concordance") is believed to promote shared understanding, no study has tested this hypothesis. We apply HL and linguistic complexity measures for patients' and physicians' respectively generated via computational linguistics to determine whether linguistic matching has clinical benefits, and for whom. We plan to classify physician-patient dyads as concordant or discordant, and explore whether concordance is associated with the patients' reports of physician communication (CAHPS), specifically using the items that ask patients about the extent to which their doctor explains things in ways that they can understand. We will further characterize each physicians' communication "style" (the predominant strategy they employ across patients), specifically exploring whether "universal precautions" (using simpler language with all patients) or "universal tailoring" (using language that matches their patients' HL) is associated with better understanding.

One of the final aims of the ECLIPPSE project is the development and evaluaton of an online, automated feedback prototype embedded in the patient portal. We plan to test the effects of this automated feedback tool with respect to its ability to reduce the physicians' SMs complexity in order to meet the low HL patients' needs. One of our objectives is to develop an interface for physicians where they can respond to messages received from patients. For SMs that are calculated to be overly complex relative to the patients' HL, physician participants will receive different versions of linguistically motivated feedback. This SM application, if successful and found to be acceptable to physicians, could have substantial benefits for their patients.

ACKNOWLEDGMENTS

This work has been supported by grants NLM R01 LM012355 from the National Institutes of Health, NIDDK Centers for Diabetes Translational Research (P30 DK092924), R01 DK065664, NICHD R01 HD46113, Institute of Education Sciences, U.S. Department of Education, through grant R305A180261 and Office of Naval Research grant (N00014–17-1–2300).

NOTES

1 linguisticanalysistools.org
2 https://www.i2b2.org/NLP/DataSets/Main.php

REFERENCES

Baayen, R. H., R. Piepenbrock, and L. Gulikers 1996. The CELEX lexical database (cd-rom).

Bailey, S. C., A. G.Brega, T. M.Crutchfield, T. Elasy, H. Herr, K. Kaphingst, … and D. Schillinger 2014. Update on health literacy and diabetes. *The Diabetes Educator* 40, no. 5: 581–604.

Badarudeen, S., and S. Sabharwal 2010. Assessing readability of patient education materials: Current role in orthopaedics. *Clinical Orthopaedics and Related Research®* 468, no. 10: 2572–2580.

Baker, D. W., D. A. DeWalt, D. Schillinger, V. Hawk, B. Ruo, K. Bibbins-Domingo, … and M. Pignone 2011. "Teach to goal": Theory and design principles of an intervention to improve heart failure self-management skills of patients with low health literacy. *Journal of Health Communication* 16, no. sup3: 73–88.

Balyan, R., S. A. Crossley, W. Brown III, A. J. Karter, D. S. McNamara, J. Y. Liu, … and D. Schillinger 2019. Using natural language processing and machine learning to classify health literacy from secure messages: The ECLIPPSE study. *PloS One* 14, no. 2: e0212488.

Balyan, R., K. S. McCarthy, and D. S. McNamara 2017. Combining machine learning and natural language processing to assess literary text comprehension. In A. Hershkovitz and L. Paquette (eds.). In Proceedings of the 10th International Conference on Educational Data Mining (EDM), Wuhan, China: International Educational Data Mining Society.

Balyan, R., K. S. McCarthy, and D. S. McNamara 2018. Comparing machine learning classification approaches for predicting expository text difficulty. In Proceedings of the 31st Annual Florida Artificial Intelligence Research Society International Conference (FLAIRS). AAAI Press, Florida.

Balyan, R., K. S. McCarthy, and D. S.McNamara 2020. Applying natural language processing and hierarchical machine learning approaches to text difficulty classification. *International Journal of Artificial Intelligence in Education* 30, no. 3: 337–370.

Barton, J. L., L. Trupin, D. Schillinger, G. Evans-Young, J. Imboden, V. M. Montori, and E. Yelin 2016. Use of low-literacy decision aid to enhance knowledge and reduce decisional conflict among a diverse population of adults with rheumatoid arthritis: Results of a pilot study. *Arthritis Care & Research* 68, no. 7: 889–898.

Berland, G. K., M. N. Elliott, L. S. Morales, J. I. Algazy, R. L. Kravitz, M. S. Broder, … and E. A. McGlynn 2001. Health information on the internet: Accessibility, quality, and readability in English and Spanish. *JAMA* 285, no. 20: 2612–2621.

BNC Consortium. 2007. *British National Corpus*. University of Oxford, Oxford Text Archive Core Collection, UK.

Boulos, M. N. K. 2005. British internet-derived patient information on diabetes mellitus: Is it readable? *Diabetes Technology & Therapeutics* 7, no. 3: 528–535.

Brach, C., B. P. Dreyer, and D. Schillinger 2014. Physicians' roles in creating health literate organizations: A call to action. *Journal of General Internal Medicine* 29, no. 2: 273–275.

Brach, C., D. Keller, L. M. Hernandez, C. Baur, R. Parker, B. Dreyer, … and D. Schillinger 2012. Ten attributes of health literate health care organizations. *NAM Perspectives, Institute of Medicine of the National Academes*.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24, no. 2: 123–140.

Brown III, W., R. Balyan, A. J. Karter, S. Crossley, W. Semere, N. D. Duran, … and D. Schillinger 2021. Challenges and solutions to employing natural language processing and machine learning to measure patients' health literacy and physician writing complexity: The ECLIPPSE study. *Journal of Biomedical Informatics* 113: 103658.

Bruce, B., and A. Rubin 1988. Readability formulas: Matching tool and task. Lawrence Erlbaum Associates, Inc.

Bruce, B., A. Rubin, and K. Starr 1981. Why readability formulas fail. *IEEE Transactions on Professional Communication*, no.1: 50–52.

Castro, C. M., C. Wilson, F. Wang, and D. Schillinger 2007. Babel babble: Physicians' use of unclarified medical jargon with patients. *American Journal Of Health Behavior* 31, no. 1: S85–S95.

Chall, J. S., and E. Dale 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

Chew, L. D., J. M. Griffin, M. R. Partin, S. Noorbaloochi, J. P. Grill, A. Snyder, … and M. VanRyn 2008. Validation of screening questions for limited health literacy in a large VA outpatient population. *Journal of General Internal Medicine* 23, no. 5: 561–566.

Coltheart, M. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A* 33, no. 4: 497–505.

Crossley, S. A. 2018. How many words needed? Using natural language processing tools in educational data mining. In Proceedings of the 10th International Conference on Educational Data Mining (EDM). pp. 630–633.

Crossley, S. A., R. Balyan, J. Liu, A. J. Karter, D. McNamara, and D. Schillinger 2021. Developing and testing automatic models of patient communicative health literacy using linguistic features: Findings from the ECLIPPSE study. *Health Communication* 36, no. 8: 1018–1028.

Crossley, S. A., R. Balyan, J. Liu, A. J. Karter, D. McNamara, and D. Schillinger 2020. Predicting the readability of physicians' secure messages to improve health communication using novel linguistic features: Findings from the ECLIPPSE study. *Journal of Communication in Healthcare* 13, no. 4: 344–356.

Crossley, S. A., J. Greenfield, and D. S. McNamara 2008. Assessing text readability using cognitively based indices. *Tesol Quarterly* 42, no. 3: 475–493.

Crossley, S., and V. Kostyuk 2017. Letting the genie out of the lamp: Using natural language processing tools to predict math performance. In International Conference on Language, Data and Knowledge. pp. 330–342. Springer, Cham.

Crossley, S. A., K. Kyle, and D. S. McNamara 2015. To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *Grantee Submission* 8, no. 1. Retrieved from https://escholarship.org/uc/item/1f21q8ck

Crossley, S. A., K. Kyle, and D. S. McNamara 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods* 48, no. 4: 1227–1237.

Crossley, S. A., K. Kyle, and D. S. McNamara 2017. Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods* 49, no. 3: 803–821.

Crossley, S., L. Paquette, M. Dascalu, D. S. McNamara, and R. S. Baker 2016. Combining click-stream data with NLP tools to better understand MOOC completion. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. pp. 6–14.

Crossley, S., R. Roscoe, and D. McNamara 2013. Using automatic scoring models to detect changes in student writing in an intelligent tutoring system. In the Twenty-Sixth International FLAIRS Conference.

Crossley, S. A., S. Skalicky, M. Dascalu, D. S. McNamara, and K. Kyle 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes* 54, no. 5-6: 340–359.

Davison, A., and R. N. Kantor 1982. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*: 187–209.

De Clercq, O., V. Hoste, B. Desmet, P. Van Oosten, M. De Cock, and L. Macken 2014. Using the crowd for readability prediction. Natural *Language Engineering* 20, no. 3: 293–325.

De Marneffe, M. C., B. MacCartney, and C. D. Manning 2006. Generating typed dependency parses from phrase structure parses. *International Conference on Language Resources and Evaluation* 6: 449–454.

DeWalt, D. A., D. W. Baker, D. Schillinger, V. Hawk, B. Ruo, K. Bibbins-Domingo, … and M. Pignone 2011. A multisite randomized trial of a single-versus multi-session literacy sensitive self-care intervention for patients with heart failure. *Journal of General Internal Medicine* 26: S57–S58. 233 Spring st, New York, 10013 USA: Springer.

DeWalt, D. A., K. A. Broucksou, V. Hawk, D. W. Baker, D. Schillinger, B. Ruo, … & M. Pignone 2009. Comparison of a one-time educational intervention to a teach-to-goal educational intervention for self-management of heart failure: Design of a randomized controlled trial. *BMC Health Services Research* 9, no. 1: 1–14.

Dumais, S., J. Platt, D. Heckerman, and M. Sahami 1998. Inductive learning algorithms and representations for text categorization. In Proceedings of the Seventh International Conference on Information and Knowledge Management. pp. 148–155.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, no. 2: 179–188.

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology* 32, no. 3: 221.

Friedman, D. B., and L. Hoffman-Goetz 2006. A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education & Behavior* 33, no. 3: 352–373.

Fry, E. 1968. A readability formula that saves time. *Journal of Reading* 11, no. 7: 513–578.

Gemoets, D., G. Rosemblat, T. Tse, and R. A. Logan 2004. Assessing readability of consumer health information: An exploratory study. *Medinfo*. pp. 869–873.

George, D. 2011. SPSS for windows step by step: A simple study guide and reference. *17.0 Update, 10/e*. Pearson Education India.

Grabeel, K. L., J. Russomanno, S. Oelschlegel, E. Tester, and R. E. Heidel 2018. Computerized versus hand-scored health literacy tools: A comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in printed patient education materials. *Journal of the Medical Library Association: JMLA* 106, no. 1: 38.

Graesser, A. C., D. S. McNamara, and J. M. Kulikowich 2011. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher* 40, no. 5: 223–234.

Graesser, A. C., D. S. McNamara, M. M. Louwerse, and Z. Cai 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36, no. 2: 193–202.

Gunning, R. 1952. *Technique of Clear Writing*. McGraw Hill.

Ha, J. F., and N. Longnecker 2010. Doctor-patient communication: A review. *Ochsner Journal* 10, no. 1: 38–43.

Harris, L. T., T. D. Koepsell, S. J. Haneuse, D. P. Martin, and J. D. Ralston 2013. Glycemic control associated with secure patient-provider messaging within a shared electronic medical record: A longitudinal analysis. *Diabetes Care* 36, no. 9: 2726–2733.

Hastie, T., R. Tibshirani, and J. Friedman 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hill-Briggs, F., K. P. Schumann, and O. Dike 2012. 5-step methodology for evaluation and adaptation of print patient health information to meet the < 5th grade readability criterion. *Medical Care* 50, no. 4: 294.

Hudson, S., R. V. Rikard, I. Staiculescu, and K. Edison 2018. Improving health and the bottom line: The case for health literacy. In Building the Case for Health Literacy: Proceedings of a Workshop. National Academies Press (US).

Institute of Medicine. 2004. *Health lIteracy: A Prescription to End Confusion*. Washington, DC: The National Academies Press.

James, G., D. Witten, T. Hastie, and R. Tibshirani 2013. *An Introduction to Statistical Learning*, Vol. 112, p. 18. New York: Springer.

Jindal, P., and J. C. MacDermid 2017. Assessing reading levels of health information: Uses and limitations of flesch formula. *Education for Health* 30, no. 1: 84.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*. pp. 137–142. Berlin, Heidelberg: Springer.

Johnson, S. B. 1999. A semantic lexicon for medical language processing. *Journal of the American Medical Informatics Association* 6, no. 3: 205–218.

Kandula, S., and Q. Zeng-Treitler 2008. Creating a gold standard for the readability measurement of health texts. In *AMIA Annual Symposium Proceedings*, Vol. 2008, p. 353. American Medical Informatics Association.

Karter, A. J., M. M. Parker, O. K. Duru, D. Schillinger, N. E. Adler, H. H. Moffet, … and J. A. Schmittdiel 2015. Impact of a pharmacy benefit change on new use of mail order pharmacy among diabetes patients: The Diabetes Study of Northern California (DISTANCE). *Health Services Research* 50, no. 2: 537–559.

Karter, A. J., M. M. Parker, H. H. Moffet, A. T. Ahmed, J. A. Schmittdiel, and J. V. Selby 2009. New prescription medication gaps: A comprehensive measure of adherence to new prescriptions. *Health Services Research* 44, no. 5p1: 1640–1661.

Kim, H., S. Goryachev, G. Rosemblat, A. Browne, A. Keselman, and Q. Zeng–Treitler 2007. Beyond surface characteristics: A new health text-specific readability measurement. In *AMIA Annual Symposium Proceedings*. pp. 418–422. Chicago, IL: American Medical Informatics Association.

Kincaid, J. P., R. P. Fishburne, Jr, R. L. Rogers, and B. S. Chissom 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch.

Kindig, D. A., A. M. Panzer, and L. Nielsen-Bohlman (Eds.). 2004. Health literacy: A prescription to end confusion. National Academies Press.

Kirsch, I. S., A. Jungeblut, L. Jenkins, and A. Kolstad 2002. *Adult Literacy in America: A First Look at the Findings of the National Adult Literacy Survey (NCES 1993–275)*. Washington, DC: U.S. Department of Education.

Krogh, A., and J. Vedelsby 1995. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*. pp. 231–238.

Kugar, M. A., A. C. Cohen, W. Wooden, S. S. Tholpady, and M. W. Chu 2017. The readability of psychosocial wellness patient resources: Improving surgical outcomes. *Journal of Surgical Research* 218: 43–48.

Kusec, S., O. Brborovic, and D. Schillinger 2003. Diabetes websites accredited by the Health on the Net Foundation Code of Conduct: Readable or not? *Studies in Health Technology and Informatics* 95: 655–660.

Kyle, K. 2016. Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. Dissertation, Georgia State University.

Kyle, K., and S. A. Crossley 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly* 49, no. 4: 757–786.

Kyle, K., S. Crossley, and C. Berger 2018. The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods* 50, no. 3: 1030–1046.

Lu, X. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15, no. 4: 474–496.

Lyles, C. R., U. Sarkar, D. Schillinger, J. D. Ralston, J. Y. Allen, R. Nguyen, and A. J. Karter 2016. Refilling medications through an online patient portal: Consistent improvements

in adherence across racial/ethnic groups. *Journal of the American Medical Informatics Association* 23, no. e1: e28–e33.

Malvern, D., B. Richards, N. Chipere, and P. Durán 2004. *Lexical Diversity and Language Development*. New York: Palgrave Macmillan.

McAndie, E., A. Gilchrist, and B. Ahamat 2016. Readability of clinical letters sent from a young people's department. *Child and Adolescent Mental Health* 21, no. 3: 169–174.

McCallum, A., and K. Nigam 1998. A comparison of event models for naive bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization* 752, no. 1: 41–48.

McCarthy, P. M. 2005. *An aSsessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)* Doctoral dissertation, Tennessee: The University of Memphis.

McLaughlin, G. H. 1969. SMOG grading-a new readability formula. *Journal of Reading* 12, no. 8: 639–646.

McNamara, D. S., S. A. Crossley, and P. M. Mccarthy 2010. Linguistic features of writing quality. *Written Communication* 27, no. 1: 57–86.

McNamara, D. S., S. A. Crossley, and R. Roscoe 2013. Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods* 45, no. 2: 499–515.

McNamara, D. S., S. A. Crossley, R. D. Roscoe, L. K. Allen, and J. Dai 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing* 23: 35–59.

McNamara, D. S., A. C. Graesser, P. M. McCarthy, and Z. Cai 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

McNamara, D. S., R. Roscoe, L. K. Allen, R. Balyan, and K. S. McCarthy 2019. Literacy: From the perspective of text and discourse theory. *Journal of Language and Education* 5, no. 3: 56–69.

Meade, C. D., J. C. Byrd, and M. Lee 1989. Improving patient comprehension of literature on smoking. *American Journal of Public Health* 79, no. 10: 1411–1412.

Miller, G. A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38, no. 11: 39–41.

Mitchell, T. M. 1997. *Machine Learning*. New York: McGraw-Hill.

Moffet, H. H., N. Adler, D. Schillinger, A. T. Ahmed, B. Laraia, J. V. Selby, … and A. J. Karter 2009. Cohort profile: The Diabetes Study of Northern California (DISTANCE)—Objectives and design of a survey follow-up study of social health disparities in a managed care population. *International Journal of Epidemiology* 38, no. 1: 38–47.

Munsour, E. E., A. Awaisu, M. A. A. Hassali, S. Darwish, and E. Abdoun 2017. Readability and comprehensibility of patient information leaflets for antidiabetic medications in Qatar. *Journal of Pharmacy Technology* 33, no. 4: 128–136.

Oliffe, M., E. Thompson, J. Johnston, D. Freeman, H. Bagga, and P. K. Wong 2019. Assessing the readability and patient comprehension of rheumatology medicine information sheets: A cross-sectional health literacy study. *BMJ Open* 9, no. 2: e024582.

Paasche-Orlow, M. K., H. A. Taylor, and F. L. Brancati 2003. Readability standards for informed-consent forms as compared with actual readability. *New England Journal of Medicine* 348, no. 8: 721–726.

Piñero-López, M. Á., P. Modamio, C. F. Lastra, and E. L. Mariño 2016. Readability analysis of the package leaflets for biological medicines available on the internet between 2007 and 2013: An analytical longitudinal study. *Journal of Medical Internet Research* 18, no. 5: e100.

Pitler, E., and A. Nenkova 2008. Revisiting readability: A unified framework for predicting text quality. In Proceedings of the 2008 conference on empirical methods in natural language processing. pp. 186–195.

Protection, P., and A. C. Act 2010. Patient protection and affordable care act. *Public Law* 111, no. 48: 759–762.

Ratanawongsa, N., A. J. Karter, M. M. Parker, C. R. Lyles, M. Heisler, H. H. Moffet, … and D. Schillinger 2013. Communication and medication refill adherence: The diabetes study of Northern California. *JAMA Internal Medicine* 173, no. 3: 210–218.

Reed, M., J. Huang, R. Brand, I. Graetz, R. Neugebauer, B. Fireman, … and J. Hsu 2013. Implementation of an outpatient electronic health record and emergency department visits, hospitalizations, and office visits among patients with diabetes. *Jama* 310, no. 10: 1060–1065.

Reed, M., J. Huang, I. Graetz, R. Brand, J. Hsu, B. Fireman, and M. Jaffe 2012. Outpatient electronic health records and the clinical care and outcomes of patients with diabetes mellitus. A*nnals of Internal Medicine* 157, no. 7: 482–489.

Rojas, R. 2013. *Neural Networks: A Systematic Introduction*. Springer Science & Business Media.

Rubin, A. 1985. How useful are readability formulas. *Reading Education: Foundations for a Literate America*: 61–77.

Sarkar, U., A. J. Karter, J. Y. Liu, H. H. Moffet, N. E. Adler, and D. Schillinger 2010. Hypoglycemia is more common among type 2 diabetes patients with limited health literacy: The diabetes study of Northern California (DISTANCE). *Journal of General Internal Medicine* 25, no. 9: 962–968.

Sarkar, U., C. R. Lyles, M. M. Parker, J. Allen, R. Nguyen, H. H. Moffet, … and A. J. Karter 2014. Use of the refill function through an online patient portal is associated with improved adherence to statins in an integrated health system. *Medical Care* 52, no. 3: 194.

Sarkar, U., J. D. Piette, R. Gonzales, D. Lessler, L. D. Chew, B. Reilly, … and D. Schillinger 2008. Preferences for self-management support: Findings from a survey of diabetes patients in safety-net health systems. *Patient Education and Counseling* 70, no. 1: 102–110.

Sarkar, U., D. Schillinger, A. López, and R. Sudore 2011. Validation of self-reported health literacy questions among diverse English and Spanish-speaking populations. *Journal of General Internal Medicine* 26, no. 3: 265–271.

Schapire, R. E., and Y. Singer 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning* 39, no. 2-3: 135–168.

Schillinger, D. 2007. Literacy and health communication: Reversing the 'inverse care law'. *The American Journal of Bioethics* 7, no. 11: 15–18.

Schillinger, D., R. Balyan, S. Crossley, D. McNamara, and A. Karter 2021. Validity of a computational linguistics-derived automated health literacy measure across race/ethnicity: Findings from the ECLIPPSE project. *Journal of Health Care for the Poor and Underserved* 32, no. 2: 347–365.

Schillinger, D., R. Balyan, S. A. Crossley, D. S. McNamara, J. Y. Liu, and A. J. Karter 2020. Employing computational linguistics techniques to identify limited patient health literacy: Findings from the ECLIPPSE study. *Health Services Research*. 56, no. 1: 132–144.

Schillinger, D., A. Bindman, F. Wang, A. Stewart, and J. Piette 2004. Functional health literacy and the quality of physician–patient communication among diabetes patients. *Patient Education and Counseling* 52, no. 3: 315–323.

Schillinger, D., K. Grumbach, J. Piette, F. Wang, D. Osmond, C. Daher, … and A. B. Bindman 2002. Association of health literacy with diabetes outcomes. *Jama* 288, no. 4: 475–482.

Schillinger, D., H. Hammer, F. Wang, J. Palacios, I. McLean, A. Tang, … and M. Handley 2008. Seeing in 3-D: Examining the reach of diabetes self-management support strategies in a public health care system. *Health Education & Behavior* 35, no. 5: 664–682.

Schillinger, D., M. Handley, F. Wang, and H. Hammer 2009. Effects of self-management support on structure, process, and outcomes among vulnerable patients with diabetes: A three-arm practical clinical trial. *Diabetes Care* 32, no. 4: 559–566.

Schillinger, D., D. McNamara, S. Crossley, C. Lyles, H. H. Moffet, U. Sarkar, … and A. J. Karter 2017. The next frontier in communication and the ECLIPPSE study: Bridging the linguistic divide in secure messaging. *Journal of Diabetes Research* 2017.

Schillinger, D., J. Piette, K. Grumbach, F. Wang, C. Wilson, C. Daher, … and A. B. Bindman 2003. Closing the loop: Physician communication with diabetic patients who have low health literacy. *Archives of Internal Medicine* 163, no. 1: 83–90.

Schumaier, A. P., R. Kakazu, C. E. Minoughan, and B. M. Grawe 2018. Readability assessment of American shoulder and elbow surgeons patient brochures with suggestions for improvement. *JSES Open Access* 2, no. 2: 150–154.

Seligman, H. K., F. F. Wang, J. L. Palacios, C. C. Wilson, C. Daher, J. D. Piette, and D. Schillinger 2005. Physician notification of their diabetes patients' limited health literacy: A randomized, controlled trial. *Journal of General Internal Medicine* 20, no. 11: 1001–1007.

Semere, W., S. Crossley, A. J. Karter, C. R. Lyles, W. Brown, M. Reed, … and D. Schillinger 2019. Secure messaging with physicians by proxies for patients with diabetes: Findings from the ECLIPPSE Study. *Journal of General Internal Medicine* 34, no. 11: 2490–2496.

Smith, F. 2012. *Understanding Reading: A Psycholinguistic Analysis of Reading and Learning to Read*. Routledge.

Smith, S. G., R. O'Conor, L. M.Curtis, K. Waite, I. J. Deary, M. Paasche-Orlow, and M. S. Wolf 2015. Low health literacy predicts decline in physical function among older adults: Findings from the LitCog cohort study. *J Epidemiol Community Health* 69, no. 5: 474–480.

Smola, A. J., and B. Schölkopf 1998. *Learning with Kernels*, Vol. 4. GMD-Forschungszentrum Informationstechnik.

Steiner, J. F., T. D. Koepsell, S. D. Fihn, and T. S. Inui 1988. A general method of compliance assessment using centralized pharmacy records: Description and validation. *Medical Care*: 814–823.

Steiner, J. F., and A. V. Prochazka 1997. The assessment of refill compliance using pharmacy records: Methods, validity, and applications. *Journal of Clinical Epidemiology* 50, no. 1: 105–116.

Sudore, R. L., C. S. Landefeld, E. J. Perez-Stable, K. Bibbins-Domingo, B. A. Williams, and D. Schillinger 2009. Unraveling the relationship between literacy, language proficiency, and patient–physician communication. *Patient Education and Counseling* 75, no. 3: 398–402.

Sudore, R. L., K. Yaffe, S. Satterfield, T. B. Harris, K. M. Mehta, E. M. Simonsick, … and D. Schillinger 2006. Limited literacy and mortality in the elderly. *Journal of General Internal Medicine* 21, no. 8: 806–812.

Thompson, P., R. T. Batista-Navarro, G. Kontonatsios, J. Carter, E. Toon, J. McNaught, … and S. Ananiadou 2016. Text mining the history of medicine. *PloS One* 11, no. 1: e0144717.

Travaline, J. M., R. Ruchinskas, and G. E. D'Alonzo Jr 2005. Patient-physician communication: Why and how. *Journal of the American Osteopathic Association* 105, no. 1: 13.

Uzuner, Ö. 2009. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association* 16, no. 4: 561–570.

Uzuner, Ö., I. Goldstein, Y. Luo, and I. Kohane 2008. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association* 15, no. 1: 14–24.

Uzuner, Ö., Y. Luo, and P. Szolovits 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association* 14, no. 5: 550–563.

Uzuner, Ö., I. Solti, and E. Cadag 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association* 17, no. 5: 514–518.

Walsh, T. M., and T. A. Volsko 2008. Readability assessment of internet-based consumer health information. *Respiratory Care* 53, no. 10: 1310–1315.

Wang, L. W., M. J. Miller, M. R. Schmitt, and F. K. Wen 2013. Assessing readability formula differences with written health information materials: application, results, and recommendations. *Research in Social and Administrative Pharmacy* 9, no. 5: 503–516.

Wilson, M. 2009. Readability and patient education materials used for low-income populations. *Clinical Nurse Specialist* 23, no. 1: 33–40.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks* 5, no. 2: 241–259.

Wu, D. T., D. A. Hanauer, Q. Mei, P. M. Clark, L. C. An, J. Lei, … and K. Zheng 2013. Applying multiple methods to assess the readability of a large corpus of medical documents. *Studies in Health Technology and Informatics* 192: 647–651.

Zeng–Treitler, Q., S. Kandula, H. Kim, and B. Hill 2012. A method to estimate readability of health content. In Proceedings of HI-KDD 2012: ACM SICKDD Workshop on Health Informatics (HI-KDD 2012), Beijing, China.

Zhang, G. P. 2000. Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 30, no. 4: 451–462.

Zheng, J., and H. Yu 2017. Readability formulas and user perceptions of electronic health records difficulty: A corpus study. *Journal of Medical Internet Research* 19, no. 3: e59.

Zheng, J., and H. Yu 2018. Assessing the readability of medical documents: A ranking approach. *JMIR Medical Informatics* 6, no. 1: e17.

Zhou, Y. Y., M. H. Kanter, J. J. Wang, and T. Garrido 2010. Improved quality at Kaiser permanente through email between physicians and patients. *Health Affairs* 29, no. 7: 1370–1375.