HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2021 June 08.

Published in final edited form as:

J Biomed Inform. 2021 January ; 113: 103658. doi:10.1016/j.jbi.2020.103658.

Challenges and solutions to employing natural language processing and machine learning to measure patients' health literacy and physician writing complexity: The ECLIPPSE study

William Brown III[a,b,c,d,*,1], **Renu Balyan**[e,i], **Andrew J. Karter**[f], **Scott Crossley**[g], **Wagahta Semere**[d], **Nicholas D. Duran**[h], **Courtney Lyles**[c,d,f], **Jennifer Liu**[f], **Howard H. Moffet**[f], **Ryane Daniels**[c], **Danielle S. McNamara**[i], **Dean Schillinger**[c,d,f]

[a]Center for AIDS Prevention Studies, University of California, San Francisco, San Francisco, CA, United States

[b]Bakar Computational Health Science Institute, University of California, San Francisco, San Francisco, CA, United States

[c]University of California San Francisco Center for Vulnerable Populations, Zuckerberg San Francisco General Hospital, San Francisco, CA, United States

[d]Department of Medicine, University of California, San Francisco, San Francisco, CA, United States

[e]State University of New York Old Westbury, NY, United States

[f]Division of Research, Kaiser Permanente Northern California, Oakland, CA, United States

[g]Department of Applied Linguistics and English as a Second Language, Georgia State University, Atlanta, GA, United States

[h]School of Social and Behavioral Sciences, Arizona State University, Glendale, AZ, United States

[i]Department of Psychology, Arizona State University, Tempe, AZ, United States

Abstract

Objective: In the National Library of Medicine funded ECLIPPSE Project (**E**mploying **C**omputational **L**inguistics to **I**mprove **P**atient-**P**rovider **S**ecure **E**mails exchange), we attempted to create novel, valid, and scalable measures of both patients' health literacy (HL) and physicians' linguistic complexity by employing natural language processing (NLP) techniques and machine learning (ML). We applied these techniques to > 400,000 patients' and physicians' secure

*Corresponding author at: Center for AIDS Prevention Studies (CAPS), Prevention Research Center (PRC), UCSF Box 0886, 550 16th Street, 3rd Floor, San Francisco, CA 94158, United States. william.brown@ucsf.edu (W. Brown).

Contributions

Study concept and design: S, C, K, B, M; acquisition of subjects and/ or data: S, C, K, Liu, M; analysis and interpretation of data: B, S, C, K, M, Liu, D; and preparation of manuscript: B, S, C, K, D, L, L, M, R, S.

[1]http://profiles.ucsf.edu/william.brown.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Danielle McNamara is the founder and owner of Adaptive Literacy Technologies. Funding from the company was not used to underwrite this research and the company will not benefit from this research.

messages (SMs) exchanged via an electronic patient portal, developing and validating an automated patient literacy profile (LP) and physician complexity profile (CP). Herein, we describe the challenges faced and the solutions implemented during this innovative endeavor.

Materials and methods: To describe challenges and solutions, we used two data sources: study documents and interviews with study investigators. Over the five years of the project, the team tracked their research process using a combination of Google Docs tools and an online team organization, tracking, and management tool (Asana). In year 5, the team convened a number of times to discuss, categorize, and code primary challenges and solutions.

Results: We identified 23 challenges and associated approaches that emerged from three overarching process domains: (1) *Data Mining* related to the SM corpus; (2) *Analyses* using NLP indices on the SM corpus; and (3) *Interdisciplinary Collaboration*. With respect to Data Mining, problems included cleaning SMs to enable analyses, removing hidden caregiver proxies (e.g., other family members) and Spanish language SMs, and culling SMs to ensure that only patients' primary care physicians were included. With respect to Analyses, critical decisions needed to be made as to which computational linguistic indices and ML approaches should be selected; how to enable the NLP-based linguistic indices tools to run smoothly and to extract meaningful data from a large corpus of medical text; and how to best assess content and predictive validities of both the LP and the CP. With respect to the Interdisciplinary Collaboration, because the research required engagement between clinicians, health services researchers, biomedical informaticians, linguists, and cognitive scientists, continual effort was needed to identify and reconcile differences in scientific terminologies and resolve confusion; arrive at common understanding of tasks that needed to be completed and priorities therein; reach compromises regarding what represents "meaningful findings" in health services vs. cognitive science research; and address constraints regarding potential transportability of the final LP and CP to different health care settings.

Discussion: Our study represents a process evaluation of an innovative research initiative to harness "big linguistic data" to estimate patient HL and physician linguistic complexity. Any of the challenges we identified, if left unaddressed, would have either rendered impossible the effort to generate LPs and CPs, or invalidated analytic results related to the LPs and CPs. Investigators undertaking similar research in HL or using computational linguistic methods to assess patient-clinician exchange will face similar challenges and may find our solutions helpful when designing and executing their health communications research.

Keywords

Health literacy; Natural language processing; Machine learning; Diabetes health care quality; Electronic health records; Digital health and health services research

1. Objective

1.1. The ECLIPPSE study and data sources

In the ECLIPPSE Project (**E**mploying **C**omputational **L**inguistics to **I**mprove **P**atient-**P**rovider **S**ecure **E**mails exchange), we created novel, valid and scalable measures of patients' health literacy (HL) and physicians' linguistic complexity by applying natural language processing (NLP) and machine learning (ML) techniques to patients' and

physicians' secure messages (SMs) sent via an electronic patient portal. By leveraging an existing large, previously untapped database of SM exchanges, we developed what we call the patient Literacy Profile (LP) and the physician Complexity Profile (CP) [1–6]. In this paper, we enumerate the challenges that we encountered while attempting to employ NLP and ML to develop the patient LP and physician CP, as well as describe the solutions that we developed and applied to address challenges in developing and validating the patient LP and physician CP. Our hope is that summarizing our experience will help facilitate the work of those interested in applying our new tools to their health system's SM data and accelerate the work of other investigators attempting to harness computational linguistic methods to assess natural language production and exchange to improve health communication and reduce related health disparities.

The data used in the ECLIPPSE Study comes from a sampling frame of more than one million SMs generated by a sample of > 12,000 ethnically diverse diabetes patients and > 15,000 clinician providers contained in The Kaiser Permanente Northern California (KPNC) Diabetes Registry [1]. The ECLIPPSE analysis focuses on the SMs exchanged between patients and their primary care physicians via the KP e-patient portal over a 10-year period between 2006 and 2015 (detailed in Fig. 1). Selected patients had previously completed the Diabetes Study of Northern California (The DISTANCE Study) survey [7–9]. Data collection methods, descriptive statistics, and detailed characteristics of the NIH-funded ECLIPPSE study, and DISTANCE Study sample, have been previously published [2,4–6,10]. This study was approved by the KPNC and UCSF Institutional Review Boards (IRBs). All analyses involved secondary data; all data were housed on a password-protected secure KPNC server that could only be accessed by Kaiser-authorized researchers and prevented copying or transferring of data.

1.2. The importance of characterizing patient health literacy and physician writing complexity

Health literacy (HL) is defined as an individual's capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions [11]. Limited HL is associated with worse health outcomes and greater health disparities [12–26]. Existing HL measurement tools present significant challenges with respect to administration and scaling because of there are time-intensive nature and/or their requirement to must be administered in person. Identification of limited HL in a more efficient manner has the potential to inform and improve healthcare, reduce communication-related disparities [27], inform quality improvement and care management initiatives [28,29], and enable targeting and tailoring of population management strategies [30]. To address this challenge we attempted to develop a novel, automated measure of HL that was generated from computational linguistic analyses of patients' written language [6]. This represents the first attempt to measure HL by assessing patients' own original written content, specifically written communications to their physicians [31–34].

Reducing physicians use of medical jargon and language complexity can reduce HL demands on patients [35–37]. Despite simple tools like Flesch – Kincaid readability level [38], there currently are no high-throughput, theory-driven tools with sufficient validity to

assess writing complexity using samples of physicians' written communications with their patients [5]. Developing a robust measure of physicians' linguistic complexity, when applied in concert with a patient HL measure, could allow researchers to measure linguistic discordance between physicians and patients, and ascertain its proximal communication consequences, as well as intermediate and long-term clinical outcomes [39]. Furthermore, such a measure could assist health systems in identifying those physicians who might benefit from additional communication training and support [40,41]. We attempted to develop a novel, automated measure of readability of health-related text that was generated from computational linguistic analyses of physicians' written language [6].

2. Materials and methods

2.1. Data sources to identify challenges and solutions

To describe challenges and solutions, we used three data sources: study documents, interviews with individual study investigators, and an online, virtual focus group. Over the five years of the project, the team carried out its regular business through biweekly, one-hour, all-team research meetings conducted by phone or Webinar, and held, in-person meetings over two days each year. A project director recorded minutes of these meetings and kept track of the research process using a combination of Google Docs tools and an online team organization, tracking, and management tool (Asana work management platform) [42]. Starting at the end of year 4 and into year 5 of the study, under the direction of one of the researchers who is expert in biomedical informatics methods (WB), different combinations of the team were convened to iteratively discuss, categorize, and code critical challenges and solutions on numerous occasions (see Table 1).

2.1.1. Documents—Google Docs was used to store and share meeting minutes and related and to promote collaborative work across sites. We reviewed unique study documents (n = 93) including: the parent grant, three published journal articles from the study and three manuscripts in review, meeting minutes (n = 67), presentation slides and notes (n = 7), documents related to LP (n = 21), documents related to CP (n = 7), IRB protocols (n = 2) NLM progress reports (n = 4), and data specification documents (n = 19). WB reviewed all documents with guidance from DS.

2.1.2. Interviews, virtual focus group, and email communications—After WB and DS introduced the idea for this study at a bi-weekly team meeting, each member of the team agreed to contribute to a comprehensive review of the methodologic process related to the ECLIPPSE project. Eight ECLIPPSE investigators were subsequently asked by WB to participate in communications regarding challenges and solutions. Interviews in the first round were conducted by WB, and were general and open-ended, with each investigator being asked which part of the project they were in charge of, what challenges they had faced, what kinds of problem-solving had been attempted and what types of solutions had been arrived at. Of these eight, five were then invited by WB to participate in a virtual, online focus group that lasted for 90 min. The purpose of the focus group was to mitigate regal bias by allowing researchers to act as sounding boards and identify those challenges and solutions that were shared between and within disciplines and teams. Participants were

asked about challenges and solutions specific to the tasks that they had to perform, or to clarify who was knowledgeable about challenges and solutions they had less involvement in. Participants were presented with a preliminary table of challenges and solutions related to both the LP and the CP – based on the review of the study documents and the first set of interviews – to stimulate recall and generate rich discussion and promote consensus-building. Based on the information generated in the focus group, three investigators plus one additional team member who had not participated in the focus group (a senior biostatistician) were then asked for follow-up communications after the virtual focus group. Three of these investigators were interviewed by WB over email and one by phone to delve deeper and to elicit more specifics about the challenges and solutions within and across study domains. Field notes were taken for all interviews; the focus group was recorded and transcribed.

2.2. Coding

All data collected from all sources were coded by WB and DS. Though these two researchers discussed potential themes based on their experience in the ECLIPPSE project, ultimately the final themes and codes were developed using a bottom-up, inductive approach. Both raters reviewed all documents, excerpts, and codes and reconciled coding discordance through discussion. Challenges were classified into thematic topic process domains as follows: data mining, analysis, and interdisciplinary collaboration; these were then placed into a matrix (Table 1).

3. Results

We encountered a total of 23 challenges and solutions implemented to overcome these challenges (Table 1). These emerged from one or more of the three overarching process domains: (1) *Data Mining* related to the SMs corpus; (2) *Analyses* of computational linguistic indices based on the SM corpus; and (3) *Interdisciplinary Collaboration*. With respect to *Data Mining*, problems included preparing and cleaning SMs to enable analyses; selecting SMs to ensure that only the appropriate clinician recipients or senders were included; and removing SMs written by formal or hidden caregiver proxies (e.g., adult child, spouse, or other family member) and Spanish language SMs. With respect to *Analyses,* critical decisions needed to be made as to which computational linguistic indices should be selected for generating the patient LP and physician CP, and how to most effectively assess the content and predictive validity of both LPs and CPs. With respect to *Interdisciplinary Collaboration*, significant effort was needed to: identify and then reconcile confusion from and differences in scientific terminologies; arrive at common understandings of tasks that needed to be completed and priorities therein; reach compromise regarding what represents a “meaningful finding ” in applied health services vs. cognitive science research; and manage expectations and tensions between developing patient LP and physician CP measures with sufficient internal validity vs. the immediate translatability/generalizability of the final LP and CP to different health care settings and populations (external validity). Herein, we explain the challenges we faced and separately describe the strategies we employed to try to overcome these challenges.

3.1. Data mining problems related to the corpus of secure messages

3.1.1. Data mining challenges—The first data mining challenge we encountered was related to data extraction. The goal was to extract both the patient and physician identification numbers (IDs) and patient and physician SMs message text. Subsequent to the data extraction challenge we then had to deal with missing and/or incorrect structural markers (i.e., punctuations, paragraph breaks, sentence markers, etc.). The final linguistic data provided to the research team for analysis, when extracted from the original database, were largely unstructured. Both the required data cleaning steps as well as the nature of SM emails (text-like statements) led to frequent absence of potentially important structural markers, such as paragraph breaks and sentence indicators necessary for certain NLP analyses. The absence of such markers created difficulties when examining linguistic features across paragraphs (e.g., lexical repetition and cohesion). Structural issues also influenced syntactic indices that rely on syntactic parsing, potentially leading to imprecise calculations and occasionally creating buffer-related issues that halted the linguistic index program's progress.

These pre-processing steps were further complicated by information including lab reports, hyperlinks, website URL's, physician auto-signatures, location address and office hours, all of which could lead to additional potential problems related to parsing, lexical analysis, analysis of discourse features and imprecise calculation of linguistic features. Another issue was that patient or physician names and phone numbers necessitated additional data security measures be taken. We also identified that physicians' SMs occasionally included what appeared to be automated content. Often known as "smart texts" or "smart phrases," these reflect standardized stock content that physicians can use by selecting from a menu of pre-determined responses (see Table 1).

Beyond the data mining problems found within the text, we also confronted problems with SM authorship on the patient side: we observed that some SMs appeared to be written by patient proxies or contained non-English data. Since the goal of the LP was to estimate patient HL, the existence of patient proxy SMs interspersed with patient SMs had to be dealt with. The occasional non-English SM had to be removed as well because NLP tools for other languages are not as advanced as they are for the English language, thus making comparable referencing of linguistic features across two or more languages impossible.

3.1.2. Data mining solutions—We matched the patients' EHR data medical record numbers (MRNs) to their KP patient portal IDs and data. We then mapped their KP patient portal message IDs to their KP patient portal message IDs in the EHR data and extracted the SM text from the notes in the physician-facing EHR. These notes in the physician-facing EHR contained the full KP patient portal SM exchange between patient and physician.

Next, to address the problem of the parser stoppages, periodic human oversight of data processing was necessary. When parser stoppages occurred, the location of the stoppage was excised, and the parser was run again.

Since the goal of the LP was to estimate patient HL, the existence of patient proxy SMs interspersed with patient SMs had to be dealt with. Prior to our study, little had been known

about patients who use caregiver proxies (e.g., adult child, spouse, or other family member) to communicate with healthcare providers on their behalf via portal secure messaging. Given proxies often write SMs informally using patients' accounts as opposed to registering for their own account, proxy communication is often hidden. As a result, we created and validated a novel algorithm "ProxyID" that specifically identified hidden proxy messages [3]. Using a threshold of > 50% proxy penetration [3] led to the exclusion of ~500 patients and ~70,000 SMs. By applying ProxyID to our corpus, we were able to identify SMs written by a formally designated as well as an informal (hidden) proxy.

The occasional non-English SM had to be removed as well because NLP tools for other languages are not as advanced as they are for the English language, thus making comparable referencing of linguistic features across two or more languages impossible. We also employed a script that identified non-English text [2,4]. The algorithm removed non-English (Spanish) text if it exceeded a threshold (>50% of SM was non-English). Thus, a small proportion of the SM corpus may have contained residual non-English text.

The data mining and pre-processing were further complicated by the need to maintain security of the confidential information. It was impractical to de-identify the voluminous data (e.g., remove patient names and phone numbers that occasionally existed in the messages). This necessitated that all storage and analysis of the data take place on a secure server behind the KPNC firewall. While the secure server represented a solution to the challenge of maintaining security on confiden tial information, the processes for receiving training and obtaining access to the secure server were understandably rigorous and time-consuming. Furthermore, occasional server connectivity problems and limitations on computational speed of analyses performed via the server portal, together created occasional delays in data processing for the non-KPNC investigators on the team.

With regards to the standardized clinical content that physicians can use by selecting from a menu of pre-determined responses, these automated text types had to be left in the corpus as it was not possible to create a generic NLP tool that could accurately identify these smart texts or phrases automatically (see Table 1). We also elected to retain these automated texts for subsequent linguistic analyses because such text was representative of some of the language used by physicians when messaging patients.

To address these possible imprecisions during LP and CP model development, we ran testing and training sets and used cross validation to try and maintain generalizability across the entire sample population (see below).

3.2. Analyses of computational linguistic indices on the secure message corpus

3.2.1. Analysis challenges—There were structural challenges to analyzing the data. Given the limitations of some of the standard NLP algorithms, some SMs were too short to enable robust linguistic analysis.

A major challenge was application of linguistic tools (available at linguisticanalysistools.org) for extracting and selecting the indices used to train the machine learning models for the LP and CP algorithms. The indices were selected from linguistic

tools that export hundreds of indices applied to the SMs exchanged between the patients and the physicians [28,30]. As such, various decisions needed to be made regarding whether and how to reduce the set of indices.

Imbalanced sample sizes in health literacy estimations were a major concern for developing the LP algorithm. For instance, compared to what we observed with respect to self-reported HL and what is known from the HL literature, there were relatively fewer people than expected who were modeled to have low LP and more people than expected to have high LP. The traditional ML algorithms such as Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) do not work well with imbalanced or skewed data.

In considering how to assess performance of both LP and CP, we faced a critical challenge due to the absence of true "gold standards" for either patient HL or physician linguistic complexity. While we did have self-reported HL as one previously validated "gold standard" for the development of the patient LP [43], it is a subjective measure that is more aligned with the construct of "HL-related self-efficacy " and is therefore somewhat limited. Insofar as our solution included developing expert ratings based on a review of SM content in a small subset of the corpus (see below), a related challenge was developing and refining the scoring rubrics and training the raters to reliably assess both patient HL and physician linguistic complexity. Deciding on a sub-sample with which to assess expert-rated HL and expert-rated physician linguistic complexity and determining thresholds for them both based on SM content presented additional challenges.

3.2.2. Analysis solutions

In order to address the minimum word requirement for processing some of the NLP algorithms, we applied a threshold wherein an SM could not contain fewer than 50 words for NLP analysis for patient secure messages.

Several approaches were used to reduce the number of linguistic indices included in the LP and CP algorithms. First, we reduced the set by applying typical filtering methods such as removing indices based on multi-collinearity, non-normal distributions, and non-normal variance (e.g., zero or near zero variance). When choosing between highly correlated indices, we selected theoretically motivated features/indices with demonstrated validity in previous writing-based studies. Second, selected the topmost important indices obtained after training the models. With the exception of the first LP model developed, these methods resulted in models for the LPs and the CPs that included from 15 to 20 indices.

We examined several methods of accounting for the imbalance that resulted from our initial analyses. Because the data we initially generated were imbalanced, the ML approach had to be adapted to different types of imbalances and the thresholds had to be set accordingly. As such, we explored whether alternative ML approaches would be more appropriate. In the end, we both refined our expert rating scoring systems and adjusted the ML algorithm scoring thresholds to balance the rating proportions. By doing so, we achieved a more balanced proportion of SMs that met the threshold versus those that were below threshold for both the LP and CP algorithms. These computational processes and their validity are detailed in papers that describe the development of our LP[2] and CP[5] algorithms. We also

explored the extent to which alternative ML approaches (such as under-sampling, oversampling or SMOTE) that correct for imbalanced data might be more appropriate. Ultimately, we decided in favor of adjusting the thresholds, but plan to explore alternative techniques in future research.

With respect to the gold standard problem, for the patient LP we applied two proxy measures: *self-reported HL* from the DISTANCE survey [43] and a novel measure of *expert-rated HL* based on review and scoring of patients' SM content on a sub-sample of patient and physician SMs [2,5]. This sub-sample was purposively sampled to contain a balanced sample of SMs across patient demographics (race, age) and self-reported HL. As a result, we generated two versions of our LP algorithms (LP-Self Report and LP-Expert) [2], and compared relative performance for each, as well as relative to more simple methods (Flesch-Kincaid) [2,4]. We found the novel LP-Self Report to be a valid measure [2] and found the expert HL rating method to have adequate inter-rater reliability, and the resultant novel LP-Expert to also be a valid measure [4]. For the physician CP, because there was no gold standard for this analysis, we developed a novel measure of expert-rated linguistic complexity based on review and scoring of physicians' SM content on this same sub-sample and found it to have adequate inter-rater reliability and validity [5]. The sub-sample contained 724 unique SM threads from 592 individual physicians. From this random sample, these physicians sent, on average, 1.23 SMs to patients; 112 of these 592 physicians messaged at least two different patients [5]. Because Excess SM length might make human ratings of physician SMS difficult, we needed to standardize the human rating process. Physicians SM threads were randomly trimmed to contain approximately ~300 words [44,45]. No individual SMs contained in the threads were truncated.

3.3. Interdisciplinary collaboration

3.3.1. Interdisciplinary collaboration challenges—Working with experts across several scientific disciplines also presented unique challenges (Table 1). For instance, similar terms often have different meanings across health services research, clinical epidemiology, cognitive science, and linguistics. This extended into defini tional differences related to tasks and methods delegated and employed, resulting in some confusion and inefficiency. Most importantly, and central to the research objectives of ECLIPPSE, among the non-linguists in the research group there was a lack of understanding of the methodologic difference between measuring the sophistication of patients' SMs (HL) vs. measuring linguistic complexity of physicians' SMs (readability). This led to debates about the value of creating two separate indices vs one common index to allow comparison between patient HL and physician CP on the same scale. Another critical trans-disciplinary related to different interpretations of the real-world significance of certain findings, and concerns about research integrity or rigor. Balancing development of methods that were optimized using available data versus developing methods that were easily adaptable to a wider range of settings, i.e., transportability, also posed a challenge. This was especially important when trying to manage and come to consensus on research and publication priorities.

3.3.2. Interdisciplinary collaboration solutions—In order to address the challenges inherent to interdisciplinary collaboration, we employed real-time and post-hoc clarification

and documentation of term and tasks (Table 1). We also organized annual in-person, two-day meetings to ensure consistency and consensus building. Biweekly video conferences and frequent communications over email helped to speed decision making and resolve terminological discrepancies. We also found it helpful to give background and context to align objectives and clarify terminologies and discipline-specific methodologies. The cognitive scientists needed to convey the nuanced differences between measuring literacy as the ability to read as opposed to as the ability to write, as well as the difference between the constructs of literacy and linguistic sophistication, all of which are critical to understanding measurement. Some of these conversations were in effect micro-training or cross-disciplinary educational sessions. For example, through a review of the literature and delivery of mini-seminars, we gained a common understanding regarding the theory-based differences between writing, reading, literacy and readability, and arrived at consensus regarding the need to develop unique measures of LP and CP. We also were conscious about and actively discussed the complexity and multifaceted nature of health constructs for those not in the medical field, which was particularly relevant when attempting to reconcile differences in models' predictive powers in health research vs. research in other disciplines. Finally, while some tasks required more negotiation, what was essential was the clear and frequent delineation of study priorities by returning to the aims of the grant and reviewing the strategy of applying computational linguistic methods to health-related outcomes. This was helpful in mitigating the tensions between the theoretical vs. applied aspects of the project.

4. Discussion

The NLP and ML strategies developed in ECLIPPSE have yielded novel high-throughput measures that can assess components of patient HL and physician linguistic complexity by analyzing written (email) messages exchanged between patients and their healthcare providers [1,2,4]. In our effort to create a generally applicable and accurate set of tools, we tested multiple linguistic analysis tools and strategies. To increase replicability of our approaches and methods, it was critical that we outline our challenges and describe our attempts to devise and implement solutions to these challenges.

Expected challenges, such as missing linguistic structural markers or the existence of text noise (e.g., clinician signatures, hyperlinks, etc.), were mostly a part of the data mining process, but nonetheless required creative solutions. This was most evident when dealing with SMs written by hidden proxy caregivers. Those challenges that were more unique to the process of assessing patient HL and physician linguistic complexity arose in the analysis phase (e.g., threshold decisions, rater selection, and training, etc.). Researchers in this field may find the articulation and resolution of these challenges to be particularly helpful, providing opportunities to act preemptively. In addition, due in part to the complexity of the construct of HL itself, overcoming problems inherent to HL measurement will likely benefit from coordination between experts in multiple disciplinary domains, the evolution of traditional tools for new and growingly sophisticated tasks, and the adaptation of methods from other disciplines for a new purpose. Those interested in engaging in interdisciplinary work in this field may also benefit from our explication of the challenges related to such

collaboration and the processes we applied to facilitate and optimize our interdisciplinary research.

Harnessing written content from the patient portal to address HL and make progress in lowering HL demands of healthcare delivery systems is a novel approach. Using qualitative methods, Alpert et. al. applied the Centers for Disease Control and Prevention's Clear Communication Index to a patient portal to identify opportunities for better patient communication and engagement [46]. In addition to describing the painstaking nature of their work, they noted the limitation of applying a single index to one portal system, which limits both its robustness as well as its translation to other clinical web portals. We recognized this challenge in our own work and attempted to address it by using NLP to broaden the diversity of lexical and syntactic indices combined with machine learning techniques to predict LP and CP. However, faced with hundreds of indices related to literacy and text difficulty, we employed standard statistical methods to reduce the number of indices combined with empirically and theoretically motivated decisions. Employing a greater diversity of linguistic tools and features, while enhancing processing efficiency and comprehensiveness, created different analytic challenges (e.g., word/character processing limits, skewed results, etc.). Finding workable solutions was critical to moving forward and was a direct result of different ideas and approaches emerging from our interdisciplinary collaboration. It is through these collaborative empirical approaches that we gained a common understanding regarding the theory-based differences between writing, reading, literacy and readability, which helped us arrive at a consensus regarding approaches to the development of novel measures of LP and CP.

Though our interdisciplinary collaborations were essential in devising solutions during this research, the collaboration process itself was not without challenges that required resolution. It is our hope that the transparency and detailed description that we provide regarding our interdisciplinary collaboration challenges and related solutions will encourage other researchers to engage with their colleagues from other disciplines. In particular, we highlight the importance of arriving at consensus regarding shared research goals and associated terminologies from the start of the study and continuing to ensure shared understanding over time. Using tools to promote collaboration was critical for our process and should be considered as soon as one engages in the research developmental process. Agreeing on collaborative needs, desires, and methods of communication may prevent various points of confusion during the conduct of the research. Doing so may enable interdisciplinary researchers to effectively navigate familiar barriers in communication, prioritization, definitions, and subjective differences in rigor.

4.1. Future work

First, to determine if the work we have carried out so far has merit beyond what we have already described, we currently are examining approaches to measure discordance between patient LP and physician CP and determine whether discordance is associated with communication-sensitive outcomes. Second, we are in the midst of developing and testing an automated feedback tool that can be deployed in real time as physicians compose their SMs to patients, so as to promote linguistic concordance for lower HL patients. Third, we plan on

comparing content of SM exchanges that are concordant vs. discordant, using qualitative methods, so as to identify whether the former demonstrate greater interactivity and linguistic evidence of "shared meaning." [1] Fourth, we plan to examine whether patients who rely on proxy caregivers, compared with matched samples who do not, have different patterns of communication with their clinicians, and explore whether proxy use is associated with differential health outcomes. These findings will have implications for how proxies are valued by healthcare systems. Fifth, given the fact that writing SMs is inherently a literacy-related task and demand, future work should examine the impacts of integrating speech-to-text technology into patient portals, and the effects of using audiovisual content, as well as testing other health system interventions that apply these new measures. Finally, insofar as HL disproportionately affects populations of lower socioeconomic status and racial and ethnicity minority subgroups, future work - to be carried out before widespread application of our new measures – should consider methods for addressing culturally specific terminology that play a significant role in communication and, by extension, HL. Since conventional literacy assessments are bounded by cultural and linguistic assumptions derived from the dominant, majority population, more research is needed to assess patient HL in a comprehensive, holistic, and unbiased manner, and to expand the assessment of reliability and validity across sub-groups of interest in order to avoid misattributing health disparities solely to limited HL. Given the broad ethnic diversity of our sample, we currently are examining the performance and predictive validity of the LP across education level and race/ethnicity.

4.2. Limitations

This evaluation of the challenges and solutions faced when creating automated measures of communicative skills using a large, health-related linguistic corpus was conducted internally by members of the research team, which raises the possibility that subjective experiences and existing team dynamics may have influenced the degree to which our findings reflect reality. Hiring an investigator with expertise in methodologic evaluation who had no prior exposure to the research was beyond the scope of our project. We believe we minimized bias by (a) having the study lead by an investigator (WB) who is both an expert in biomedical informatics as well as a team member who joined the project late in year 2 and did not have primary responsibilities related to the ECLIPPSE Project's main deliverables; and (b) reviewing and coding nearly 100 project records that documented study processes – including challenges and solutions – across all five years. However, since this paper was conceptualized *post hoc*, four years into the ECLIPPSE Project, the documents that were reviewed were not developed, organized, or preserved in such a way as to systematically enable a recount of challenges and solutions. Thus, some documents may not have fully reflected all relevant challenges in the study. Similarly, the researchers of this study were asked to provide a retrospective account of their experience with certain challenges, and their solutions for attempting to resolve those challenges. As with any qualitative study that involves a retrospective account, there is a possibility of recall bias. Relatedly, the coding of the challenges and solutions into broader categories reflects how the two raters interpreted the materials provided based on their unique perspectives. However, we attempted to reduce any variation by employing a consensus process between the two coders, and also by

providing opportunities for the entire research team to comment and suggest revisions to our codes and categories within each domain in an ongoing fashion.

5. Conclusions

Characterization of patient HL and development of physician linguistic complexity profiles that can be automated and scaled required interdisciplinary collaboration, and our experience can inform future efforts by other groups. Interdisciplinary collaboration demands ongoing attention to reconcile differences in mental models, research methods, and meaning derived from analyses. Failure to attend to such differences can lead to research inefficiencies and an inability to answer important research questions in biomedical informatics. Agreeing on a set of research goals, terminology, and selection of collaboration tools that are available to all team members should be determined and agreed upon from the outset.

Developing novel NLP algorithms for the classification of patient HL and physician linguistic complexity requires multiple iterations and variations. When harnessing a large email dataset, identification of appropriate corpora should involve a pragmatic selection of specific and relevant patient and provider cohort and associated messages. Signifi cant attention must be paid to data cleaning to enable large scale analyses of secure message exchanges derived from electronic patient portals. Careful selection of linguistic indices is essential and should be based on theory related to the research question. Validation of new measures generated though natural language processing and machine learning requires multiple approaches. Data parsing methods should be high-throughput and extensible. Multiple analyses should be expected and even encouraged for cross-validation and verification of results. Various analyses should be systematic and clearly defined. Team management requires multiple communication methods to facilitate open exchange of ideas and the development of common understandings and consensus development. Employing an iterative process – to define and redefine terms; track changes in study design and execution; and interpret and reconcile differences in the significance of findings between linguistics and health services research fields – can help resolve interdisciplinary challenges that arise when creating and executing NLP and ML architectures and programming processes.

Acknowledgments

This work was part of a larger parent study, ECLIPPSE, funded by the NIH/NLM (R01LM012355). This work was also supported by “The Health Delivery Systems Center for Diabetes Translational Research (CDTR),” funded by the NIH/NIDDK (P30DK092924). William Brown III was also supported by the NIH/NLM (R01LM013045), the Agency for Healthcare Research and Quality (K12HS026383), and the National Center for Advancing Translational Sciences of the NIH (KL2TR001870) throughout various parts of the research and writing process. The content is solely the responsibility of the authors and does not necessarily represent the official views of NLM, NIDDK, AHRQ or the NIH. We would like to acknowledge Dr. Aaron Likens and Dr. Jianmin Dai for the programming work they did for the ECLIPPSE study.

References

[1]. Schillinger D, McNamara D, Crossley S, Lyles C, Moffet HH, Sarkar U, et al., The next frontier in communication and the ECLIPPSE study: bridging the linguistic divide in secure messaging, J. Diabetes Res. 2017 (2017) 1348242. [PubMed: 28265579]

[2]. Balyan R, Crossley SA, Brown W, Karter AJ, McNamara DS, Liu JY, et al., Using natural language processing and machine learning to classify health literacy from secure messages: The ECLIPPSE study, PLoS ONE 14 (2) (2019), e0212488. [PubMed: 30794616]

[3]. Semere W, Crossley S, Karter AJ, Lyles CR, Brown W, Reed M, et al., Secure messaging with physicians by proxies for patients with diabetes: findings from the ECLIPPSE study, J. Gen. Intern. Med. 34 (11) (2019 11) 2490–2496. [PubMed: 31428986]

[4]. Crossley SA, Balyan R, Liu J, Karter AJ, McNamara D, Schillinger D, Developing and testing automatic models of patient communicative health literacy using linguistic features: findings from the ECLIPPSE study, Health Commun. (2020 3 2) 1–11.

[5]. Crossley S, Balyan R, Karter AJ, Liu J, McNamara DS, Schillinger D, Predicting the Readability of Physicians' Secure Messages to Improve Health Communication Using Novel Linguistic Features: The ECLIPPSE Study, J Commun Healthc, 2020, In Press.

[6]. Schillinger D, Balyan R, Crossley SA, McNamara DS, Liu JY, Karter AJ, Employing computational linguistics techniques to identify limited patient health literacy: Findings from the ECLIPPSE study, Health Serv. Res. (2020). 9 23.

[7]. Ratanawongsa N, Karter AJ, Parker MM, Lyles CR, Heisler M, Moffet HH, et al., Communication and medication refill adherence: the Diabetes Study of Northern California, Feb 11, JAMA Intern. Med. 173 (3) (2013) 210–218. [PubMed: 23277199]

[8]. Chew LD, Griffin JM, Partin MR, Noorbaloochi S, Grill JP, Snyder A, et al., Validation of screening questions for limited health literacy in a large VA outpatient population, J. Gen. Intern. Med. 23 (5) (2008 5) 561–566. [PubMed: 18335281]

[9]. Moffet HH, Adler N, Schillinger D, Ahmed AT, Laraia B, Selby JV, et al., 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, Int. J. Epidemiol. 38 (1) (2009 2) 38–47. [PubMed: 18326513]

[10]. Cemballi AG, Karter AJ, Schillinger D, Liu JY, McNamara DS, Brown W, et al., Descriptive examination of secure messaging in a longitudinal cohort of diabetes patients in the ECLIPPSE study, J. Am. Med. Inform. Assoc. JAMIA. 2020 11 24.

[11]. Hudson S, Rikard RV, Staiculescu I, Edison K, Improving health and the bottom line: the case for health literacy [Internet], National Academies Press (US) (2018) [cited 2020 Feb 18]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK518850/.

[12]. Doubova SV, Infante C, Villagrana-Gutiérrez GL, Martínez-Vega IP, Pérez-Cuevas R, Adequate health literacy is associated with better health outcomes in people with type 2 diabetes in Mexico, Psychol. Health Med. 24 (7) (2019) 853–865. [PubMed: 30706719]

[13]. Yilmazel G, Health literacy, mammogram awareness and screening among tertiary hospital women patients, J. Cancer Educ. Off. J. Am. Assoc. Cancer Educ. 33 (1) (2018) 89–94.

[14]. Mazor KM, Williams AE, Roblin DW, Gaglio B, Cutrona SL, Costanza ME, et al., Health literacy and pap testing in insured women, J. Cancer Educ. Off. J. Am. Assoc. Cancer Educ. 29 (4) (2014 12) 698–701.

[15]. Castro-Sanchez E,Vila-Candel ´R, Soriano-Vidal FJ, Navarro-Illana E, Díez-Domingo J, Influence of health literacy on acceptance of influenza and pertussis vaccinations: a cross-sectional study among Spanish pregnant women, BMJ Open [Internet] (2018). 7 1 [cited 2020 Feb 18];8(7). Available from: https://bmjopen.bmj.com/content/8/7/e022132.

[16]. Literacy I of M (US) C on H, Nielsen-Bohlman L, Panzer AM, Kindig DA. The Extent and Associations of Limited Health Literacy [Internet]. National Academies Press (US); 2004 [cited 2020 Feb 18]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK216036/.

[17]. Halladay JR, Donahue KE, Cené CW, Li Q, Cummings DM, Hinderliter AL, et al., The association of health literacy and blood pressure reduction in a cohort of patients with hypertension: the heart healthy lenoir trial, Patient Educ. Couns. 100 (3) (2017 3) 542–549. [PubMed: 27776790]

[18]. Olesen K, Reynheim ALF, Joensen L, Ridderstråle M, Kayser L, Maindal HT, et al., Higher health literacy is associated with better glycemic control in adults with type 1 diabetes: a cohort study among 1399 Danes, BMJ Open Diabetes Res. Care [Internet] (2017). 8 1 [cited 2020 Feb 18];5(1). Available from: https://drc.bmj.com/content/5/1/e000437.

[19]. Krishnan S, Rohman A, Welter J, Dozor AJ, Relationship between health literacy in parents and asthma control in their children: a prospective study in a diverse suburban population, Pediatr Allergy Immunol Pulmonol. 31 (4) (2018 12 1) 221–225.

[20]. Navarra A-M, Neu N, Toussi S, Nelson J, Larson EL, Health Literacy and adherence to antiretroviral therapy among HIV-infected youth, J. Assoc. Nurses AIDS Care JANAC, 2013 2 21.

[21]. Tique JA, Howard LM, Gaveta S, Sidat M, Rothman RL, Vermund SH, et al., Measuring health literacy among adults with HIV infection in mozambique: development and validation of the HIV literacy test, AIDS Behav. 21 (3) (2017) 822–832. [PubMed: 26961538]

[22]. Bauer AM, Schillinger D, Parker MM, Katon W, Adler N, Adams AS, et al., Health literacy and antidepressant medication adherence among adults with diabetes: the diabetes study of Northern California (DISTANCE), J. Gen. Intern. Med. 28 (9) (2013 9) 1181–1187. [PubMed: 23512335]

[23]. Karter AJ, Subramanian U, Saha C, Crosson JC, Parker MM, Swain BE, et al., Barriers to insulin initiation: the translating research into action for diabetes insulin starts project, Diabetes Care 33 (4) (2010 4) 733–735. [PubMed: 20086256]

[24]. van der Heide I, Poureslami I, Mitic W, Shum J, Rootman I, FitzGerald JM, Health literacy in chronic disease management: a matter of interaction, J. Clin. Epidemiol. 102 (2018) 134–138. [PubMed: 29793001]

[25]. Schillinger D, Grumbach K, Piette J, Wang F, Osmond D, Daher C, et al., Association of health literacy with diabetes outcomes, JAMA 288 (4) (2002 7 24) 475–482. [PubMed: 12132978]

[26]. Haun JN, Patel NR, French DD, Campbell RR, Bradham DD, Lapcevic WA, Association between health literacy and medical care costs in an integrated healthcare system: a regional population based study, BMC Health Serv. Res. [Internet] (2015). 6 27 [cited 2020 Feb 18];15. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4482196/.

[27]. Seligman HK, Wang FF, Palacios JL, Wilson CC, Daher C, Piette JD, et al., Physician notification of their diabetes patients' limited health literacy. A randomized, controlled trial, J. Gen. Intern. Med. 20 (11) (2005 11) 1001–1007. [PubMed: 16307624]

[28]. DeWalt DA, Schillinger D, Ruo B, Bibbins-Domingo K, Baker DW, Holmes GM, et al., Multisite randomized trial of a single-session versus multisession literacy-sensitive self-care intervention for patients with heart failure, Circulation 125 (23) (2012 6 12) 2854–2862. [PubMed: 22572916]

[29]. Karter AJ, Parker MM, Duru OK, Schillinger D, Adler NE, Moffet HH, et al., Impact of a pharmacy benefit change on new use of mail order pharmacy among diabetes patients: the diabetes study of Northern California (DISTANCE), Health Serv. Res 50 (2) (2015 4) 537–559. [PubMed: 25131156]

[30]. Brach C, Keller D, Hernandez LM, Baur C, Parker R, Dreyer B, et al., Ten attributes of health literate health care organizations, NAM Perspect [Internet] (2012). 6 19 [cited 2020 Apr 9]; Available from: https://nam.edu/perspectives-2012-ten-attributes-of-health-literate-health-care-organizations/.

[31]. Allen L, Dascalu M, McNamara DS, Crossly S, Trausan-Matu S, Modeling individual differences among writers using readerbench, in: EDULEARN16 Proceedings: 8th International Conference on Education and New Learning Technologies [Internet]. IATED Academy; 2016 [cited 2020 Oct 26]. p. 5269–79. Available from: https://research.ou.nl/en/publications/modeling-individual-differences-among-writers-using-readerbench.

[32]. Allen LK, Snow EL, Crossley SA, Jackson GT, McNamara D, Reading comprehension components and their relation to writing, Annee Psychol. 114 (4) (2014 12 1) 663–691.

[33]. Crossley SA, Allen LK, Snow EL, McNamara DS, Incorporating learning characteristics into automatic essay scoring models: what individual differences and linguistic features tell us about writing quality, J. Educ. Data Min. 8 (2) (2016) 1–19.

[34]. Schoonen R, Are reading and writing building on the same skills? the relationship between reading and writing in L1 and EFL, Read Writ Interdiscip J 32 (3) (2019 3) 511–535.

[35]. Castro CM, Wilson C, Wang F, Schillinger D, Babel babble: physicians' use of unclarified medical jargon with patients, Am. J. Health Behav. 31 (Suppl 1) (2007 10) S85–S95. [PubMed: 17931142]

[36]. Schillinger D, Bindman A, Wang F, Stewart A, Piette J, Functional health literacy and the quality of physician-patient communication among diabetes patients, Patient Educ. Couns. 52 (3) (2004 3) 315–323. [PubMed: 14998602]

[37]. Schillinger D, Piette J, Grumbach K, Wang F, Wilson C, Daher C, et al., Closing the loop: physician communication with diabetic patients who have low health literacy, Arch. Intern. Med. 163 (1) (2003 1 13) 83–90. [PubMed: 12523921]

[38]. Grabeel KL, Russomanno J, Oelschlegel S, Tester E, Heidel RE, Computerized versus hand-scored health literacy tools: a comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in printed patient education materials, J. Med. Libr. Assoc. JMLA 106 (1) (2018 1) 38–45. [PubMed: 29339932]

[39]. Institute of Medicine (US) Committee on Health Literacy. Health Literacy: A Prescription to End Confusion [Internet]. Nielsen-Bohlman L, Panzer AM, Kindig DA, editors. Washington (DC): National Academies Press (US); 2004 [cited 2020 Apr 9]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK216032/.

[40]. Kim H, Goryachev S, Rosemblat G, Browne A, Keselman A, Zeng-Treitler Q, Beyond surface characteristics: a new health text-specific readability measurement, AMIA Annu. Symp. Proc. 2007 (2007) 418–422.

[41]. Oliffe M, Thompson E, Johnston J, Freeman D, Bagga H, Wong PKK, Assessing the readability and patient comprehension of rheumatology medicine information sheets: a cross-sectional Health Literacy Study, BMJ Open [Internet] (2019). 2 5 [cited 2020 Apr 9];9(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6377552/.

[42]. Asana. Manage your team's work, projects, & tasks online · Asana [Internet]. Asana. [cited 2020 Apr 6]. Available from: https://asana.com/.

[43]. Sarkar U, Karter AJ, Liu JY, Moffet HH, Adler NE, Schillinger D, Hypoglycemia is more common among type 2 diabetes patients with limited health literacy: the Diabetes Study of Northern California (DISTANCE), J. Gen. Intern. Med. 25 (9) (2010 9) 962–968. [PubMed: 20480249]

[44]. Crossley S, Kostyuk V, Letting the genie out of the lamp: using natural language processing tools to predict math performance, in: Gracia J, Bond F, McCrae JP, Buitelaar P, Chiarcos C, Hellmann S (Eds.), Language, Data, and Knowledge. Cham: Springer International Publishing; 2017. p. 330–42. (Lecture Notes in Computer Science).

[45]. Crossley S, Paquette L, Dascalu M, McNamara DS, Baker RS, Combining click-stream data with NLP tools to better understand MOOC completion, in: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge [Internet], Edinburgh, United Kingdom: Association for Computing Machinery; 2016 [cited 2020 Apr 7]. p. 6–14. (LAK '16). Available from: 10.1145/2883851.2883931.

[46]. Alpert JM, Desens L, Krist AH, Aycock RA, Kreps GL, Measuring health literacy levels of a patient portal using the CDC's clear communication index, Health Promot Pract. 18 (1) (2017 1) 140–149. [PubMed: 27188894]

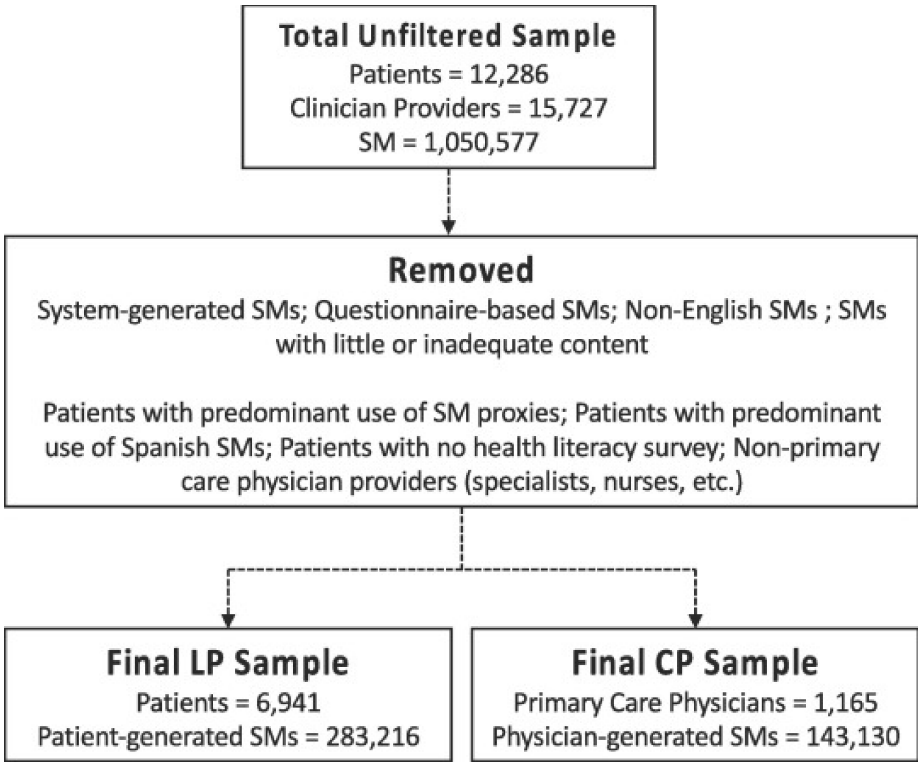

Fig. 1.
Patient, physician and secure message samples flowchart.

J Biomed Inform. Author manuscript; available in PMC 2021 June 08.

Table 1

Challenges and solutions matrix for patient literacy profile (LP) and physician complexity profile (CP).

Process	Challenge(s)	Solution(s)	LP	CP
Data mining	Formal linguistic structural markers (i.e., punctuations, new line characters, sentence markers) were missing, particularly because the corpus involved email/text-like messages.	We extracted and analyzed the data in such a way that some of these linguistic structural markers could be retrieved; others of these markers were still missing and could not be handled.	X	X
	Website URLs, Hyperlinks, lab codes, page numbers were present in the SMs.	We developed algorithm that removed some of these features.	X	X
	Automated (pre-written stock) and physician auto signature text was sometimes contained in physician messages.	Auto text was retained in the corpus because it proved impossible to create an NLP method that could reliably identify these “smart texts,” phrases or signatures automatically. We also elected to retain these automated texts because they are representative of the language used by physicians when messaging patients.		X
	Some SMs were from patients and others were from their proxies (formal and informal, or hidden).	We identified SMs written by formally registered proxies and removed them from the corpus. Created an algorithm that predicted SMs written by hidden proxies, which were removed from the corpus based on a threshold derived from additional research [3].	X	
	There were some non-English (Spanish) messages.	A program was created that identified non-English (Spanish) text; SMs with 50% or more Spanish text were removed from the corpus.	X	X
	There were no clear distinctions between individual messages and threads, precluding detailed analysis of individual SMs.	All the SMs from a patient or a physician were aggregated into a single corpus for creating LPs and CPs. For work that compared LPs and CPs within patient-physician dyads, aggregation occurred at the level of the dyad.	X	X
	Brief SMs could not be linguistically analyzed by our ML algorithms.	We created a cutoff whereby SMs <50 words were excluded.	X	X
	When running linguistic tools some indices would occasionally cause the program to stop functioning due to parser problems.	We noted the stop point identified by the tool and excised the short section that caused the offense; the tool was re-run from that point on	X	
	Excess text length made human ratings of physician SMs difficult. We needed to standardize the human rating process.	Physicians SM threads were randomly trimmed to contain approximately ~ 300 words. No individual SMs contained in the threads were truncated.		X
Analysis	Application of tools for extraction of linguistic indices was challenging for many reasons, including the nature of the corpus being email exchanges rather than written prose and because of the large number of indices available related to literacy and text difficulty. The SMs content also contained test results and lab reports that did not return meaningful linguistic features.	Human intervention was frequently necessary to run the NLP tools and eventually to extract meaningful data for the large corpus and/large number of linguistic indices. Several combinations of linguistic indices, starting with the most basic and commonly used to more complex and theory-driven combinations were tested.	X	X
	While developing LPs to assess patients’ HL, we found that the resulting data were imbalanced/skewed.	We adjusted the default thresholds to create more balanced data in line with distribution of previously published HL research. Implemented and compared different sampling techniques; ensemble approaches; assigned different weight/cost to misclassification; tested new ML techniques explicitly designed for imbalanced data.	X	
	There was a lack of “gold standards” to assess the content validity/performance of the HL measure (patient LP) and the linguistic complexity measure (physician CP).	In addition to using self-reported HL, we created a novel measure of expert-rated HL wherein literacy and health experts reviewed a purposive sub-sample of patient SMs using a Likert scale to classify their HL. We also created a parallel expert-rated linguistic complexity measure to serve as a gold standard for the physician CP and applied it to the same sub-sample [2,5].	X	X
	When identifying physician CP scores some of the ML scores were biased towards the class that had a higher number of instances as compared to the class having fewer instances.	We raised the probability threshold of our ML algorithm for the class with more instances and lowered the probability threshold for the class with fewer instances. We also refined our expert rating scoring system.		X

Process	Challenge(s)	Solution(s)	LP	CP
	Identification of appropriate ML techniques among many options of varying complexity was challenging.	We selected the two simpler ML techniques, rather than those that were complex, because they yielded similar results.	X	X
Interdisciplinary collaboration	Definitional differences of tasks and methods were confusing.	We implemented real-time clarification and documentation of terms whenever possible.	X	X
	It was difficult communicating domain-specific constructs across disciplines.	We carried out frequent video conference communications, and communications over email or other forms of text and verbal communication. Shared relevant disciplinary literature.	X	X
	There was a lack of understanding of the methodologic difference between measuring the sophistication of patients' SMs (HL) vs. measuring linguistic complexity of physicians' SMs (readability).	We arrived at consensus regarding the theory-based differences between writing, reading, literacy and readability through a review of the literature and mini-seminars.	X	X
	There was confusion regarding when steps were completed and/or when tasks were ready to be handed over.	We frequently refined terminology and refined or reinforced decisions previously reached.	X	X
	There was use of similarly sounding or spelled words that have different meanings (Homonymy) across scientific disciplines.	We clarified and documented meanings of terms in real time.	X	X
	There were research priority differences.	We negotiated and defined overall study objectives prioritized team needs over individual team member/or sub-team priorities.	X	X
	There were subjective differences in "scientific rigor".	We engaged in biweekly video and annual in-person conference communication for group consensus of appropriate definitions, thresholds and methods to achieve scientific rigor, with a particular focus on differences between explanation of variance in computational linguistics vs. health services research. Shared relevant disciplinary literature.	X	X