Proceedings of the First workshop on Language Technologies for Government and Public Admnistration (LT4Gov), pages 1–6
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Development of Natural Language Processing Tools to Support Determination of Federal Disability Benefits in the U.S.

Bart Desmet[1]* **Julia Porcino**[1]* **Ayah Zirikly**[1]*
Denis Newman-Griffis[1,2] **Guy Divita**[1] **Elizabeth Rasch**[1]
[1]Rehabilitation Medicine Dept., Clinical Center, National Institutes of Health, Bethesda, MD
[2]Dept. of Computer Science and Engineering, The Ohio State University, Columbus, OH
{bart.desmet, julia.porcino, ayah.zirikly, denis.griffis, guy.divita, elizabeth.rasch}@nih.gov

Abstract

The disability benefits programs administered by the US Social Security Administration (SSA) receive between 2 and 3 million new applications each year. Adjudicators manually review hundreds of evidence pages per case to determine eligibility based on financial, medical, and functional criteria. Natural Language Processing (NLP) technology is uniquely suited to support this adjudication work and is a critical component of an ongoing inter-agency collaboration between SSA and the National Institutes of Health. This NLP work provides resources and models for document ranking, named entity recognition, and terminology extraction in order to automatically identify documents and reports pertinent to a case, and to allow adjudicators to search for and locate desired information quickly. In this paper, we describe our vision for how NLP can impact SSA's adjudication process, present the resources and models that have been developed, and discuss some of the benefits and challenges in working with large-scale government data, and its specific properties in the functional domain.

Keywords: disability, health, machine learning, NLP, information extraction

1. Introduction

The United States Social Security Administration (SSA) administers the largest federal programs for disability benefits in the US, serving over 15 million individuals (SSA Office of the Chief Actuary, 2019b; Social Security Administration, 2019). The SSA programs provide benefits to those individuals who are unable "to engage in any substantial gainful activity by reason of any medically determinable physical or mental impairment(s) which can be expected to result in death or which has lasted or can be expected to last for a continuous period of not less than 12 months" (Social Security Administration, 2012).

In order to determine whether an individual meets this definition of disability, SSA uses a five step process, illustrated in Figure 1. The first step is used to determine whether the individual meets the financial eligibility criteria. The second step looks at whether the applicant's alleged impairments are sufficiently severe. The third step evaluates whether the applicant meets certain medical criteria. If these criteria are met, the applicant will receive benefits. Otherwise, the case proceeds to the fourth and fifth steps, where SSA considers the individual's remaining functional capacity and the ability to work. Thus, both medical and functional information are critical to SSA's business process. To gather this information, adjudicators solicit medical records from the applicant's medical providers. This often results in hundreds or even thousands of pages of medical records for a single applicant, which the adjudicator must review manually to determine whether there is sufficient evidence to make a determination. This business process is further strained by the volume of applications – approximately 2 to 3 million new applications each year – and an aging work force where greater numbers of adjudicators will be retiring (SSA Office of the Chief Actuary, 2019a; United States Government Accountability Office, 2018).

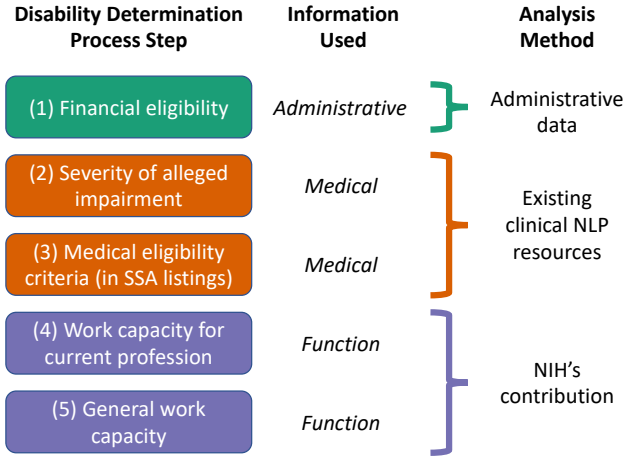

Figure 1: Illustration of the SSA disability determination process, indicating the primary type of information used at each step and relevant analytic methods.

In an effort to manage these challenges and better support adjudicators, the SSA has invested in developing natural language processing (NLP) systems for efficiently processing medical records. In addition, the SSA has recognized the importance of engaging external domain experts in order to introduce new perspectives and address key challenges. Through an inter-agency agreement with the National Institutes of Health (NIH), the two agencies have established a collaboration to develop novel NLP tools that particularly target information on function to help improve SSA's business process. This paper outlines the vision for these NLP tools at SSA, the current state of that vision, and what lessons have been learned.

*Equal contribution.

2. Vision for NLP in Disability Determination

The introduction of NLP into SSA's business process serves two critical goals: providing decision support and building a foundation for business intelligence. Decision support includes using NLP models to quickly identify information pertinent to a case, alerting adjudicators when documents contain relevant information, as well as providing tools that allow adjudicators to search for and locate desired information. Abbott et al. (2017) discussed the use of NLP to identify severely ill applicants to the Compassionate Allowance (CAL) initiative at SSA. On the other hand, business intelligence offers case support by checking for consistency of evidence when medical records are coming from different providers and covering months or even years of medical history. Developing systems for business intelligence also allows for a more global picture of data and business processes, such as by detecting fraud and making information more readily available for research purposes. The NIH-SSA collaboration has focused on decision support, where SSA's 5-step decision process offers an opportunity to combine the expertise of the two agencies.

Steps 2 and 3 of SSA's adjudication process are primarily concerned with medical information, such as documented symptoms, diseases, and disorders. A wide variety of NLP tools have been developed for identifying this information (Kreimeyer et al., 2017), and have proven useful even for identifying rare diseases (Udelsman et al., 2019). While there are known challenges in adapting medical NLP systems to language from the diversity healthcare providers interacting with a national consumer like SSA (Carrell et al., 2017), these tools nonetheless present significant potential to reduce adjudicator burden in reviewing medical evidence.

Steps 4 and 5, however, are concerned primarily with information on physical and mental function. Function, as conceptualized in the World Health Organization's International Classification of Functioning, Disability and Health (ICF), is determined not only by medical factors, but also by environmental and personal factors, and by the activities and social roles an individual chooses to engage in (World Health Organization, 2001). Anner et al. (2012) showed that the ICF framework is effective for evaluating disability. However, functioning information poses distinct problems for NLP, including inconsistent documentation standards, a lack of ontological and terminological resources capturing functional concepts, and a paucity of available data for NLP development and analysis (Newman-Griffis et al., 2019a). NIH's expertise in conceptualization and analysis of function thus offered a synergistic opportunity to focus on developing novel tools and resources to address these challenges in capturing functioning information with NLP.

The remainder of this paper describes NIH's initial research and development of NLP technologies for functional information.

3. Implementation

For initial research and development, NIH has focused on mobility reports, one of the most frequent areas of functional limitation involved in disability cases (Courtney-Long et al., 2015). Several types of NLP technologies have been developed for both document-level and case-level support, including information extraction and document ranking technologies, as well as the automated creation of terminologies supporting identification of functioning information.

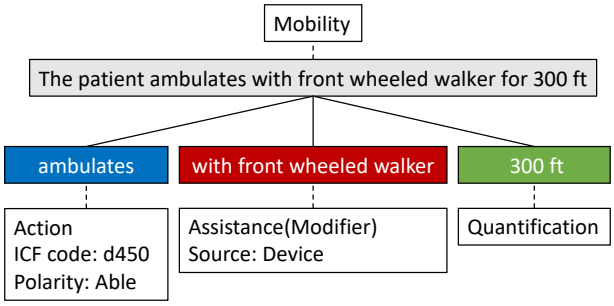

Figure 2: Annotation example of a Mobility report with subentities and attributes.

3.1. Data

Since functional information relevant to a claimant's allegations is primarily present in free text without structured codes associated with it, finding such information is a more time-consuming process for the adjudicators. In our developed models we focus on finding *activity reports* (Newman-Griffis et al., 2019a) that are relevant to a claimant's functional status. Examples of such information for mobility include `The patient is able to walk using a cane` and `The pt requires assistance to transfer from bed to chair.`

For the initial phases of research, we built our resources using data from NIH Clinical Center medical records as surrogate to SSA data. The NIH data are a rich source of information about function for terminology discovery and are often cleaner than SSA records.

A team of rehabilitation and medical experts developed schemas and guidelines for annotating mobility information. Spans of text related to a claimant's mobility status were marked in a corpus of 400 English-language physical therapy notes, provided by the Office of Biomedical Translational Research Information System (Cimino et al., 2014, BTRIS). Additional subentities and attributes were marked, as summarized in Figure 2.

Annotation results are presented in Table 1. Pairwise inter-annotator agreement as measured on a doubly-annotated set of 200 documents ranged from 96 to 98% F1 score on overlapping text spans (Thieu et al., 2017).

The resulting 400 annotated notes served as the gold standard for automatic Mobility report detection, and were randomly assigned to an 80/20 split into training and test sets.

3.2. NER Modeling

NIH introduced multiple information extraction baseline models that cast the problem as a named entity recognition (NER) task, where named entities are the functional information reports.

Type	Count	IAA (F1)
Mobility	4631	0.980
Action	4527	0.980
Assistance	2517	0.960
Quantification	2303	0.982
Score Definition	303	0.980

Table 1: Annotation results for the Mobility domain on 400 PT notes, and inter-annotator agreement on 200 doubly annotated PT notes.

As a baseline model, we used Conditional Random Fields (CRF) (Finkel et al., 2005) with an extensive list of features such as word shape, part-of-speech (POS) tags, word clusters, etc. Additionally, we test Bidirectional Long Short-Term Memory (BiLSTM-CRF) models, given their popularity and high performance in NER (Lample et al., 2016) and patient notes deidentification tasks (Dernoncourt et al., 2017). We tested both architectures to build mobility recognition models that handle the full mobility report span and its subentities. Both the CRF and Bi-LSTM-CRF models show promising results with respective token-level F1-scores of 82% and 78% for the mobility reports. Additionally, the models yield good results for subentities, with 75% and 83% token-level F1-score for *Action* mentions, which contain the most salient information for mobility-related queries.

These results are considerably lower than what NER systems typically achieve. For instance, state-of-the-art performance on the CoNLL 2003 dataset is 93.5% F1-score (Baevski et al., 2019). While this discrepancy can be partially attributed to the comparatively limited amount of training examples, we believe this is also caused by the challenging nature of the task, the large data variability and presence of noise (e.g. OCR). We refer to Newman-Griffis and Zirikly (2018) for further description and analysis of the results on a subset of the annotated reports.

To complement these modeling strategies, which yield high-precision predictions but suffer in recall, NIH also developed a recall-focused model that uses contextual information to estimate the likelihood that each token in a document is part of a mobility report (Newman-Griffis and Fosler-Lussier, 2019). This approach consistently identified over 90% of relevant tokens in NIH documents, though with an accompanying increase in false positives necessitating post review. Preliminary evaluation on SSA data has shown similar results; qualitative review of system outputs on diverse document types suggests effective generalization with only a small decrease in precision. These different strategies therefore offer useful alternatives for applications that may emphasize high-confidence predictions (e.g., document classification) or high-coverage (e.g., evidence retrieval).

3.3. Polarity Classification

Identifying relevant information is a key first step to help the adjudicators in their decision process. However, the next step in that process is providing the polarity of the functional report. For instance, given the mobility report in Figure 2, the polarity associated with the mobility action mention *ambulates* is *able*. The four polarity values in our annotation schema are *able*, *unable*, *unclear*, and *none*. Our proposed models range from rule-based systems, conventional machine learning techniques using random forests and support vector machines (SVM) to feed-forward (FF) and convolutional (CNN) neural network models. In addition we employ ensemble models that use majority voting between SVM and CNN, and a FF model that dynamically chooses output from the rule-based, SVM and CNN systems. Our proposed models predict the ability of a functional activity with 88% F1-score, as opposed to 69% for the *unable* label. This large gap in performance is mainly due to the imbalanced nature of the dataset. For further details about these models and analysis, we refer to Newman-Griffis et al. (2019b).

3.4. Document Ranking

Document-level information extraction technologies also offer an opportunity to support case-level processes, particularly document triage and prioritization. NIH has investigated using mobility reports extracted using NER models to rank a set of documents by the amount of predicted mobility information in each. These experiments yielded strong correlation with the true number of mobility reports in each document, indicating that NER technologies present significant utility for assisting case-level review of documents (Newman-Griffis and Fosler-Lussier, 2019).

3.5. Terminology Extraction

Terminologies and ontologies have been heavily developed and used for NLP in the clinical and biomedical domains. Examples of such repositories are the Unified Medical Language System (UMLS) (Bodenreider, 2004), the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) (Donnelly, 2006) and the Human Phenotype Ontology (Robinson et al., 2008). SNOMED CT terminologies, for instance, provide over 90% coverage of the commonly used terms in medical problem lists (Elkin et al., 2006).

Given the utility of terminologies and the lack of any for the functioning domain, we developed them for multiple functioning domains including Mobility. A particular challenge for building these terminologies is that relevant terms in these domains are often not medical, but highly frequent and ambiguous. As a result, they need to be captured as multi-word units that include sufficient context (e.g. `able to walk around`), and the many different surface realizations of a concept needs to be generated to increase recall. We used neural models to expand seed terminology lists to achieve a partial-match coverage of 88% against annotated data.

4. Discussion

This project to develop models and tools for functional information to support SSA's business process has provided insight into the benefits and challenges of collaborations between federal agencies. At the same time, this is only a first step in the work to improve the decision-making process. In this section, we discuss some of the implications

of collaborating across agencies, technical challenges, and future work aimed at addressing them.

4.1. Government Collaborations

The collaboration between SSA and NIH brings together expertise and knowledge across federal agencies to leverage process insights while providing new perspectives on ways to inform the disability determination process. There is a lot of work that goes into forming and maintaing such a relationship to ensure that the collaboration supports the mission of each agency and offers value to both. In particular, since SSA provides services to the American public, it is paramount that the collaboration protects the interests and privacy of those individuals who apply for benefits. In the US government, the Privacy Act protects information about individuals that is "retrieved by personal identifiers such as a name, social security number, or other identifying number or symbol" (Health and Human Services, 2019). SSA includes information about the Privacy Act as part of the disability benefits application, as well as any other form that collects information from an applicant (Social Security Administration, 1998). The Privacy Act prohibits the sharing of this information except if covered by one of twelve exceptions. These exceptions include use for research and statistical purposes, which therefore allows SSA to share these data with NIH as part of the collaborative effort to "enhance the decision-making process in the Social Security program" (Social Security Administration, 2020). While this exception allowed SSA to share these data, since the NIH is a research institute, we also sought the necessary human subjects' protection determinations for accessing and conducting research with the data. By leveraging the regulation processes across both agencies, we ensure that the necessary checks and balances are in place for protecting the data and the individuals the agencies serve.

4.2. Technical Challenges of SSA Data

While having access to these data is critical in order to develop systems that best suit SSA's business process, working with SSA records poses many challenges. SSA collects and generates enormous amounts of data for each applicant, and these data are often heterogeneous, noisy and fluid. Applicants' data include medical records from across the country and from all kinds of providers. Such a geographically diverse set of documents, with regional differences in use of language, and the evolution of language and medical jargon over time pose additional hurdles for developing NLP models.

Finding function information within this corpus inherently comes with challenges posed by the genre, where the terminology is under-specified and telegraphic at best, and text is often semi-structured. These properties exacerbate problems of scoping and ambiguity inherent in natural language, and make the genre resistant to traditional NLP techniques. Figure 3 illustrates these challenges with an example from the function domain. *Range of motion* (*ROM*), *within functional limits* (*WFL*) and *external rotation [strength]* (*ER*) are examples of telegraphic and ambiguous terminology. The example also contains two slot and value structures, for *ROM* and *Strength*. Strength observations are not enumerated (*all extremities*), and the shorthand *10/10 for all except* presents scoping issues, as it modifies the truth propositions from the previous statement. Improvements to any of these issues in the function domain are applicable more broadly. To that end, we are building systems to address scoping and decompose structured text using function as the use case.

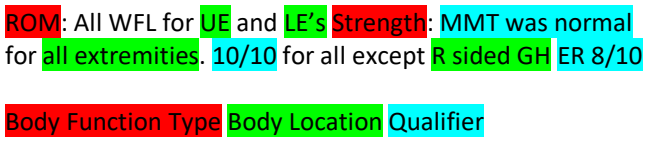

Figure 3: Example of terminological and structural ambiguity from the function domain.

5. Future Work

In ongoing work, we are developing classification models for other functional domains, tuning and validating them on SSA data, and supporting their integration at SSA.

5.1. From Demonstration to Deployment

Translating novel innovations in informatics research into operational practice in health systems faces a wide variety of challenges (Goldstein et al., 2004; Scott et al., 2018). A key challenge posed by current technologies lies in translating software designed for research and demonstration, which must be easily modifiable and typically focuses on small, controlled datasets, into products ready for enterprise-level deployment, demanding much greater robustness and the ability to process large-scale data rapidly. In NLP, two primary factors limit this translation: computational requirements and engineering environments. Cutting-edge technologies such as BERT (Devlin et al., 2019) require GPU capability for effective use, and present high demands for disk space and memory in processing and storing results; this imposes significant burden in procuring and maintaining sufficient computational resources to support the tools used. In addition, many current deep learning technologies use libraries implemented in the Python programming language, whereas Java is often the language of choice in secure government and enterprise environments, and for many medical NLP tools designed for large-scale use. Deployment might therefore necessitate re-implementation or interoperability layers.

6. Conclusion

Disability benefits case adjudication is an area of government functioning where human language technologies have the potential to improve service quality and cut costs. In an effort to address challenges with adjudicator case load, the US Social Security Administration is pursuing NLP solutions and reaching out to external partners with domain expertise that can help address the most challenging components. The SSA-NIH inter-agency agreement has been a success in bringing together experts from multiple domains, defining a modern vision and delivering tangible results that can improve SSA's business processes.

Acknowledgments

The authors would like to thank Pei-Shu Ho and Jonathan Camacho Maldonado for their annotation efforts. This research was supported in part by the Intramural Research Program of the National Institutes of Health, Clinical Research Center and through an Inter-Agency Agreement with the US Social Security Administration.

Bibliographical References

Abbott, K., Ho, Y.-Y., and Erickson, J. (2017). Automatic health record review to help prioritize gravely ill social security disability applicants. *Journal of the American Medical Informatics Association*, 24(4):709–716.

Anner, J., Schwegler, U., Kunz, R., Trezzini, B., and de Boer, W. (2012). Evaluation of work disability and the international classification of functioning, disability and health: what to expect and what not. *BMC Public Health*, 12(1):470.

Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., and Auli, M. (2019). Cloze-driven pretraining of self-attention networks. *arXiv preprint arXiv:1903.07785*.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Carrell, D. S., Schoen, R. E., Leffler, D. A., Morris, M., Rose, S., Baer, A., Crockett, S. D., Gourevitch, R. A., Dean, K. M., and Mehrotra, A. (2017). Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *Journal of the American Medical Informatics Association*, 24(5):986–991, sep.

Cimino, J. J., Ayres, E. J., Remennik, L., Rath, S., Freedman, R., Beri, A., Chen, Y., and Huser, V. (2014). The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): design, contents, functionality and experience to date. *Journal of Biomedical Informatics*, 52:11–27.

Courtney-Long, E. A., Carroll, D. D., Zhang, Q. C., Stevens, A. C., Griffin-Blake, S., Armour, B. S., and Campbell, V. A. (2015). Prevalence of disability and disability type among adults – United States, 2013. *MMWR. Morbidity and mortality weekly report*, 64(29):777–783, jul.

Dernoncourt, F., Lee, J. Y., and Szolovits, P. (2017). NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *arXiv preprint arXiv:1705.05487*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121:279.

Elkin, P. L., Brown, S. H., Husser, C. S., Bauer, B. A., Wahner-Roedler, D., Rosenbloom, S. T., and Speroff, T. (2006). Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. In *Mayo Clinic Proceedings*, volume 81, pages 741–748. Elsevier.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Goldstein, M. K., Coleman, R. W., Tu, S. W., Shankar, R. D., O'Connor, M. J., Musen, M. A., Martins, S. B., Lavori, P. W., Shlipak, M. G., Oddone, E., Advani, A. A., Gholami, P., and Hoffman, B. B. (2004). Translating research into practice: Organizational issues in implementing automated decision support for hypertension in three medical centers. *Journal of the American Medical Informatics Association*, 11(5):368–376, 09.

Health and Human Services. (2019). The Privacy Act. `https://www.hhs.gov/foia/privacy/index.html`.

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M., and Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, 73:14 – 29.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Newman-Griffis, D. and Fosler-Lussier, E. (2019). HARE: a flexible highlighting annotator for ranking and exploration. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 85–90, Hong Kong, China, November. Association for Computational Linguistics.

Newman-Griffis, D. and Zirikly, A. (2018). Embedding transfer for low-resource medical named entity recognition: A case study on patient mobility. In *Proceedings of the BioNLP 2018 workshop*, pages 1–11, Melbourne, Australia, July. Association for Computational Linguistics.

Newman-Griffis, D., Porcino, J., Zirikly, A., Thieu, T., Camacho Maldonado, J., Ho, P.-S., Ding, M., Chan, L., and Rasch, E. (2019a). Broadening horizons: the case for capturing function and the role of health informatics in its use. *BMC Public Health*, 19(1):1288.

Newman-Griffis, D., Zirikly, A., Divita, G., and Desmet, B. (2019b). Classifying the reported ability in clinical mobility descriptions. *arXiv preprint arXiv:1906.03348*.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615.

Scott, P., Dunscombe, R., Evans, D., Mukherjee, M.,

and Wyatt, J. (2018). Learning health systems need to bridge the 'two cultures' of clinical informatics and data science. *Journal of innovation in health informatics*, 25(2):126–131, jun.

Social Security Administration. (1998). GN 03301.020 Privacy Act. https://secure.ssa.gov/apps10/poms.nsf/lnx/0203301020#b.

Social Security Administration. (2012). Disability Evaluation Under Social Security. https://www.ssa.gov/disability/professionals/bluebook/general-info.htm.

Social Security Administration. (2019). Annual report of the supplemental security income program. https://www.ssa.gov/oact/ssir/SSI19/ssi2019.pdf.

Social Security Administration. (2020). GN 03316.130 Disclosure Without Consent for Research and Statistical Purposes. https://secure.ssa.gov/apps10/poms.nsf/lnx/0203316130.

SSA Office of the Chief Actuary. (2019a). Disabled worker beneficiary statistics by calendar year, quarter, and month. https://www.ssa.gov/oact/STATS/dibStat.html (accessed: 02.12.2020).

SSA Office of the Chief Actuary. (2019b). Social Security Beneficiary Statistics. https://www.ssa.gov/oact/STATS/DIbenies.html (accessed: 02.12.2020).

Thieu, T., Camacho, J., Ho, P.-S., Porcino, J., Ding, M., Nelson, L., Rasch, E., Zhou, C., Chan, L., Brandt, D., et al. (2017). Inductive identification of functional status information and establishing a gold standard corpus: A case study on the mobility domain. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2319–2321. IEEE.

Udelsman, B., Chien, I., Ouchi, K., Brizzi, K., Tulsky, J. A., and Lindvall, C. (2019). Needle in a haystack: Natural language processing to identify serious illness. *Journal of Palliative Medicine*, 22(2):179–182. PMID: 30251922.

United States Government Accountability Office. (2018). Social Security Administration: Continuing leadership focus needed to modernize how SSA does business. https://www.gao.gov/products/gao-18-432t.

World Health Organization. (2001). *The International Classification of Functioning, Disability and Health: ICF*. World Health Organization, Geneva.