Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Jeanne Marie Powell	Date

Text as data:
Leveraging natural language processing to decode Parkinson's disease.

By
Jeanne M. Powell
Doctor of Philosophy

Graduate Division of Psychology

_____
J. Lucas McKay, Ph.D, M.S.C.R
Advisor

_____
Philip Kragel, Ph.D
Committee Member

_____
Donna Maney, Ph.D
Committee Member

_____
Hillary Rodman, Ph.D
Committee Member

_____
Abeed Sarker, Ph.D
Committee Member

Accepted:

_____
Kimberly Jacob Arriola, Ph.D, MPH
Dean of James T. Laney School of Graduate Studies

_____
Date

Text as data:
Leveraging natural language processing to decode Parkinson's disease.

By

Jeanne M. Powell
B.A., Cornell University, 2017
M.A., Emory University, 2021

Advisor: J. Lucas McKay, Ph.D., M.S.C.R.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in Psychology
2024

Abstract

Text as data:
Leveraging natural language processing to decode Parkinson's disease.
By Jeanne M. Powell

This dissertation explores the use of natural language processing (NLP) to enhance understanding of Parkinson's disease (PD) by analyzing text data from varied sources. It begins with examining academic publications to address the historical underrepresentation of women in PD research, revealing that despite NIH policies promoting sex and gender inclusivity, women's participation in studies remains disproportionately low compared to the prevalence of PD among women. To streamline the tracking of sex inclusion in research, we developed an automated tool using GPT-4-Turbo, which efficiently extracts data on sex inclusion from publications. We then address the issue of falls among PD patients by creating a classifier that categorizes falls based on patient descriptions, offering insights into biomechanical causes. This approach allows for more targeted fall prevention research. Finally, we analyze patient and caregiver experiences shared on Facebook to capture real-time discussions about disease onset, progression, and complications. This pipeline provides a direct source of patient-reported information, overcoming the limitations of traditional data collection methods. In summary, this work demonstrates NLP's capability to improve PD research and care by enhancing sex/gender representation monitoring in clinical studies, facilitate large-scale research on the prevention of falls, and extract insights from patient experiences, showcasing NLP's extensive potential in advancing research on PD.

Text as data:
Leveraging natural language processing to decode Parkinson's disease.

By

Jeanne M. Powell
B.A, Cornell University, 2017
M.A., Emory University, 2021

Advisor: J. Lucas McKay, Ph.D., M.S.C.R.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in Psychology
2024

Acknowledgements

This dissertation reflects not only my efforts but the guidance, support, and invaluable contributions of several individuals and groups whose expertise and encouragement were indispensable throughout my time here at Emory University.

I extend deep gratitude to my advisor, Dr. J. Lucas McKay, whose professionalism, focus, and exceptional leadership have significantly contributed to my growth, both professionally and personally. His guidance has imparted lessons I aim to incorporate into my future career path.

Thank you to my committee members for their invaluable insights and guidance throughout my academic journey. Special appreciation goes to Drs. Abeed Sarker and Donna Maney, who have been indispensable mentors and coauthors, enriching my research experience. I also owe a debt of gratitude to my former advisors, Drs. Patricia Jensen and Alex Ophir, whose support not only guided me to graduate school but also sustained me throughout my academic journey. I thank my research assistants—Sarah Palmer, Catherine Jones, Grey Skelton, Imari Genias, Anissa Patel, and Sofia Hefner—and collaborators, including Dr. Christine Esper, Dr. Richa Triapthi, Sahithi Lakamana, Mia White, and Yuting Guo, for their diligent work.

Thank you to my family—my mother Mary, sister Catherine, and in memory of my father John—and to my partner, Andy, for their unwavering support throughout my education. I also extend my gratitude to my friends, both near and far, for creating a supportive network that has made Atlanta not just a place of study, but a home.

Lastly, a heartfelt thank you to the individuals with Parkinson's disease and their caregivers, whose willingness to share their experiences has been the cornerstone of this research. Your generosity and openness have been instrumental in advancing our understanding of Parkinson's disease.

Table of Contents

Chapter 1: General Introduction

As the second most common neurodegenerative disorder, Parkinson's disease (PD) stands as a significant challenge in global health, impacting over six million people worldwide (Dorsey et al., 2018). Despite its prevalence, a cure remains elusive, and the underlying causes of PD are not fully understood. Most PD cases are idiopathic, meaning their origins are unknown, and the onset, presentation, and progression of this disease varies widely across individuals (Blonder, 2018). Such variability underscores the necessity for personalized, multidisciplinary care strategies. Further, this complexity emphasizes the urgent need for continued research to explore PD's multifaceted nature.

This dissertation focuses on utilizing unstructured text data to gain insights into PD. Text data, with its inherent complexity, presents significant challenges for quantitative analysis and medical understanding. Here, natural language processing (NLP) plays a crucial role, offering a means to convert unstructured text into a structured format that can be more easily analyzed.

In the subsequent sections of this introductory chapter, further background on PD and NLP will be provided. Following this background, the objectives for each data-driven chapter will be outlined, focusing on extracting insights from three distinct text data sources: clinical research publications on PD, prompted patient narratives, and unprompted social media text.

Introduction to Parkinson's Disease

PD is a progressive neurodegenerative disorder characterized by the degeneration of neurons in the substantia nigra, a brain region essential for movement control (McDonald et al., 2018). In particular, the death of dopaminergic neurons reduces the availability of dopamine, which in turn reduces muscle coordination, leading to PD's motor symptoms (Xia & Mao, 2012). Additionally, protein aggregates called Lewy bodies accumulate in neurons (Bloem et al., 2021;

Kalia & Lang, 2015); their role in disease progression is still under investigation (Chu et al., 2024).

PD manifests through a spectrum of motor and non-motor symptoms, contributing significantly to morbidity. For a PD diagnosis, an individual must exhibit at least two of the three cardinal motor symptoms: resting tremor, rigidity (stiffness), and bradykinesia (slowness of movement) (Postuma et al., 2015). People with PD may also experience falls, dysarthria (difficulty speaking), dysphagia (difficulty swallowing), postural instabilities, medication response fluctuations, and dyskinesias (involuntary and erratic muscle movements) (Bloem et al., 2021). Beyond motor symptoms, PD encompasses a range of non-motor manifestations, including memory issues, mood disorders, sleep disturbances, and autonomic dysfunction (e.g., constipation, difficulties with bladder control, and orthostatic hypotension) (Bloem et al., 2021). These non-motor symptoms play a substantial role in the disability associated with the disease (Lang, 2007). The variability in the manifestation and progression of symptoms, which generally unfold slowly and cause accumulating disability over time (Bloem et al., 2021; Kalia & Lang, 2015), underscores the necessity of a comprehensive understanding for effective management of PD.

Current therapeutic strategies focus on management of symptoms, primarily through physiotherapy and pharmacotherapy. Physiotherapy aims to improve a range of motor symptoms, including altered gait and balance issues. Treatment outcomes can vary based on the specific approach utilized (Radder et al., 2020). Pharmacotherapy, initiated upon the onset of functional impairment, involves multiple classes of drugs that target motor symptoms (Connolly & Lang, 2014). These medications come with their own sets of complications, necessitating personalized treatment plans based on factors like demographics and health history. For example, while

dopamine agonists are effective at alleviating some of the motor symptoms of PD, people with a history of obsessive-compulsive disorder are especially at risk of developing an impulse control disorder as a result of this treatment regimen (Connolly & Lang, 2014).

Sex/Gender Differences

Despite the common perception that PD predominantly affects men, recent findings suggest a more balanced prevalence between men and women. A meta-analysis by Zirra et al. (2022) indicated that the difference in prevalence might be smaller than previously thought and also varies by geography. In 2016, about 47.5% of people with PD were women (Dorsey et al., 2018).

Although the prevalence of PD is fairly balanced by sex/gender, there is evidence to suggest that sex/gender may explain some of the variability observed among people with PD. Men, particularly those aged 50-69, are reported to have a higher susceptibility to PD (Pringsheim et al., 2014) and women tend to experience the onset of the disease at a later age than their male counterparts (Bloem et al., 2021).

Sex/gender may also explain some of the variability observed in regards to symptoms of PD. Men with PD generally experience a greater number of disability years and are more likely to develop cognitive impairments and daytime sleepiness than are women (Deuschl et al., 2020; Iwaki et al., 2021). On the other hand, women with PD are at a higher risk of experiencing dyskinesias and depression and less likely to present with akinesia or rigid features compared to men (Iwaki et al., 2021).

In terms of treatment and healthcare access, women with PD often face challenges. They experience delays in receiving healthcare services and are less frequently provided with healthcare than men (Fullard et al., 2017, 2018). Furthermore, women are underrepresented in

neurosurgical interventions like deep brain stimulation and pallidotomy, even though they are more prone to conditions like dyskinesias, where such interventions could be beneficial (Hariz & Hariz, 2000; Iwaki et al., 2021).

The underrepresentation of women extends to PD clinical trials as well. For example, 55.7% of randomized clinical trials on PD published between 2010 and 2016 included a study population that consisted of 59% men or higher (Tosserams et al., 2018). This skew in data potentially compromises the generalizability of research findings, and by extension, the efficacy of treatment strategies across groups. Recognizing this gap, Pavon et al. (2010) advocated for a heightened emphasis on investigating sex/gender disparities within PD research. Despite these calls for action, studies specifically analyzing data by sex/gender remain scarce (Iwaki et al., 2021).

Fall Events

One particularly burdensome complication of PD is falling, which is highly prevalent and a major research priority for both patients and clinical providers (Deane et al., 2014; Politis et al., 2010). A fall is often defined as an event where a person unintentionally lands on the ground or a lower level, regardless of whether or not the event resulted in injury (Bloem et al., 2001; Maki et al., 1994). Falls are a public health burden, costing people significantly in their time, money, and health (Burns & Kakara, 2018; Florence et al., 2018; Haddad et al., 2019; Stack & Ashburn, 1999). For example, falls were the leading cause of injury-related death among older adults in the United States between 2007 and 2016 (Burns & Kakara, 2018). These high costs could be minimized with a better understanding of the causes of falls and subsequent implementation of preventative measures.

People with PD are more likely both to fall at least once or be frequently fallers than healthy older adults (HOAs) (Bloem et al., 2001; Stack & Ashburn, 1999). Falling in this population can be incapacitating, often resulting in soft tissue injuries and can be disabling even early in disease progression (Bloem et al., 2001). For these reasons, it is of particular importance to predict and prevent falls in this population. Falls are heterogeneous and can result from multiple types of biomechanical perturbations, including perturbations to an individual's base of support (BoS; e.g., trips) or center of mass (CoM; e.g., overextension during bending) (Maki et al., 1994). BoS falls are more common in HOAs compared to CoM falls (Maki et al., 1994). However, the opposite holds true in subpopulations of people, such as people with PD, where disease-related postural instability results in more CoM falls (Bloem et al., 2001). Although it has been well established that falls are multifactorial (Lach et al., 1991; Maki et al., 1994), many studies still use all-cause falls as an outcome measure (e.g., Asai et al., 2022; Lord et al., 2017; McKay et al., 2018, 2019, 2021).

Stack and Ashburn (1999) published their findings from interviews of community-dwelling people with PD about circumstances surrounding their fall events. They concluded that the routines that challenged individuals, while somewhat overlapping (e.g., turning), can vary considerably. As a follow up, Ashburn et al. (2008) recruited people with PD to fill out fall diaries for six months. The authors identified six primary activity-cause pairs responsible for more than half of the falls, which are scenarios likely encountered by all individuals with PD. These include freezing during standing or walking and postural instability when bending or reaching, during transfers, and while walking. They also found that falls when standing were more prominent in this study than in their earlier study, a finding they suspect reflects the more advanced stages of disease in individuals in the latter study. Their findings

suggested that not only are fall contexts extremely individualized, the circumstances in which people fall change as PD progresses.

Each person's experience of PD is shaped by their own biological and social determinants of health which can impact their risk of falls (Bloem et al., 2021; McKay et al., 2018). For example, older adult women are more likely to fall than older adult men (Gale et al., 2016). This difference is partially explained by mechanisms such as older women, on average, being in less physical shape than older men (Pereira et al., 2013). Additionally, certain treatments for PD, such as dopaminergic treatments that cause dyskinesias, can increase the risk of falling (Robinson et al., 2005).

Summary

PD is a highly heterogeneous disorder with variable symptom onset, presentation, and progression, necessitating personalized care approaches (Bloem et al., 2021). This variability may be partially explained by factors related to sex/gender (Iwaki et al., 2021). Moreover, falls represent a significant complication for individuals with PD, contributing to substantial morbidity and highlighting the need for targeted prevention strategies (Bloem et al., 2001). The disease's heterogeneity, coupled with sex/gender disparities and the high prevalence of falls, underscores the imperative for in-depth research to unravel PD's complexities.

Introduction to Machine Learning and Natural Language Processing

Machine learning (ML) techniques are used to autonomously identify patterns within data without explicit programming. Unlike traditional inferential statistics, which aim to make population-wide inferences from sample data, ML is used to detect predictive patterns from data that could then be applied to predict outcomes for new, previously unseen data (Bzdok et al.,

2018). Natural language processing (NLP) is a field that intersects with ML, utilizing ML techniques among others to understand, interpret, and generate human language.

ML methods can broadly be classified into two categories: supervised learning, where models are trained to identify patterns in data using known outcomes as ground-truth labels, and unsupervised learning, where models discern patterns from the input data without any predefined output labels (De et al., 2023).

Text Classification

Figure 1 illustrates a ML workflow for supervised text classification, encompassing steps from data acquisition to deployment. It demonstrates a binary classification problem in which descriptions of falls are categorized into class 1 (blue) or class 2 (green). In ML and NLP, inputs into algorithms are known as features (De et al., 2023), which are predominantly representations of human language in this context.

Text Preprocessing. To enable computers to process language, it must be converted into numerical form. This conversion process is depicted from **Figure 1a** to **1c**. Initially, relevant texts for training the model are identified and acquired. This collection of texts is referred to as a corpus. Text preprocessing then aims to standardize and simplify text data, as shown in **Figure 1b**. This step can involve converting text to lowercase, removing punctuation, eliminating stopwords, which are words that contribute no semantic value to the text (e.g., “a”, “is”, “the”; Smelyakov et al., 2020), and normalizing words through stemming, which reduces words to their root forms (Porter, 1980). An alternative to text stemming is text lemmatization, which converts words to their base form (Smelyakov et al., 2020). These measures address the challenges posed by variability in language by, for example, treating different cases of the same word or various conjugations of a verb as identical.

Feature Extraction. Following preprocessing, the text and other variables are transformed into a feature matrix, depicted in **Figure 1c**. This depiction is of a bag-of-words approach (Juluru et al., 2021), during which text is broken down into individual, stemmed words, also referred to as tokens, and the count of each token in a fall description is numerically represented. An alternative representation of textual data not shown in the figure is the n-gram approach (Majumder et al., 2002), which captures sequences of n items—unigrams (one word), bigrams (two words), and trigrams (three words)—to better represent context than the bag-of-words model by considering word order. Non-textual features, such as participant age, can also be incorporated into the feature matrix (**Figure 1c**). For supervised learning tasks, the class of each entry is known, appended to the feature matrix, and used for training the model to distinguish between the classes based on predictive features.

Model Training and Evaluation. The model is trained to find an optimal way to divide data into two categories based on provided features, with this division referred to as a decision boundary. Following the establishment of a decision boundary by the classifier, akin to a support vector machine as depicted in **Figure 1d**, we evaluate model performance through metrics derived from a confusion matrix, contrasting actual versus predicted classifications (**Figure 1e**). Evaluation metrics such as accuracy, sensitivity/recall, specificity, precision, and F_1-score provide comprehensive insights into model performance, which are essential for assessing the effectiveness of NLP models (Vujovic, 2021). These metrics reflect the balance between different outcomes, like reducing false positives versus false negatives, tailored to application needs. In medical settings, when binary classifiers are used to detect illness, the ideal performance metric may vary. In some cases, it may be important to prioritize accurate identification of sick individuals and to minimize false positives among the healthy (high

sensitivity). In other instances, it may be more important to ensure that healthy individuals are accurately identified, reducing misclassification of sick individuals (high specificity) (Florkowski, 2008).

Model performance can be improved by revising the feature set or exploring alternative classifiers. Shallow classifiers, such as Support Vector Machine, Naive Bayes, and k-Nearest Neighbors, resemble the basic tools in a toolkit, learning directly from labeled examples to categorize new information based on learned patterns. Deep classifiers include architectures like neural networks that draw inspiration from the human brain's ability to recognize patterns, employing layers of neurons to parse complex patterns in large datasets. This deep architecture endows neural networks with a profound ability to grasp the nuances of human language, making them particularly effective in interpreting intricate language data (Goldberg, 2016). At the cutting edge of NLP technology stand transformers, which represent a significant leap forward in model sophistication. By analyzing the full context of text and the relationships between words, transformers use a mechanism known as multi-head attention to simultaneously focus on various aspects of a sentence. This mechanism allows them to capture the complete meaning of text with remarkable accuracy (Vaswani et al., 2017).

Model Deployment. Once the model achieves satisfactory performance, it is prepared for deployment on new data. For example, as illustrated in **Figure 1f**, a yellow triangle is classified as Class 2 due to its position relative to the decision boundary. However, it is crucial to remember that a model's effectiveness and generalizability depend heavily on its training data. As highlighted by Bender et al. (2021), large language models (LLMs) can act as "stochastic parrots," reproducing sequences of words from their training data (Bender et al., 2021). This replication can include the biases—sometimes harmful—present in the training material; these

biases can still occur even when simpler architectures are employed. Therefore, deploying NLP models, particularly in medical contexts, requires careful evaluation of both their benefits and potential harms. For example, Omiye et al. (2023) found that four commercially available LLMs will promote race-based medicine that has been discredited when prompted for certain medical advice. Given that the training materials used to train LLMs are not transparent, it is challenging to ensure these models do not perpetuate harmful biases. Rigorous testing, continuous monitoring, and the inclusion of diverse datasets are essential to mitigate these risks and enhance the reliability of NLP applications in healthcare.

Additional NLP Tasks

Text classification is just one application of NLP. Among many other techniques, NLP can also be used for information retrieval and extraction. The objective of information retrieval is to identify and retrieve a subset of textual documents from a larger collection that are pertinent to a specific query, often through keyword search (Piskorski & Yangarber, 2013). Information extraction is an application of NLP that aims to extract structured data from unstructured free text (Piskorski & Yangarber, 2013). This task can be nontrivial given the complexity and potential ambiguity of language. For example, target information may need to be extracted from multiple locations in a document or inferred from ambiguous text (Huttunen et al., 2002). Promisingly, the GPT family of models perform well at this task (Foppiano et al., 2024).

Summary

ML and NLP are powerful tools for analyzing human language, enabling computers to extract meaningful patterns from text. The process involves preparing text data through standardization and simplification techniques, such as stemming (Porter, 1980) and lemmatization (Smelyakov et al., 2020), to make it understandable for algorithms. Feature

extraction methods, like bag-of-words (Juluru et al., 2021) and n-grams (Majumder et al., 2002), then transform this prepared text into a numerical format that ML models can process. These models range from shallow classifiers to deep neural networks and transformers, each suited to different types of language analysis tasks. Evaluation metrics help determine a model's effectiveness, focusing on accuracy and the ability to distinguish between different outcomes. The deployment of these models must be approached with caution to avoid perpetuating biases present in the training data (Bender et al., 2021).

Objective and Structure of the Dissertation

This dissertation aims to leverage unstructured text data to extract insights about PD, focusing on three types of text data: clinical research publications on PD, prompted patient narratives, and unprompted social media text.

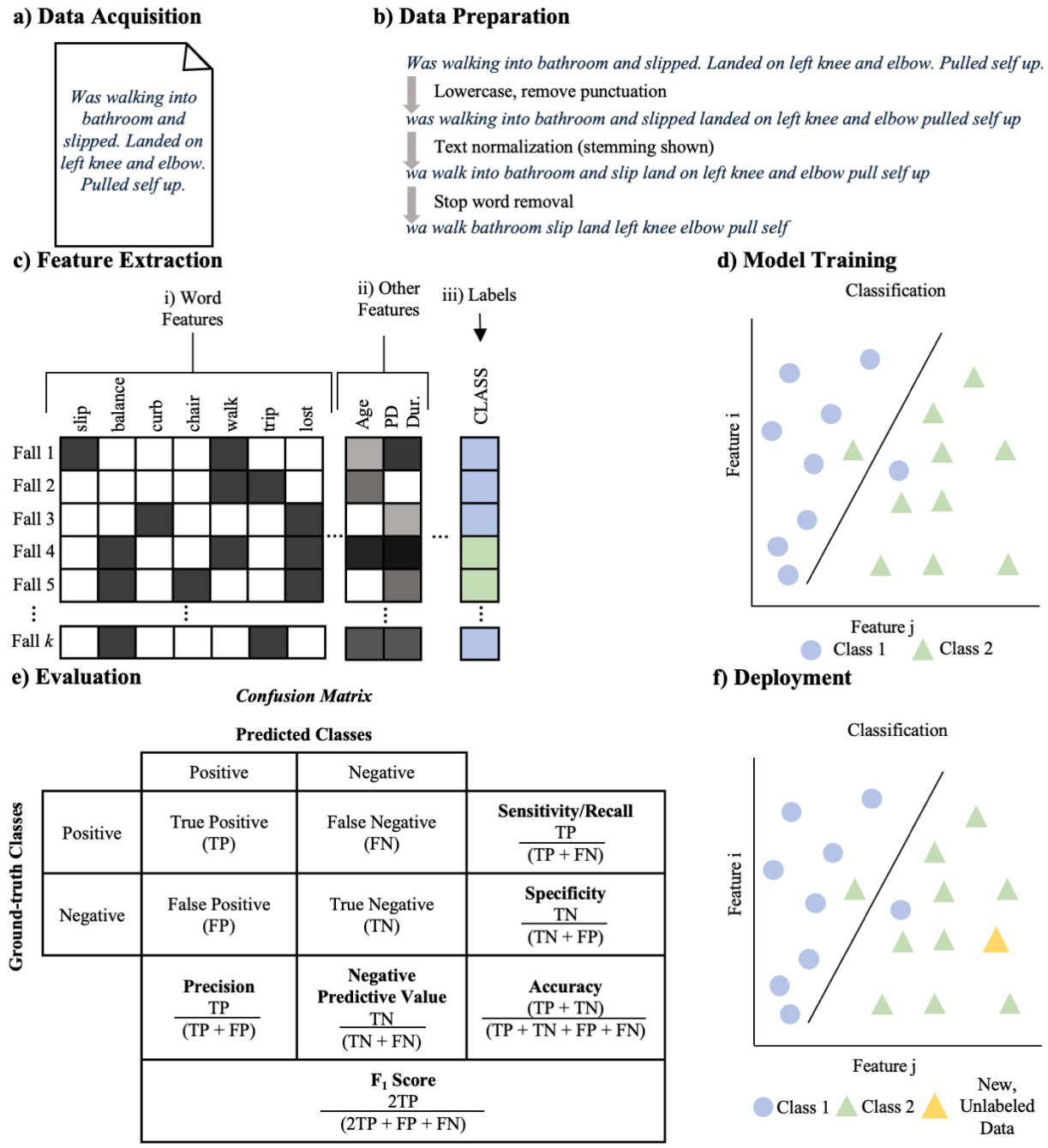
Chapters 2 and 3 utilize published clinical research. The first project's objective is to assess the adequacy of female representation in PD clinical research and examine the evolution of inclusion practices and the consideration of sex/gender as a variable over time. The second project extends this analysis by automating the extraction of sex-related information from articles, facilitating scalable analysis. Chapter 4 explores patient-provided narratives about fall events, aiming to develop a text classifier to categorize falls by biomechanical cause. Chapter 5 investigates narratives from individuals with PD and their caregivers shared on Facebook, assessing the medical relevance of shared lived experiences. Last, **Appendix A** further clarifies the constructs of sex and gender.

The dissertation concludes with a discussion of findings, future research directions, and the potential of NLP tools in customizing PD care.

Figures

Figure 1

Machine Learning Workflow for Classifying Text



Note. This workflow illustrates an example of the sequential steps involved in processing and classifying text data: **(a)** Acquisition of raw descriptions of falls, **(b)** Preprocessing steps

including conversion to lowercase, punctuation removal, word normalization via stemming, and stop word elimination, **(c)** Feature extraction where (i) text data is transformed using a bag-of-words approach, (ii) demographic data features are integrated, and (iii) accurate labels are assigned to the dataset, **(d)** Model training, employing a support vector machine to delineate decision boundaries effectively between classes, **(e)** Model evaluation through metrics such as precision, sensitivity (recall), accuracy, and F_1-score, with potential iterations of steps **(c)** and **(d)** for performance enhancement, and **(f)** Deployment of the model for predicting the categories of new, unseen data (yellow triangle). This figure was inspired by (De et al., 2023). Abbreviations: Parkinson's disease duration (PD Dur.)

References

Asai, T., Oshima, K., Fukumoto, Y., Yonezawa, Y., Matsuo, A., & Misu, S. (2022). The association between fear of falling and occurrence of falls: A one-year cohort study. *BMC Geriatrics*, *22*(1), 393. https://doi.org/10.1186/s12877-022-03018-2

Ashburn, A., Stack, E., Ballinger, C., Fazakarley, L., & Fitton, C. (2008). The circumstances of falls among people with Parkinson's disease and the use of falls diaries to facilitate reporting. *Disability and Rehabilitation*, *30*(16), 1205–1212. https://doi.org/10.1080/09638280701828930

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bloem, B. R., Grimbergen, Y. A. M., Cramer, M., Willemsen, M., & Zwinderman, A. H. (2001). Prospective assessment of falls in Parkinson's disease. *Journal of Neurology*, *248*(11), 950–958. https://doi.org/10.1007/s004150170047

Bloem, B. R., Okun, M. S., & Klein, C. (2021). Parkinson's disease. *The Lancet*, *397*(10291), 2284–2303. https://doi.org/10.1016/S0140-6736(21)00218-X

Blonder, L. X. (2018). Historical and cross-cultural perspectives on Parkinson's disease. *Journal of Complementary & Integrative Medicine*, *15*(3), 1–15. https://doi.org/10.1515/jcim-2016-0065

Burns, E., & Kakara, R. (2018). Deaths from falls among persons aged ≥65 years—United States, 2007–2016. *Morbidity and Mortality Weekly Report*, *67*(18), 509–514. https://doi.org/10.15585/mmwr.mm6718a1

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234. https://doi.org/10.1038/nmeth.4642

Chu, Y., Hirst, W. D., Federoff, H. J., Harms, A. S., Stoessl, A. J., & Kordower, J. H. (2024). Nigrostriatal tau pathology in parkinsonism and Parkinson's disease. *Brain*, *147*, 444–457. https://doi.org/10.1093/brain/awad388

Connolly, B. S., & Lang, A. E. (2014). Pharmacological treatment of Parkinson disease: A review. *Journal of the American Medical Association*, *311*(16), 1670. https://doi.org/10.1001/jama.2014.3654

De, A., Bhatia, K. P., Volkmann, J., Peach, R., & Schreglmann, S. R. (2023). Machine learning in tremor analysis: Critique and directions. *Movement Disorders*, *38*(5), 717–731. https://doi.org/10.1002/mds.29376

Deane, K. H. O., Flaherty, H., Daley, D. J., Pascoe, R., Penhale, B., Clarke, C. E., Sackley, C., & Storey, S. (2014). Priority setting partnership to identify the top 10 research priorities for the management of Parkinson's disease. *BMJ Open*, *4*(12), e006434. https://doi.org/10.1136/bmjopen-2014-006434

Deuschl, G., Beghi, E., Fazekas, F., Varga, T., Christoforidi, K. A., Sipido, E., Bassetti, C. L., Vos, T., & Feigin, V. L. (2020). The burden of neurological diseases in Europe: An analysis for the Global Burden of Disease Study 2017. *The Lancet Public Health*, *5*(10), e551–e567. https://doi.org/10.1016/S2468-2667(20)30190-0

Dorsey, E. R., Elbaz, A., Nichols, E., Abbasi, N., Abd-Allah, F., Abdelalim, A., Adsuar, J. C., Ansha, M. G., Brayne, C., Choi, J.-Y. J., Collado-Mateo, D., Dahodwala, N., Do, H. P., Edessa, D., Endres, M., Fereshtehnejad, S.-M., Foreman, K. J., Gankpe, F. G., Gupta, R., … Murray, C. J. L. (2018). Global, regional, and national burden of Parkinson's disease,

1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, *17*(11), 939–953. https://doi.org/10.1016/S1474-4422(18)30295-3

Florence, C. S., Bergen, G., Atherly, A., Burns, E., Stevens, J., & Drake, C. (2018). Medical costs of fatal and nonfatal falls in older adults: Medical costs of falls. *Journal of the American Geriatrics Society*, *66*(4), 693–698. https://doi.org/10.1111/jgs.15304

Florkowski, C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: Communicating the performance of diagnostic tests. *Clinical Biochemist Reviews*, *29*(Suppl (i)), S83–S87.

Foppiano, L., Lambard, G., Amagasa, T., & Ishii, M. (2024). *Mining experimental data from materials science literature with large language models* (No. arXiv:2401.11052). arXiv. http://arxiv.org/abs/2401.11052

Fullard, M. E., Thibault, D. P., Hill, A., Fox, J., Bhatti, D. E., Burack, M. A., Dahodwala, N., Haberfeld, E., Kern, D. S., Klepitskava, O. S., Urrea-Mendoza, E., Myers, P., Nutt, J., Rafferty, M. R., Schwalb, J. M., Shulman, L. M., & Willis, A. W. (2017). Utilization of rehabilitation therapy services in Parkinson disease in the United States. *Neurology*, *89*(11), 1162–1169. https://doi.org/10.1212/WNL.0000000000004355

Fullard, M. E., Thibault, D. P., Todaro, V., Foster, S., Katz, L., Morgan, R., Kern, D. S., Schwalb, J. M., Urrea Mendoza, E., Dahodwala, N., Shulman, L., & Willis, A. W. (2018). Sex disparities in health and health care utilization after Parkinson diagnosis: Rethinking PD associated disability. *Parkinsonism & Related Disorders*, *48*, 45–50. https://doi.org/10.1016/j.parkreldis.2017.12.012

Gale, C. R., Cooper, C., & Aihie Sayer, A. (2016). Prevalence and risk factors for falls in older men and women: The English Longitudinal Study of Ageing. *Age and Ageing*, *45*(6),

789–794. https://doi.org/10.1093/ageing/afw129

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345–420. https://doi.org/10.1613/jair.4992

Haddad, Y. K., Bergen, G., & Florence, C. S. (2019). Estimating the economic burden related to older adult falls by state. *Journal of Public Health Management and Practice*, *25*(2), E17–E24. https://doi.org/10.1097/PHH.0000000000000816

Hariz, G.-M., & Hariz, M. I. (2000). Gender distribution in surgery for Parkinson's disease. *Parkinsonism & Related Disorders*, *6*(3), 155–157. https://doi.org/10.1016/S1353-8020(00)00009-2

Huttunen, S., Yangarber, R., & Grishman, R. (2002). Complexity of event structure in IE scenarios. *Proceedings of the 19th International Conference on Computational Linguistics*, *1*, 1–7. https://doi.org/10.3115/1072228.1072393

Iwaki, H., Blauwendraat, C., Leonard, H. L., Makarious, M. B., Kim, J. J., Liu, G., Maple‐Grødem, J., Corvol, J., Pihlstrøm, L., Nimwegen, M., Smolensky, L., Amondikar, N., Hutten, S. J., Frasier, M., Nguyen, K. H., Rick, J., Eberly, S., Faghri, F., Auinger, P., … Nalls, M. A. (2021). Differences in the presentation and progression of Parkinson's disease by sex. *Movement Disorders*, *36*(1), 106–117. https://doi.org/10.1002/mds.28312

Juluru, K., Shih, H.-H., Keshava Murthy, K. N., & Elnajjar, P. (2021). Bag-of-words technique in natural language processing: A primer for radiologists. *RadioGraphics*, *41*(5), 1420–1426. https://doi.org/10.1148/rg.2021210025

Kalia, L. V., & Lang, A. E. (2015). Parkinson's disease. *The Lancet*, *386*(9996), 896–912. https://doi.org/10.1016/S0140-6736(14)61393-3

Lach, H. W., Reed, A. T., Arfken, C. L., Miller, J. P., Paige, G. D., Birge, S. J., & Peck, W. A.

(1991). Falls in the elderly: Reliability of a classification system. *Journal of the American Geriatrics Society*, *39*(2), 197–202. https://doi.org/10.1111/j.1532-5415.1991.tb01626.x

Lang, A. E. (2007). The progression of Parkinson disease: A hypothesis. *Neurology*, *68*(12), 948–952. https://doi.org/10.1212/01.wnl.0000257110.91041.5d

Lord, S., Galna, B., Yarnall, A. J., Morris, R., Coleman, S., Burn, D., & Rochester, L. (2017). Natural history of falls in an incident cohort of Parkinson's disease: Early evolution, risk and protective features. *Journal of Neurology*, *264*(11), 2268–2276. https://doi.org/10.1007/s00415-017-8620-y

Majumder, P., Mitra, M., & Chaudhuri, B. B. (2002). N-gram: A language independent approach to IR and NLP. *International Conference on Universal Knowledge and Language*, *2*.

Maki, B. E., Holliday, P. J., & Topper, A. K. (1994). A prospective study of postural balance and risk of falling in an ambulatory and independent elderly population. *Journal of Gerontology*, *49*(2), M72–M84. https://doi.org/10.1093/geronj/49.2.M72

McDonald, C., Gordon, G., Hand, A., Walker, R. W., & Fisher, J. M. (2018). 200 Years of Parkinson's disease: What have we learnt from James Parkinson? *Age and Ageing*, *47*(2), 209–214. https://doi.org/10.1093/ageing/afx196

McKay, J. L., Hackney, M. E., Factor, S. A., & Ting, L. H. (2019). Lower limb rigidity Is associated with frequent falls in Parkinson's disease. *Movement Disorders Clinical Practice*, *6*(6), 446–451. https://doi.org/10.1002/mdc3.12784

McKay, J. L., Lang, K. C., Bong, S. M., Hackney, M. E., Factor, S. A., & Ting, L. H. (2021). Abnormal center of mass feedback responses during balance: A potential biomarker of falls in Parkinson's disease. *PLOS ONE*, *16*(5), e0252119. https://doi.org/10.1371/journal.pone.0252119

McKay, J. L., Lang, K. C., Ting, L. H., & Hackney, M. E. (2018). Impaired set shifting is associated with previous falls in individuals with and without Parkinson's disease. *Gait & Posture*, *62*, 220–226. https://doi.org/10.1016/j.gaitpost.2018.02.027

Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *Npj Digital Medicine*, *6*(1), 195. https://doi.org/10.1038/s41746-023-00939-z

Pavon, J. M., Whitson, H. E., & Okun, M. S. (2010). Parkinson's disease in women: A call for improved clinical studies and for comparative effectiveness research. *Maturitas*, *65*(4), 352–358. https://doi.org/10.1016/j.maturitas.2010.01.001

Pereira, C. L. N., Baptista, F., & Infante, P. (2013). Men older than 50 yrs are more likely to fall than women under similar conditions of health, body composition, and balance. *American Journal of Physical Medicine & Rehabilitation*, *92*(12), 1095–1103. https://doi.org/10.1097/PHM.0b013e31829b49eb

Piskorski, J., & Yangarber, R. (2013). Information extraction: Past, present and future. In T. Poibeau, H. Saggion, J. Piskorski, & R. Yangarber (Eds.), *Multi-source, Multilingual Information Extraction and Summarization* (pp. 23–49). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28569-1_2

Politis, M., Wu, K., Molloy, S., G. Bain, P., Chaudhuri, K. R., & Piccini, P. (2010). Parkinson's disease symptoms: The patient's perspective. *Movement Disorders*, *25*(11), 1646–1651. https://doi.org/10.1002/mds.23135

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K.,

Litvan, I., Lang, A. E., Halliday, G., Goetz, C. G., Gasser, T., Dubois, B., Chan, P., Bloem, B. R., Adler, C. H., & Deuschl, G. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Movement Disorders*, *30*(12), 1591–1601. https://doi.org/10.1002/mds.26424

Pringsheim, T., Jette, N., Frolkis, A., & Steeves, T. D. L. (2014). The prevalence of Parkinson's disease: A systematic review and meta-analysis. *Movement Disorders*, *29*(13), 1583–1590. https://doi.org/10.1002/mds.25945

Radder, D. L. M., Lígia Silva De Lima, A., Domingos, J., Keus, S. H. J., Van Nimwegen, M., Bloem, B. R., & De Vries, N. M. (2020). Physiotherapy in Parkinson's disease: A meta-analysis of present treatment modalities. *Neurorehabilitation and Neural Repair*, *34*(10), 871–880. https://doi.org/10.1177/1545968320952799

Robinson, K., Dennison, A., Roalf, D., Noorigian, J., Cianci, H., Bunting-Perry, L., Moberg, P., Kleiner-Fisman, G., Martine, R., Duda, J., Jaggi, J., & Stern, M. (2005). Falling risk factors in Parkinson's disease. *NeuroRehabilitation*, *20*(3), 169–182. https://doi.org/10.3233/NRE-2005-20304

Smelyakov, K., Karachevtsev, D., Kulemza, D., Samoilenko, Y., Patlan, O., & Chupryna, A. (2020). Effectiveness of preprocessing algorithms for natural language processing applications. *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)*, 187–191. https://doi.org/10.1109/PICST51311.2020.9467919

Stack, E., & Ashburn, A. (1999). Fall events described by people with Parkinson's disease: Implications for clinical interviewing and the research agenda. *Physiotherapy Research International*, *4*(3), 190–200. https://doi.org/10.1002/pri.165

Tosserams, A., Araújo, R., Pringsheim, T., Post, B., Darweesh, S. K. L., IntHout, J., & Bloem, B. R. (2018). Underrepresentation of women in Parkinson's disease trials. *Movement Disorders*, *33*(11), 1825–1826. https://doi.org/10.1002/mds.27505

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need* (No. arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Vujovic, Ž. Đ. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6). https://doi.org/10.14569/IJACSA.2021.0120670

Xia, R., & Mao, Z.-H. (2012). Progression of motor symptoms in Parkinson's disease. *Neuroscience Bulletin*, *28*(1), 39–48. https://doi.org/10.1007/s12264-012-1050-z

Zirra, A., Rao, S. C., Bestwick, J., Rajalingam, R., Marras, C., Blauwendraat, C., Mata, I. F., & Noyce, A. J. (2022). Gender differences in the prevalence of Parkinson's disease. *Movement Disorders Clinical Practice*, mdc3.13584. https://doi.org/10.1002/mdc3.13584

Chapter 2: Persistent Underrepresentation of Women in Parkinson's Disease Clinical Research:

A Systematic Review of Research Before and After Major NIH Policy Changes

Jeanne M. Powell[1], Sarah S. Palmer[2], Catherine L. Jones[3], Grey Skelton[1], Imari Genias[1], Anissa Patel[1], Sofia Hefner[4], Mia S. White[1], Donna L. Maney[1,5], J. Lucas McKay[1, 4]

[1] Emory University, Atlanta, GA 30322, USA

[2] Boston University, Boston, MA 02118, USA

[3] University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[4] Georgia Institute of Technology, Atlanta, GA 30332, USA

[5] Harvard Radcliffe Institute, Harvard University, Cambridge, MA, United States

Abstract

Historically, women have been underrepresented in clinical research. In this systematic review, we examined sex/gender representation in clinical research on Parkinson's disease (PD) across three pivotal time periods (1982–1992, 2009, 2021) correlating with significant NIH policy shifts: the 1993 NIH Revitalization Act and the 2016 NIH Policy on Sex as a Biological Variable (SABV). Our primary objective was to evaluate the representation of women in PD clinical trials and track changes in sex/gender inclusion, data disaggregation, and sex/gender-specific analysis over time. We focused on primary clinical research involving pharmacological or physiotherapeutic interventions in PD, with a particular emphasis on United States-based studies for longitudinal analysis. Comprehensive database searches were conducted, culminating in the final search in February 2023. Risk of bias was not assessed, and a meta-analysis was not performed; instead, we employed descriptive and statistical analyses. Results from 465 studies showed that despite an overall increase in the proportion of studies that include men and women, there is a persistent underrepresentation of women in PD research. Further, we observed no change in data disaggregation over time, despite NIH policy mandates. These results highlight the need for more inclusive research practices in PD clinical research. The review was not registered.

Persistent Underrepresentation of Women in Parkinson's Disease Clinical Research: A Systematic Review of Research Before and After Major NIH Policy Changes

Parkinson's disease (PD), a pervasive neurodegenerative disorder for which there is no cure, continues to pose significant challenges for affected individuals (Rajith & Angiel, 2023). The heterogeneous nature of PD in the onset, progression, and manifestation of symptoms necessitates a tailored approach to management, which has historically been largely based on pharmacological and physiotherapeutic interventions (Bloem et al., 2021). Importantly, the variation in effectiveness of these interventions can be partially explained by demographic variables like sex/gender, among others (Cerri et al., 2019).

Historically, clinical research, including that on PD, has seen an underrepresentation of women, leading to a significant gap in sex/gender-related health data (National Academies of Sciences, Engineering, and Medicine, 2022; Tosserams et al., 2018). In response to this androcentric focus on health research, the National Institutes of Health (NIH) enacted policies that included the 1993 NIH Revitalization Act (Public Law 103-43) and the 2016 Policy on Sex as a Biological Variable (SABV) (NOT-OD-15-102). The 1993 Revitalization Act mandated the inclusion of women and minority groups in all NIH-funded clinical studies (Kelty et al., 2012). The 2016 SABV policy further reinforced this mandate, requiring researchers to consider sex/gender in study design, analysis, and reporting in clinical as well as basic and preclinical non-human vertebrate research. The SABV policy aims to enhance scientific rigor and transparency but does not necessarily compel researchers to focus on sex/gender differences. Instead, it mandates that researchers disaggregate their data by sex, or provide raw data, irrespective of whether their study is statistically powered to detect sex/gender differences (Geller et al., 2018; Clayton, 2018).

In this study, we conducted a systematic review of PD clinical research from 1982–1992, 2009, and 2021, time periods surrounding these NIH policy changes. We assessed the proportion of women in clinical trials on PD, specifically in pharmacological and physiotherapeutic sub-domains of PD research, and examined the progression of sex/gender inclusion, data disaggregation, and sex/gender-specific analysis over time.

We aimed to answer the following questions:

1. Does the proportion of women in PD clinical research published in 2021 reflect the prevalence of women in the PD population?
2. Does the proportion of women participants in studies published in 2021 differ between the pharmacological and physiotherapeutic sub-domains of PD clinical research?
3. Has the number of PD studies reporting the inclusion of men and women changed over time?
4. Has there been a change in the proportion of women with PD in clinical studies over time?
5. Has the rate of disaggregation of data by sex/gender changed over time in clinical PD research?
6. Has the rate of analysis of data considering sex/gender changed over time in clinical PD research?

In our review, we employ the term 'sex/gender'. The NIH defines 'sex' as a construct based on anatomy, physiology, genetics, and hormones (https://orwh.od.nih.gov/sex-gender). The NIH defines 'gender' as a social construct, primarily influenced by cultural expectations that are associated with certain sex traits (https://orwh.od.nih.gov/sex-gender). These terms are often

used inconsistently and interchangeably in biomedical literature, including the literature we reviewed (Blakeman & Fillman, 2021). Given that the studies in our corpus typically used one or both terms without clear operationalization, we here use the joint term 'sex/gender' to encompass both dimensions, which are not separable for most research purposes. Of note, no studies in our corpus reported the inclusion of any sexual and gender minorities. Please refer to Appendix A of this dissertation for a more detailed discussion of the use of the terms ‘sex’ and ‘gender’ in biomedical research.

Materials and Methods

Eligibility Criteria

We established inclusion criteria for our systematic review to assess shifts in women's representation in PD clinical research, focusing on time periods that should reflect responses to NIH policy changes. Our criteria required studies to be: 1) published in 1982-1992, 2009, or 2021, 2) primary clinical research, including post-hoc analyses of full datasets but excluding meta-analyses, and 3) focused on the efficacy of pharmacological or physiotherapeutic interventions in PD patients.

We selected articles on the basis of their official publication date within these years to maintain consistency. For the pre- and post-1993 NIH policy era (1982-1992 or 2009), we included studies if the first or last authors were affiliated with U.S.-based institutions, assuming their significant influence over study design and adherence to US policies. In contrast, for 2021 publications, we included studies regardless of the research location to assess contemporary practices related to sex/gender inclusion, analysis, and disaggregation in the field of clinical research on PD. **Figure 1** provides a visual overview of our inclusion criteria and screening process.

Exclusions included case studies with single-participant data and epidemiological studies examining factors leading to PD. We also excluded studies not involving PD patients, those centered on cognitive behavioral therapy, and those using non-conventional interventions like vibrational, light, electric therapy, or dietary supplements.

Database Search Strategy

A comprehensive search was conducted on PubMed, Embase, Cochrane, CINAHL, and Web of Science. We initially focused on literature from 1985, 2009, and 2021, but repeated our search on February 8, 2023 to include papers published from 1982 to 1992 due to a limited yield. Search queries are detailed in **Appendix A**. Results from the database searches were uploaded to Covidence (Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia; available at www.covidence.org) for screening and automated duplicate removal. At the full-text screening stage, papers were acquired using Emory University Library access. For studies not available through institutional access, papers were acquired through interlibrary loan.

Screening and Data Extraction Process

Papers were independently assessed by two researchers at each stage (Title/Abstract Screening, Full Text Screening, and Data Extraction), with discrepancies resolved through discussion or third-party consultation. Papers were selected for inclusion based on criteria outlined above in the Eligibility Criteria section and depicted in **Figure 1.**

Data extraction was executed using a custom Python module (https://github.com/jeannempowell/SABVinPD/blob/main/Jupyter/SABV_PD_DataExtract.ipynb). The details of each paper, along with its supplemental files, were consolidated into a dataframe. This dataframe was paired with file paths through Python scripting, facilitating semi-automation of the data extraction process. Briefly, we used the Python module to iterate

through all papers and their supplements. When presented with the materials for each study, the data extractor was guided to input pertinent information into a form that was generated from the column names in our data extraction spreadsheet. Subsequently, these acquired data were automatically recorded within a dataframe, which was then stored as an Excel file following each instance of data extraction.

We extracted data related to each study's publication year, author affiliations, funding, intervention, participant demographics, rationale for use of a single sex/gender (if applicable), analyses by sex/gender, claims of sex/gender differences, and disaggregation of results by sex. See **Appendix B** for a full description of each study question, possible values, and considerations.

Statistical Analyses

Our study's methodological framework was geared towards exploring sex/gender disparities in PD research. We employed a blend of descriptive and statistical techniques. The analysis was bifurcated into two main strands: a cross-sectional analysis of studies published in 2021, and a longitudinal analysis encompassing United States-based (U.S-based) studies from 1982-1992, 2009, and 2021.

We conducted statistical tests in R (version 4.3.1), including logit-transformed ANOVA and Fisher's Exact Test, to explore trends and patterns in sex/gender representation. Post-hoc analyses with Bonferroni correction were applied where applicable. The analysis code is available on GitHub (https://github.com/jeannempowell/SABVinPD).

Cross-Sectional Analysis (2021)

The cross-sectional analysis included clinical studies on PD from 2021, regardless of geographic origin, with a focus on the proportion of female participants in each study. We

categorized the studies by intervention type to discern trends in sex/gender representation across different treatment modalities. For the statistical analysis, we computed logit-transformed proportions of female PD patients relative to the total PD patient population in each study, facilitating robust examination of bounded proportion data.

Longitudinal Analysis (U.S.-Based Research)

The longitudinal component targeted U.S.-based studies, assuming more consistency in cultural norms and policy influences over time compared to international studies. This analysis delved into inclusion by sex/gender, race/ethnicity reporting trends, participant sex/gender proportions, and the extent of data disaggregation by sex/gender. A key focus was on NIH-funded studies to gauge trends in federally funded research.

Inclusion by Sex/Gender. We evaluated the inclusion by sex/gender in clinical studies over time. Studies were categorized based on whether they included men only, women only, or men and women, and any studies that did not clearly specify the inclusion by sex/gender were further scrutinized. We applied Fisher's Exact Test to determine if there were statistically significant changes in the inclusion practices by sex/gender across different time periods.

Participant Sex/Gender Proportions. We analyzed the proportions of female participants in the included studies. These proportions were logit-transformed to enable robust statistical analysis of bounded data. ANOVA was conducted to compare these logit-transformed proportions across different time periods, focusing on whether the proportion of female participants has changed over time, particularly in NIH-funded studies.

Data Disaggregation by Sex/Gender. We assessed the extent to which data were disaggregated by sex/gender in the studies. This involved categorizing studies based on whether they reported descriptive statistics, statistical outputs, or raw data disaggregated by sex/gender.

Fisher's Exact Test was utilized to examine whether the practice of data disaggregation has changed over the years.

Analysis by Sex/Gender. Here, we used a Fisher's Exact Test to determine whether researchers' consideration of sex/gender in their analyses has changed over time. This consideration includes incorporating sex/gender as a main factor in models, as a covariate, and performing separate analyses by sex/gender.

Race/Ethnicity Reporting. Given the NIH Revitalization Act of 1993 also mandated the inclusion of minorities in clinical research, we analyzed trends in race/ethnicity reporting. We noted whether studies reported race/ethnicity data. We conducted Fisher's Exact Test to evaluate whether race/ethnicity reporting changed over time.

Prisma Checklist Adherence and Deviations

The PRISMA 2020 Checklist and PRISMA 2020 for Abstracts Checklist can be found in **Appendix C**. We did not register this review prior to its initiation. A separate protocol was not developed. This manuscript and its appendices constitute the protocol.

Our study diverged from PRISMA guidelines due to its focus and methodology. We did not assess risk of bias, as our primary aim was to examine participant recruitment and disaggregation of results by sex/gender, rather than the quality of evidence or outcomes in the reviewed papers. Similarly, heterogeneity and subgroup analyses were not central to our objectives, although they may be considered in future research for a more detailed examination of intervention types or demographic factors influencing sex/gender representation in PD clinical research.

We did not conduct meta-analysis or other complex syntheses; focusing instead on descriptive and statistical analyses. These analyses were categorized into cross-sectional analyses

of 2021 studies and longitudinal analyses of U.S.-based research published from 1982 to 1993, in 2009, and in 2021. Due to this descriptive approach, we did not apply methods for reporting bias assessment or certainty in the body of evidence. Specifically, we did not use formal certainty assessment tools, as our goal was to overview the current state of sex/gender representation in PD research in light of policy changes and evolving research practices rather than to provide a certainty or confidence assessment of clinical outcomes.

Results

In total, 8,481 studies were returned from our database search. After duplicate removal, 2,962 studies were screened, leading to the exclusion of 1,782 studies. A further assessment of 1,180 full-text articles resulted in 715 exclusions. The reasons for exclusions at the screening and full-text stages included case studies with only one participant, epidemiological studies, studies not testing physiotherapeutic or pharmacological interventions in PD, studies not classified as primary clinical research, publications outside the specified years, non-English publications, non- U.S.-based studies for certain years, studies focusing on cost-effectiveness, trials testing supplements, and trial protocols. These categories are not mutually exclusive; only one exclusion reason was selected for each paper, but some papers could have been excluded for multiple reasons. Ultimately, 465 studies were included in the review. See **Figure 2** for our PRISMA flow diagram and an additional box in the diagram provides a detailed breakdown of the number of studies excluded for each specified reason.

Cross Sectional Analyses

Our cross-sectional analysis encompassed 312 papers published in 2021, with 70 of these featuring first and/or last authors from U.S.-based institutions.

Question 1: Does the proportion of women in PD clinical research published in 2021 reflect the prevalence of women in the PD population?

Of the 312 studies published in 2021, 294 reported the number of participants with PD disaggregated by sex/gender. Using these 294 papers, we examined whether women were proportionally represented in PD clinical research in 2021 relative to the proportion of women in the general PD population (47.5%) (Dorsey et al., 2018).

In our dataset, the average proportion of female participants across the studies was 0.39. When we applied a logit transformation to these proportions, the mean value was -0.556, with a 95% confidence interval ranging from -0.691 to -0.422. Separately, the logit-transformed proportion of women among people with PD was calculated as -0.1. This value does not fall within the logit-transformed 95% confidence interval for the proportion of female participants in our studies. Consequently, this indicates that the actual proportion of women in these studies, which corresponds to a mean proportion significantly lower than 0.475, deviates from the expected proportion of female PD participants (see **Figure 3**).

Question 2: Does the proportion of women participants in studies published in 2021 differ between the pharmacological and physiotherapeutic sub-domains of PD clinical research?

In this analysis, we analyzed the same set of 294 papers as in Question 1 to determine whether there is a differential representation of women in clinical research focusing on physiotherapeutic versus pharmacological interventions for PD. The mean proportion of female PD participants was 0.401 in studies involving physiotherapeutic interventions and 0.383 in those with pharmacological interventions, as illustrated in **Figure 4**. To compare these two groups, we applied a logit transformation to the proportions and conducted a t-test. The results

indicated no statistically significant difference in the representation of female participants with PD between the two types of interventions ($t(261.39) = 0.7418$, $p = .459$).

Longitudinal Analyses (U.S.-based research only)

In our longitudinal analysis, we evaluated 223 papers authored by U.S.-affiliated researchers, comprising 98 papers from 1982-1992, 55 from 2009, and 70 from 2021. For some analyses, we focused on NIH-funded research within this U.S.-affiliated dataset. Within the NIH-funded dataset, there were 72 U.S.-based, NIH-funded papers reporting the sex/gender composition of participants with PD, spanning three periods: 1982-1992 (n = 26), 2009 (n = 22), and 2021 (n = 24). **Figure 5** presents a detailed categorization of these papers based on NIH funding status, distinguishing between NIH-funded, non-funded, and those without specified funding sources.

Question 3: Has the proportion of PD studies reporting the inclusion of men and women changed over time?

For this analysis, we analyzed the 223 papers authored by researchers affiliated with US institutions from our dataset. Two papers in our corpus hinted at the inclusion of at least one sex/gender through contextual clues but failed to clearly define the extent of participation for potentially other sex/gender groups (e.g., using gendered pronouns when discussing one participant's outcomes). We labeled these papers as "unspecified" given we could not tell definitively which sexes/genders were included.

To evaluate temporal trends in the reporting of sex/gender in PD clinical research, we utilized Fisher's Exact Test. This analysis focused on studies that either exclusively included men, included both men and women, or did not specify the sex/gender composition of

participants (see **Figure 6**). The analysis revealed a statistically significant change ($p = .049$) over time in the reporting of sex/gender composition in PD studies.

To further dissect these variations, we conducted post-hoc pairwise comparisons across distinct time periods (1982-1992 vs. 2009, 1982-1992 vs. 2021, and 2009 vs. 2021) and within sex/gender inclusion categories for time bins showing notable differences. After performing six such post-hoc analyses, we adjusted the p-values using the Bonferroni correction to account for multiple comparisons.

Our analysis revealed a difference in the reporting of sex/gender inclusion between studies published during the earliest (1982-1992) and the latest (2021) time frames ($p_{adj} = .0497$). Specifically, studies from 1982 to 1992 were significantly less likely to report the inclusion of men and women and instead not specify sex/gender-based inclusion compared to those published in 2021 ($p_{adj} = .0268$, Odds Ratio = 0.0928). No significant differences were observed in sex/gender inclusion reporting between the periods of 1982-1992 and 2009 ($p_{adj} = 1$) or between 2009 and 2021 ($p_{adj} = 1$).

We next examined whether the studies that included only men addressed or justified this choice. Out of the nine studies that included only male participants, three acknowledged this choice as a limitation; two of these studies were published in 2021, and one in 2009. Additionally, two of these three studies, one each from 2021 and 2009, explicitly mentioned the exclusive use of male participants in their abstracts.

Question 4: Has there been a change in the proportion of women with PD in clinical studies over time?

Question 4a: Has there been a change in the participation of women with PD in U.S.-based clinical studies over time? In this analysis, we focused on 198 papers from our

dataset that explicitly specified the sex/gender composition of participants with PD and were authored by researchers affiliated with US institutions. To assess whether the proportion of women with PD in these studies has increased over time, we conducted a logit-transformed ANOVA.

Our analysis revealed that the proportion of women participants with PD, when logit-transformed, did not differ significantly across different time periods ($F(2, 195) = 0.257$, $p = .774$) (see **Figure 7**).

Question 4b: Has there been a change in the participation of women with Parkinson's Disease in U.S.-based, NIH-funded clinical studies over time? In this study, we analyzed 72 U.S.-based, NIH-funded papers reporting the sex/gender composition of participants with PD. Our logit-transformed analysis of these NIH-funded PD clinical studies indicated no significant differences in the proportion of female participants across these time frames ($F(2, 69) = 0.698$, $p = .501$) (see **Figure 8**).

Question 5: Has the rate of disaggregation of data by sex/gender changed over time in clinical PD research?

Question 5a: Has the rate of disaggregation of data by sex/gender changed over time in U.S.-based clinical PD research? For this analysis, we focused on the 196 papers from our dataset that included men and women and were authored by researchers affiliated with US institutions. Given the rare occurrence of disaggregation of data by sex/gender, we collapsed all types of data disaggregation (descriptive statistics (1982-1992: n = 1; 2009: n = 1; 2021: n = 2), statistical outputs (2009: n = 1; 2021: n = 5), and raw data (1982-1992: n = 9; 2009: n = 3; 2021: n = 5) into one category.

We conducted a Fisher's Exact Test to assess whether there was a statistically significant difference in the number of studies disaggregating data by sex/gender or providing raw data with individual sex/gender information over time (**Figure 9**). There was no significant change over time in the number of PD clinical research papers that either disaggregated data by sex/gender or provided raw data with individual sex/gender information (p = .505).

Question 5b: Has the rate of disaggregation of data by sex/gender changed over time in U.S.-based, NIH-funded clinical PD research? For this analysis, we focused on the 73 papers from our dataset that included men and women, were authored by researchers affiliated with US institutions, and had received NIH funding.

We conducted a Fisher's Exact Test to test for a statistically significant difference in the number of studies disaggregating their data by sex/gender over time (**Figure 10**). There was no significant difference in the disaggregation of data by sex/gender in U.S.-based, NIH-funded PD clinical research over time (p = 0.165).

Question 6: Has the rate of analysis of data considering sex/gender changed over time in clinical PD research?

Question 6a: Has the rate of analysis of data considering sex/gender changed over time in U.S.-based clinical PD research? For this analysis, we considered the 214 papers in our corpus that were U.S.-based and included men and women with PD. Of those papers, nine included sex/gender as a main factor in at least one analysis (1982-1992: n = 2; 2009: n = 1; 2021: n = 6) and eight included sex/gender as a covariate in at least one analysis (2009: n = 3; 2021: n = 5).

We conducted a Fisher's Exact Test to determine whether the proportion of studies that performed sex/gender-based analyses (those that include sex/gender as a main factor) has changed over time in our corpus. The Fisher's Exact Test was not significant (p = .109), indicating no change over time in the use of sex/gender as a main factor in analyses. (**Figure 11**).

We identified four papers that made claims about sex/gender differences or a lack thereof without statistical evidence. Three papers committed the Difference in Sex-Specific Significance (DISS) error (Maney & Rich-Edwards, 2023). Two of these papers, both published in 2021, reported a difference between sexes/genders after conducting statistical analyses within each sex/gender; however, the authors did not compare the sexes/genders directly with each other. In another paper, published in 2009, the authors stated there was no difference by sex/gender but did not specify the statistical methods used to support this assertion. Additionally, one paper published in 1982, reported a difference between men and women without presenting any statistical analyses to substantiate this claim.

Question 6b: Has the rate of analysis of data considering sex/gender changed over time in NIH-funded, U.S.-based clinical PD research? For this analysis, we considered the 81 papers in our corpus that were U.S.-based, included men and women with PD, and received funding from the NIH. Of those studies, three included sex/gender as a main factor (1982-1992: n = 1; 2021: n = 2) and four included sex/gender as a covariate (2009: n = 2; 2021: n = 2). Two studies published in 2021 committed DISS errors.

We conducted a Fisher's Exact Test to determine whether the proportion of studies that performed sex/gender-based analyses with sex/gender as a main factor changed over time in

NIH-funded, U.S.-based studies. The Fisher's Exact Test was not significant ($p = .480$), indicating no change in sex/gender-based analyses over time.

Analysis of Race/Ethnicity Reporting Over Time

Although an analysis of race/ethnicity reporting was secondary to our primary research objectives, we felt it important to assess this type of reporting over time given that the 1993 NIH Revitalization Act mandated the inclusion of both women *and* minorities in clinical research (Kelty et al., 2012). Our analysis aimed to investigate trends in race/ethnicity reporting in PD clinical research across different time periods. Our analysis was conducted on the 223 papers in our corpus published by authors affiliated with U.S.-based institutions.

We conducted Fisher's Exact Test to determine whether race/ethnicity reporting has changed over time. Our analysis indicated significant changes in race/ethnicity reporting across the time periods ($p < 0.001$), particularly between 1982-1992 vs. 2021 and 2009 vs. 2021 ($p_{adj} < .001$ and $p_{adj} = .003$, respectively) (**Figure 12**).

Discussion

In our study, we evaluated the representation of women and consideration of sex/gender in PD clinical research, particularly in the longitudinal context spanning significant policy changes by the NIH. Our findings offer insights into the landscape of PD research with respect to sex/gender and, to a lesser degree, race/ethnicity inclusion.

In our cross-sectional analysis of 2021 PD clinical research, we identified a statistically significant underrepresentation of women with PD in research. The participation of women in PD research was significantly lower than the proportion of women in the general PD population (Dorsey et al., 2018). We found no significant differences in the representation of women across pharmacological and physiotherapeutic sub-domains in 2021. This uniformity suggests that the

underrepresentation of women is a general issue in PD research, rather than being confined to specific types of interventions. This conclusion echoes that of Tosserams and colleagues (2018), who reported a similar sex/gender disparity in PD randomized clinical trials between 2010 and 2016.

Our longitudinal analysis spanning 1982-1992, 2009, and 2021 showed significant temporal improvements in the reporting of sex/gender inclusion, consistent with broader trends in clinical research showing a general increase in female inclusion, particularly following the 1993 NIH Revitalization Act (Beery & Zucker, 2011; Helmuth, 2000; National Academies of Sciences, Engineering, and Medicine, 2022).

The SABV policy mandates disaggregation of results by sex/gender (NOT-OD-15-102). Our study, however, revealed no significant change in the presentation of disaggregated data over time. This finding highlights a gap between policy objectives and research practices, suggesting that despite policy mandates, data disaggregation by sex/gender is not increasing in PD clinical research. Our research supports Helmuth's (Helmuth, 2000) observation that, although there has been a rise in the participation of women in clinical research, only a small proportion of NIH-funded studies report results separately for men and women.

We found no change in the rate of sex/gender-based analyses over time. This finding suggests that the requirements to consider sex/gender in research design and analysis have had no significant impact. Consistent with this finding, Geller and colleagues (2018) found that even when high-impact journals publish NIH-funded randomized controlled trials with substantial female enrollment, these studies often do not comply with NIH guidelines regarding the analysis and reporting of data by sex/gender.

Although our research mainly focused on sex/gender disparities, we also recognize the significance of racial and ethnic representation in clinical studies. It is well established that race as a social construct can explain some of the variability observed in health outcomes. For example, Chen and colleagues (2023) provide an in-depth analysis of disparities across breast cancer management stages, revealing significant diagnostic and treatment delays for Black, Hispanic, Asian, and Native American women. They highlight the interplay between these disparities and broader societal inequalities, such as socioeconomic status and access to healthcare, demonstrating how race as a social construct may influence health outcomes. Importantly, access to healthcare may be impeded by a lack of representation in clinical trials; insurance companies may restrict coverage of new therapies to only the demographic populations included in the clinical trial (National Academies of Sciences, Engineering, and Medicine, 2022).

Despite broader trends in clinical research indicating a decrease in the proportion of white participants in FDA-approved drug trials—from 84% in 2014 to 73.7% in 2020—the representation of minority groups remains low (National Academies of Sciences, Engineering, and Medicine, 2022). Our findings reflect a similar pattern, with a lack of comprehensive racial and ethnic representation in PD research. While our assessment of race/ethnicity *reporting* in PD clinical research from 1982-1992 to 2021 showed marked improvements, our study demonstrated that only a quarter of PD research reports on participants' race/ethnicity, and a majority of those studies reported predominantly Caucasian cohorts.

In summary, our study highlights both the challenges and the gradual advancements in addressing sex/gender and, to a lesser degree, race/ethnicity representation in PD clinical research. The findings underscore the necessity for ongoing efforts to enhance inclusivity in research practices.

Limitations of the Evidence Included in the Review

While our review was geographically inclusive for the 2021 cross-sectional analysis, it primarily focused on U.S.-based studies for longitudinal analysis. This approach could introduce a selection bias, limiting the global applicability of our findings. Furthermore, this review did not consider sex/gender inclusion policies of other agencies within the US or in other countries, which might influence PD research practices globally. The scope, concentrated on pharmacological and physiotherapeutic interventions in PD, might not capture the full spectrum of PD research.

Implications of the Results for Practice, Policy, and Future Research

Our findings have significant implications for clinical practice, policy formulation, and future research in PD. It highlights the need for intentional recruitment of women and minorities, including sexual and gender minorities, in PD research. The prevalent overlooking of sexual minorities, as discussed by the Committee on Improving the Representation of Women and Underrepresented Minorities in Clinical Trials and Research (2022), necessitates a more inclusive approach.

The sex/gender imbalance in research participation can be influenced by societal and behavioral factors, which emphasizes the need for equitable recruitment strategies. For example, patients who continue to drive with PD are more likely to be men (Crizzle et al., 2012). Further, the DATATOP (Parkinson study group, 1989) investigators noted that the disproportionate representation of men in their study may reflect a higher tendency for risk-taking among men and an increased likelihood for men to have caretaking spouses. This disparity underscores the necessity of implementing equitable recruitment strategies that effectively address these underlying factors. Moreover, recruiting a diverse participant pool demands considerable time

and effort to establish trust and overcome barriers to access (National Academies of Sciences, Engineering, and Medicine, 2022). The restrictions imposed by Institutional Review Boards on participant compensation to prevent coercion can inadvertently contribute to this challenge, particularly affecting certain demographics (National Academies of Sciences, Engineering, and Medicine, 2022). Consequently, researchers must strive to ethically ensure representativeness in their study cohorts.

Regarding policy, our findings suggest that, in practice, NIH policies, particularly the 2016 SABV policy, are not universally adhered to. In particular, the policy to present data disaggregated by sex/gender is largely ignored in the field of clinical research in PD.

In conclusion, our study indicates a need for more inclusive research practices and policies that not only encourage the participation of diverse groups but also ensure that their data are analyzed and reported in a manner that truly reflects their unique experiences and needs in PD research.

Data Availability Statement

All data, code, and statistical analyses is available and can be found at https://github.com/jeannempowell/SABVinPD.

Acknowledgements

We would like to thank the Radcliffe Research Partners Alicetonia Nwamah, Avni Kamat, and Madison Echols, and the Harvard Radcliffe Institute for their help in detecting DISS errors in our corpus. Authors have no competing interests or funding sources to report.

Author Contributions

JMP: Conceptualization, Project administration, Supervision, Methodology, Search Strategy, Paper Acquisition, Title/Abstract Screening, Full-text Screening, Data Extraction, Validation, Visualization, Analysis, Writing – Original Draft, Writing – Review and Editing

SSP: Data Extraction, Validation, Writing – Original Draft, Writing – Review and Editing

CLJ: Data Extraction, Validation

GS: Title/Abstract Screening, Full-text Screening, Validation, Data Extraction

IG: Title/Abstract Screening, Full-text Screening, Validation

AP: Data Extraction, Validation

SH: Full-text screening, Data Extraction, Validation

MSW: Search Strategy, Paper Acquisition

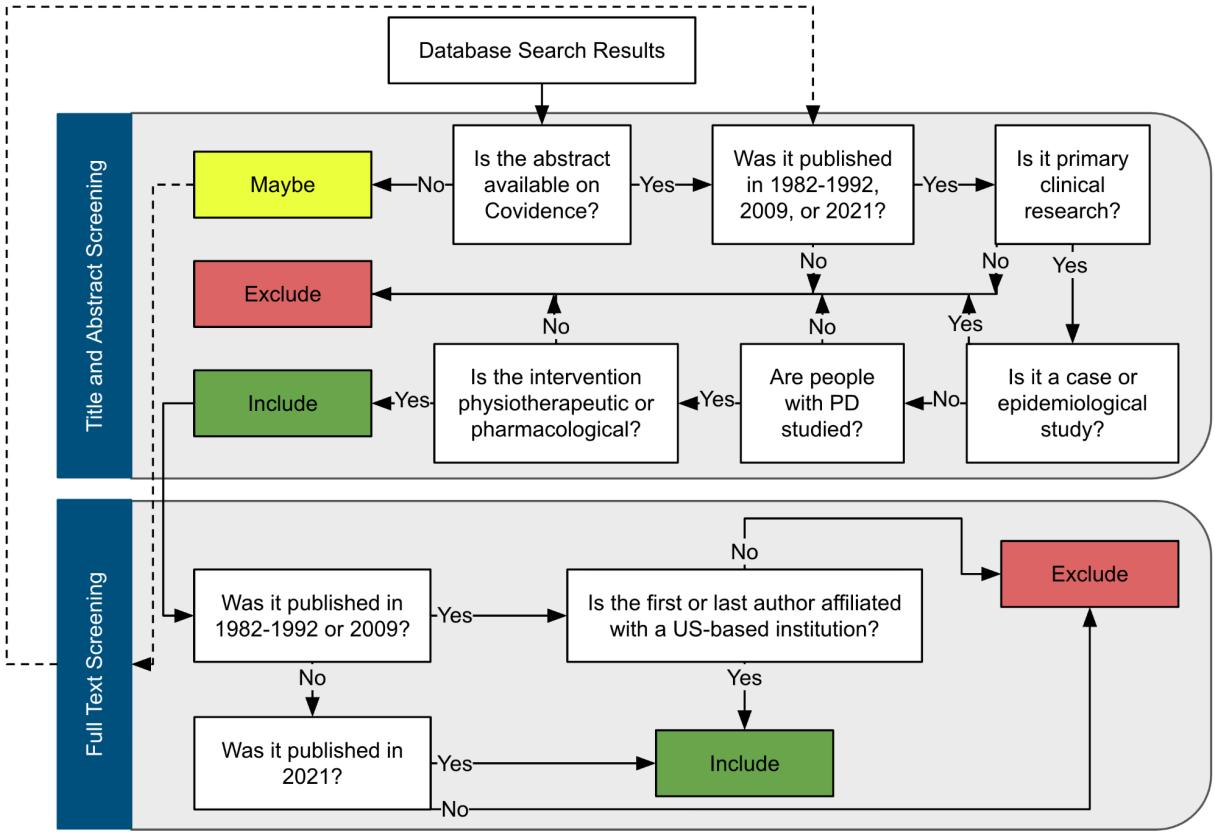
DLM: Conceptualization, Analysis, Writing – Original Draft, Writing – Review and Editing

JLM: Conceptualization, Methodology, Validation, Analysis, Writing – Original Draft, Writing – Review and Editing

Figures

Figure 1

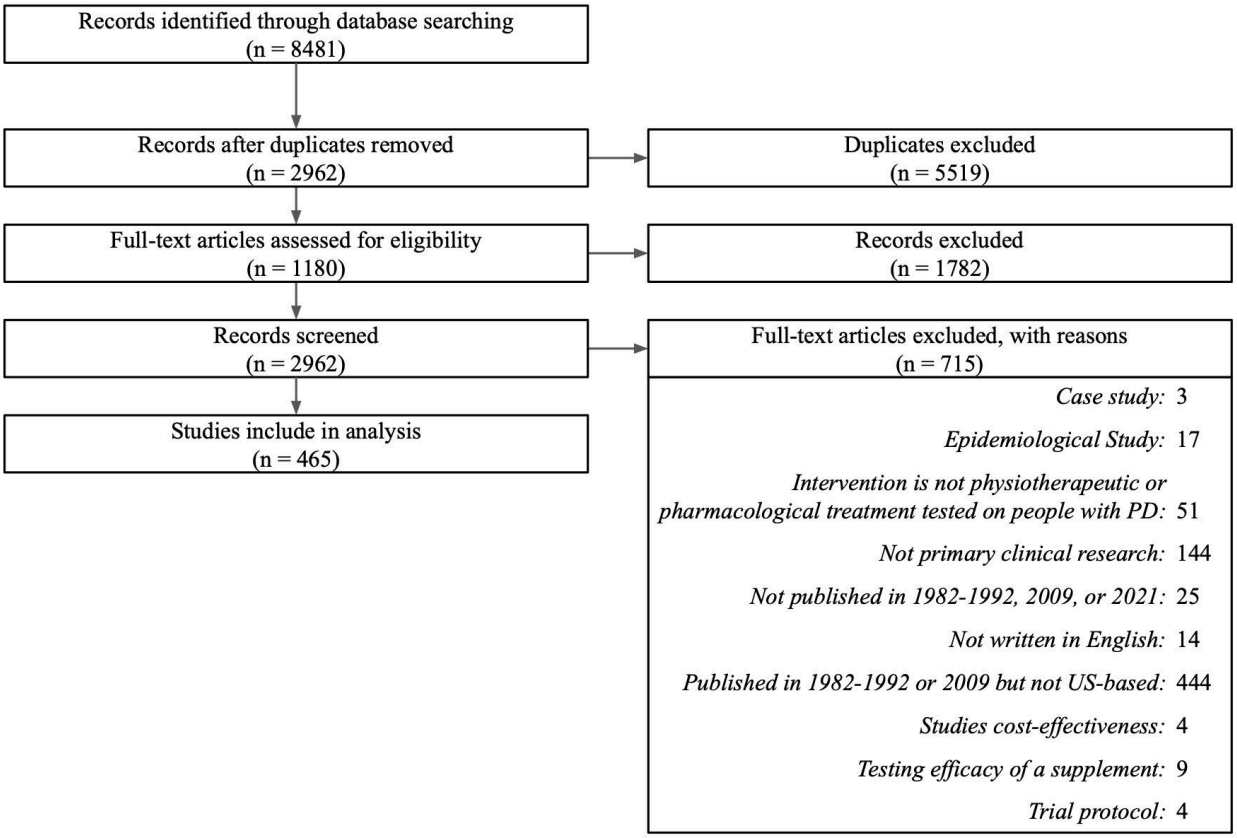
Overview of Inclusion and Exclusion Criteria for Systematic Review on Women's Representation in Parkinson's Disease Clinical Research.



Note. This figure illustrates the timeframe for study selection (1982-1992, 2009, 2021) and the criteria applied: inclusion of primary clinical research studies (excluding meta-analyses, case, and epidemiological studies), with a focus on pharmacological or physiotherapeutic interventions in PD. It also details the distinction in author affiliation requirements for pre- and post-1993 NIH policy era studies and the exclusion of non-PD patient studies.

Figure 2

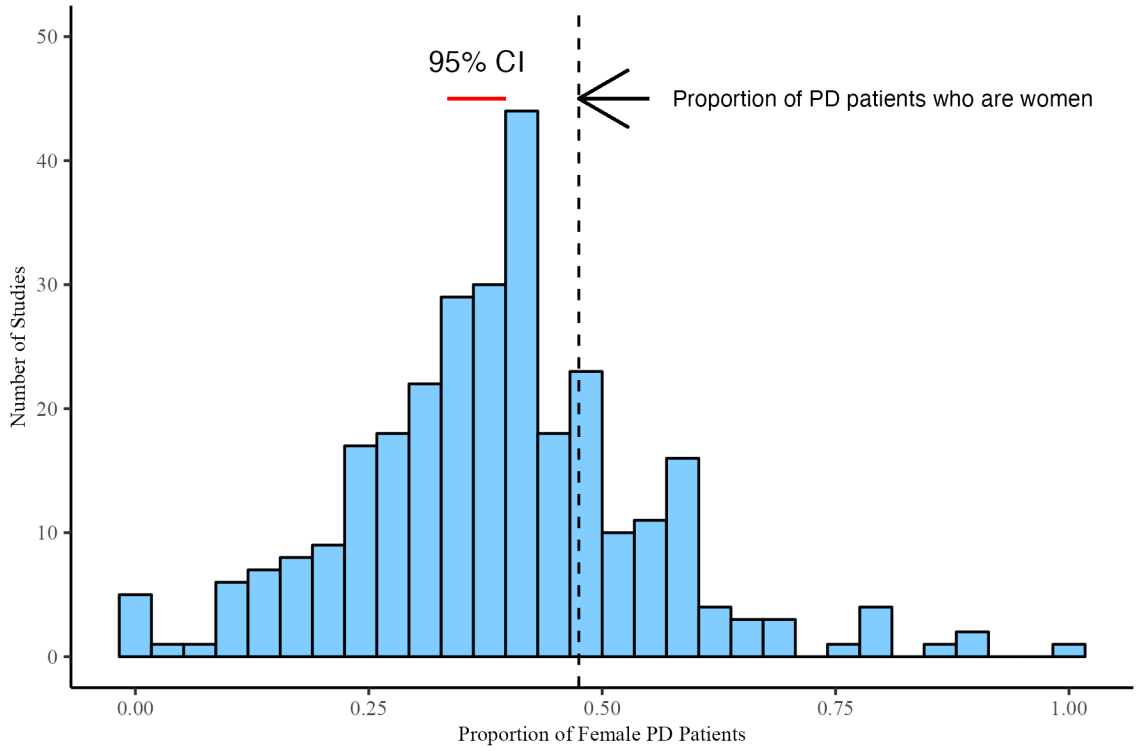
PRISMA Flow Diagram of Literature Screening and Study Selection Process



Note. This PRISMA flow diagram displays the literature screening and study selection process. Initially, 8,481 studies were identified, decreasing to 2,962 after removing duplicates. Post-screening and full-text assessment excluded 1,782 and 715 studies, respectively, for reasons such as non-relevance to physiotherapeutic or pharmacological treatments in PD, non-primary clinical research, and inappropriate publication characteristics. The final review included 465 studies.

Figure 3

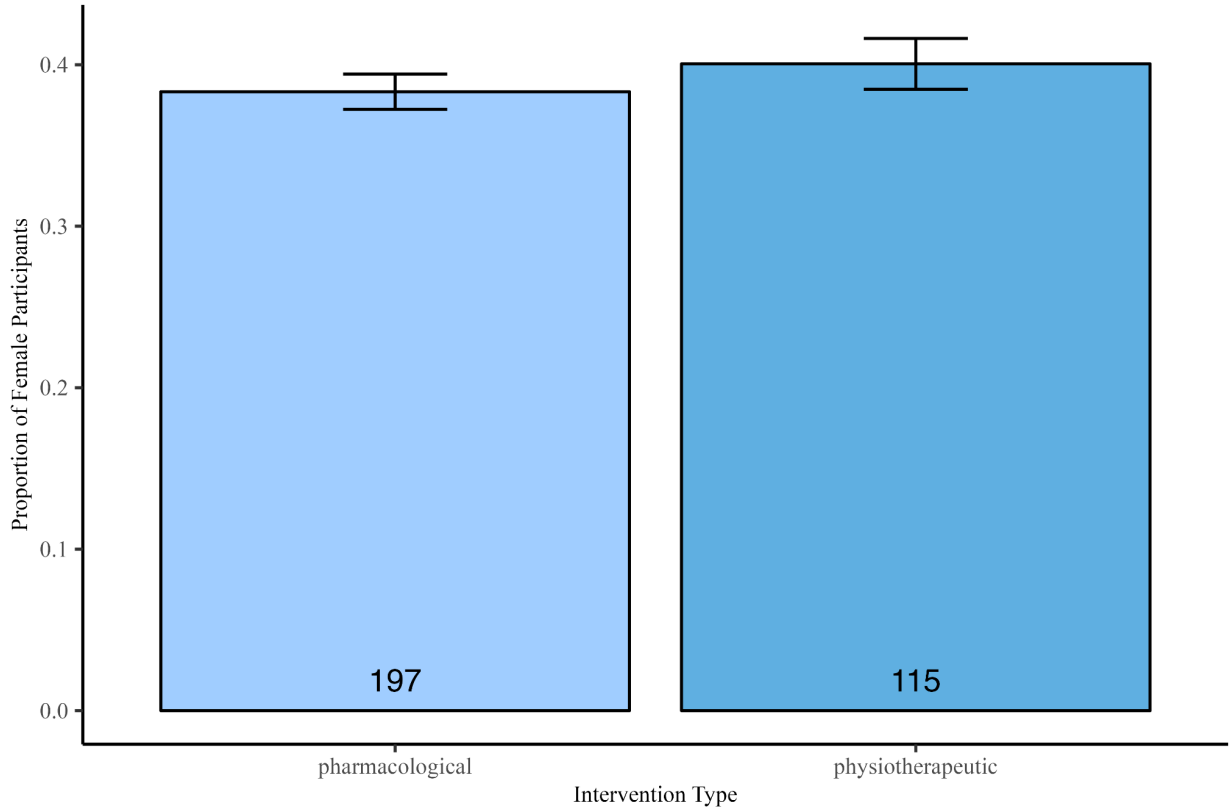
Women are underrepresented as participants in PD clinical research studies (2021)



Note. Bars indicate study counts per female participation range. The vertical dashed line at 0.475 marks the proportion of women in the PD population (Dorsey et al., 2018). Logit-transformed data indicate a deviation from this benchmark, with a reverse-transformed 95% confidence interval ranging from 0.334 to 0.396.

Figure 4

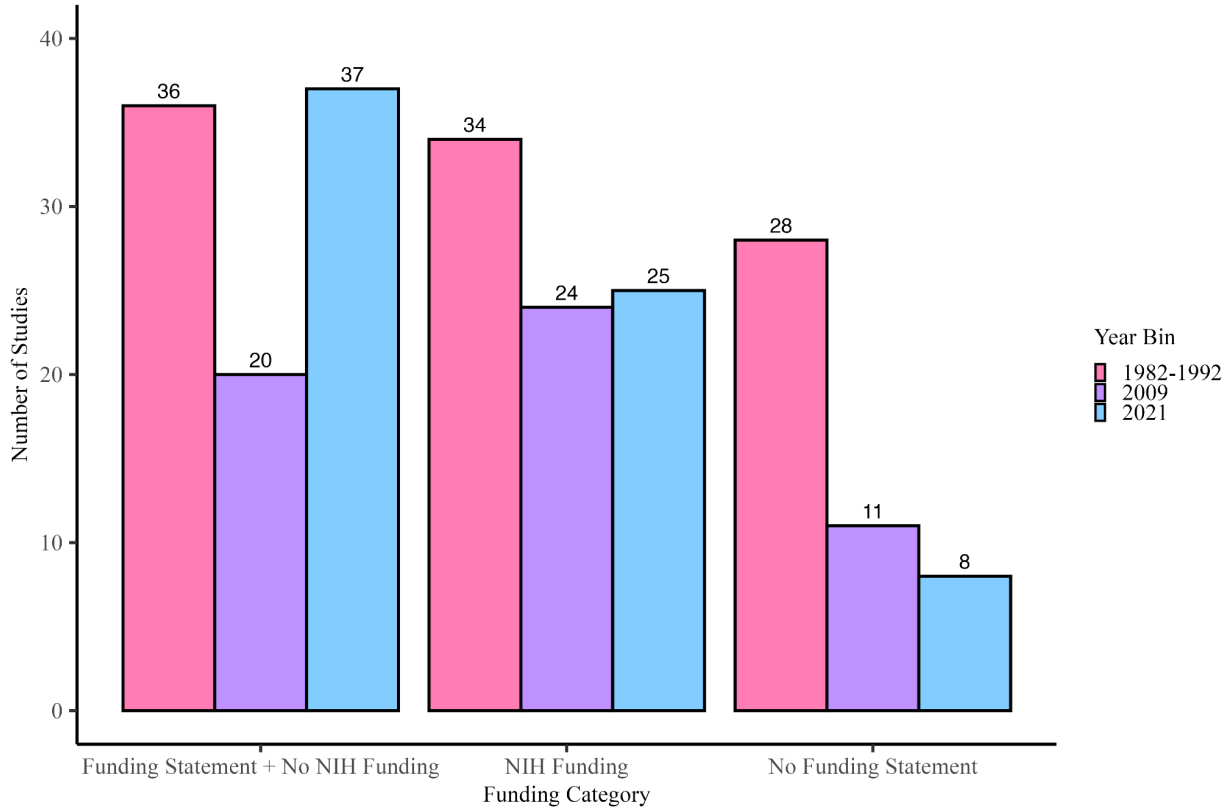
Comparison of Female Participation in PD Clinical Research: Physiotherapeutic vs Pharmacological Interventions (2021)



Note. This bar plot displays the proportion of female participants in PD research across physiotherapeutic and pharmacological interventions in 2021. The error bars represent the standard error. Statistical analysis showed no significant difference in female representation between these intervention types ($t(261.39) = 0.7418$, $p = 0.459$).

Figure 5

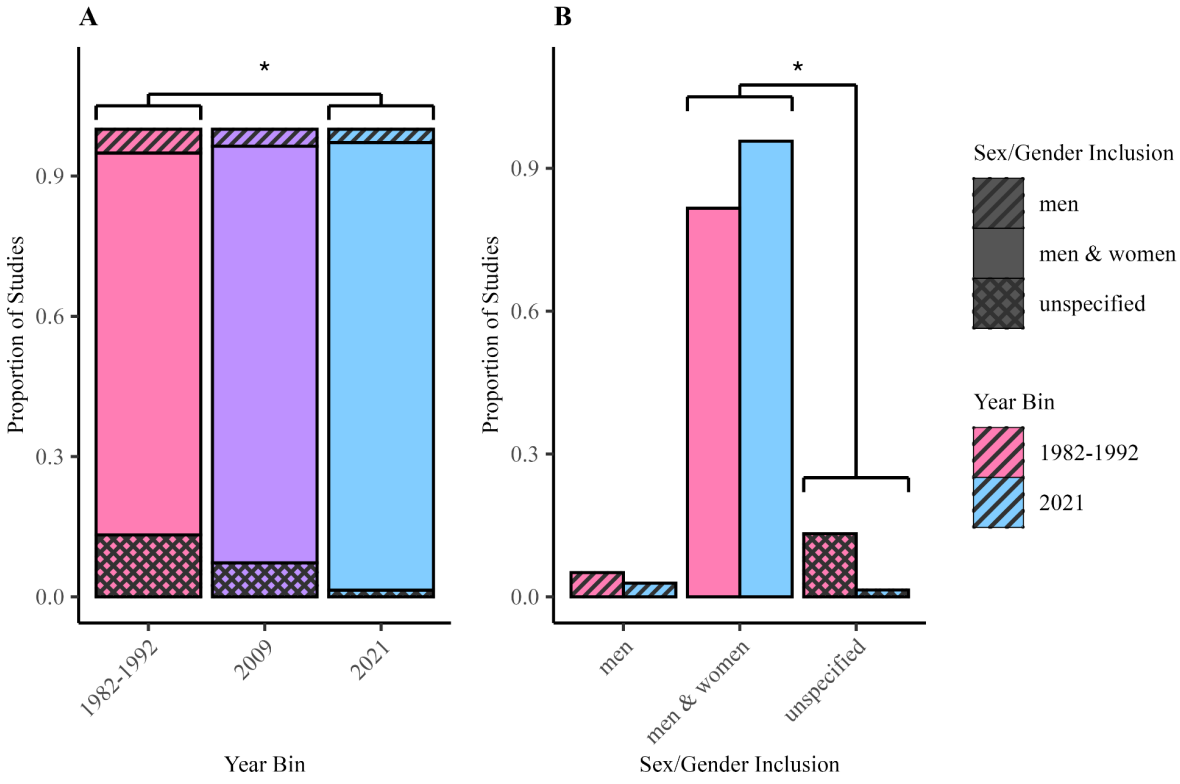
U.S.-based PD Studies Disaggregated by Funding Category



Note. This bar chart categorizes U.S.-based PD studies by NIH funding status across three distinct time periods: 1982-1992 (pink), 2009 (purple), and 2021 (blue).

Figure 6

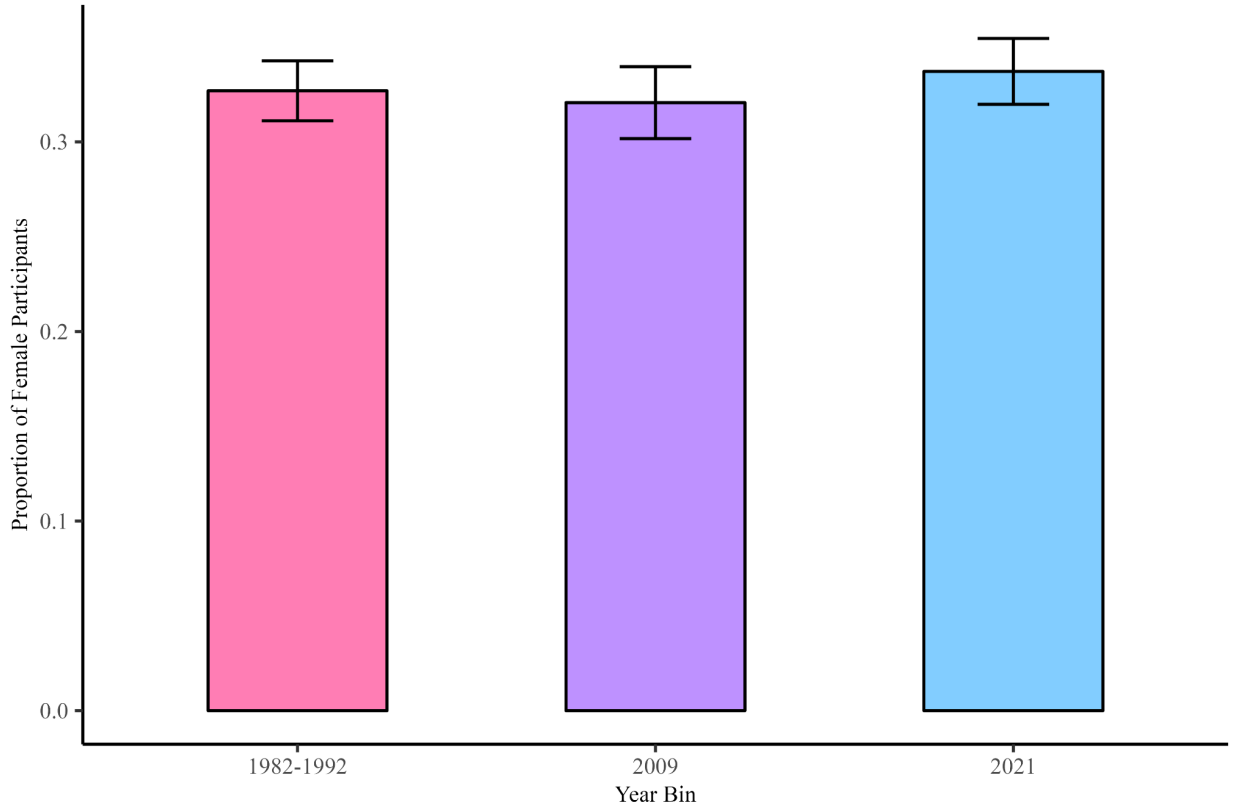
Trends in Sex/Gender Inclusion in U.S.-Based PD Clinical Research (1982-2021)



Note. This figure illustrates the outcome of our analysis on the temporal trends in the reporting of sex/gender in PD clinical research. The omnibus Fisher's Exact Test was statistically significant, indicating a change in the sex/gender inclusion reported in these papers over time ($p = 0.049$). **A**) Our analysis revealed a difference in the reporting of sex/gender inclusion between studies published during the earliest (1982-1992) and the latest (2021) time frames ($p_{adj} = 0.0497$). **B**) Specifically, studies from 1982 to 1992 were significantly less likely to report the inclusion of both sexes and instead not specify sex/gender-based inclusion compared to those published in 2021 ($p_{adj} = 0.0268$, Odds Ratio (OR) = 0.0928).

Figure 7

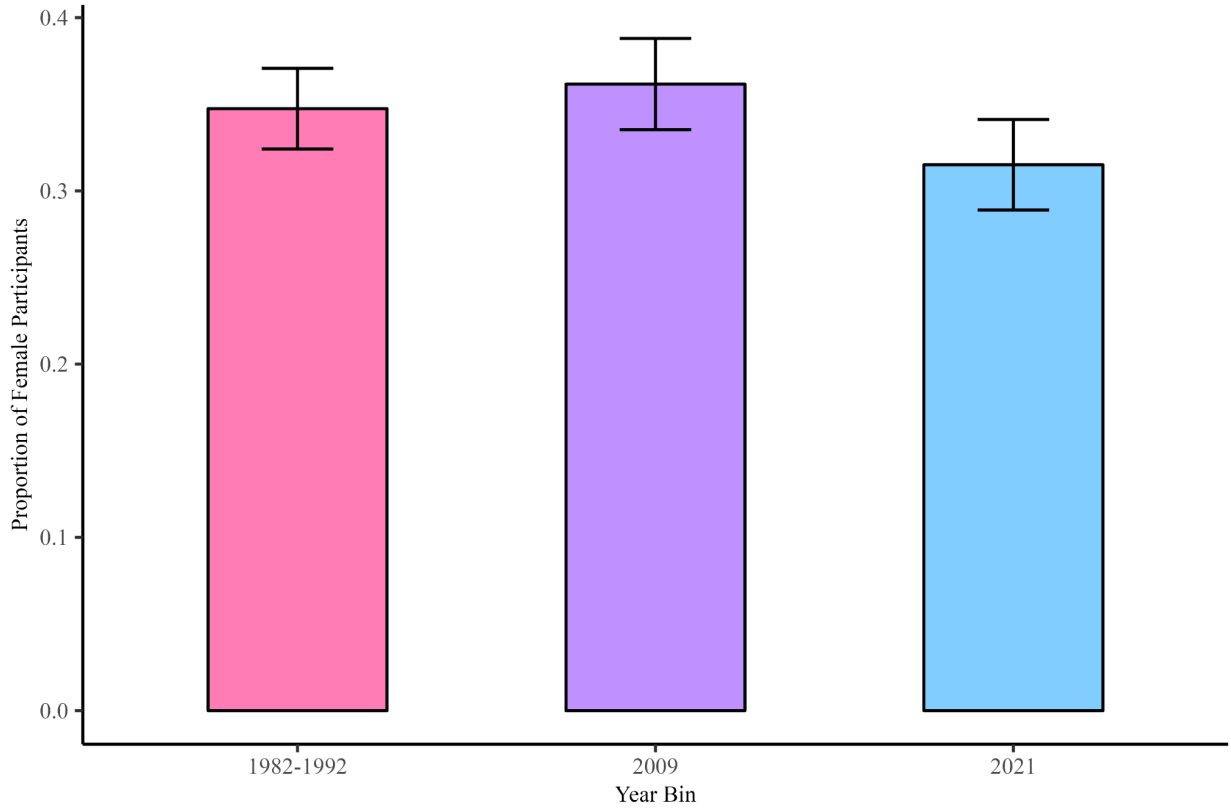
Trends in Female Participation in U.S.-Based Parkinson's Disease Clinical Studies (1982-2021)



Note. This bar chart illustrates the mean proportion of female participants in U.S.-based PD clinical studies across three time periods: 1982-1992, 2009, and 2021. The bars, differentiated by color for each period, represent the average proportion of female participants. Error bars indicate the standard error. According to our logit-transformed ANOVA, the differences in female participation rates across these time periods were not statistically significant ($F(2, 195) = 0.31$, $p = .734$), suggesting a consistent level of female recruitment in PD research over the years.

Figure 8

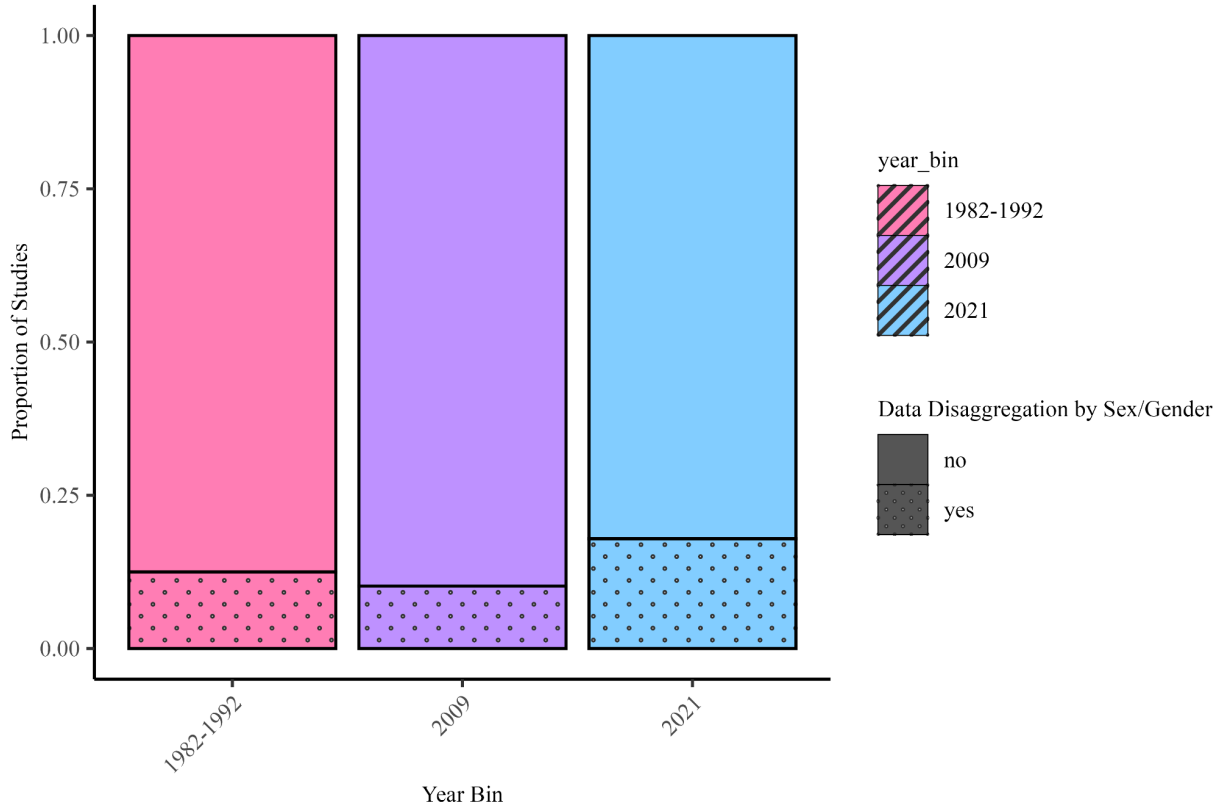
Female Participation Trends in NIH-Funded US PD Clinical Studies (1982-2021)



Note. This bar chart illustrates the mean proportion of female participants in NIH-funded PD clinical studies across three distinct time periods: 1982-1992, 2009, and 2021. The bars, colored to differentiate each time period, display the average proportion of female participants with PD. Error bars represent the standard error for each period. Our analysis showed no significant change in the proportion of female participants in NIH-funded PD studies over the specified periods ($F(2, 69) = 0.698$, $p = .501$).

Figure 9

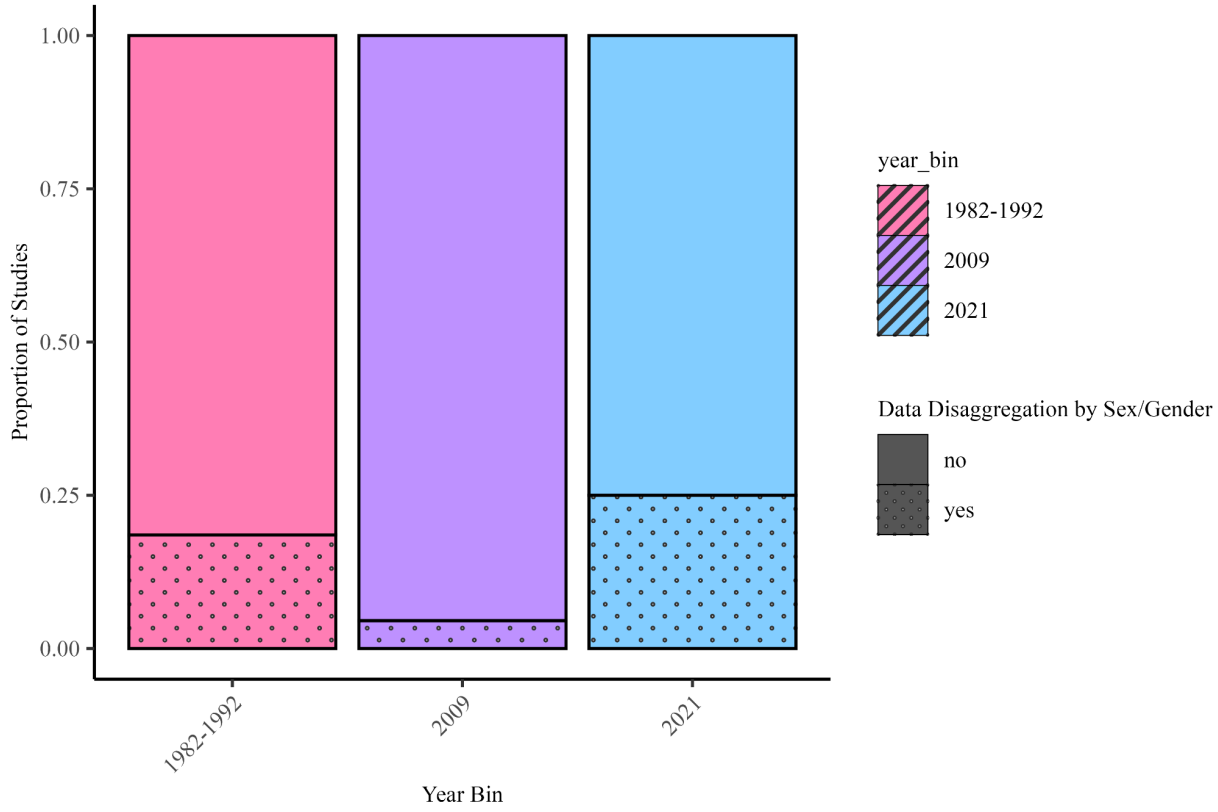
Temporal Trends in Sex/Gender Data Disaggregation in U.S.-Based PD Clinical Research (1982-2021)



Note. This line chart depicts the trend in sex/gender data disaggregation in U.S.-based PD clinical studies from 1982 to 2021. It displays the percentage of studies in which results were disaggregated by sex/gender. Each type of line represents a category of data treatment (disaggregated or not), with color denoting different time periods. The Fisher's Exact Test (p = 0.505) indicated no significant shifts in the practice of sex/gender data disaggregation in PD clinical research across these periods.

Figure 10

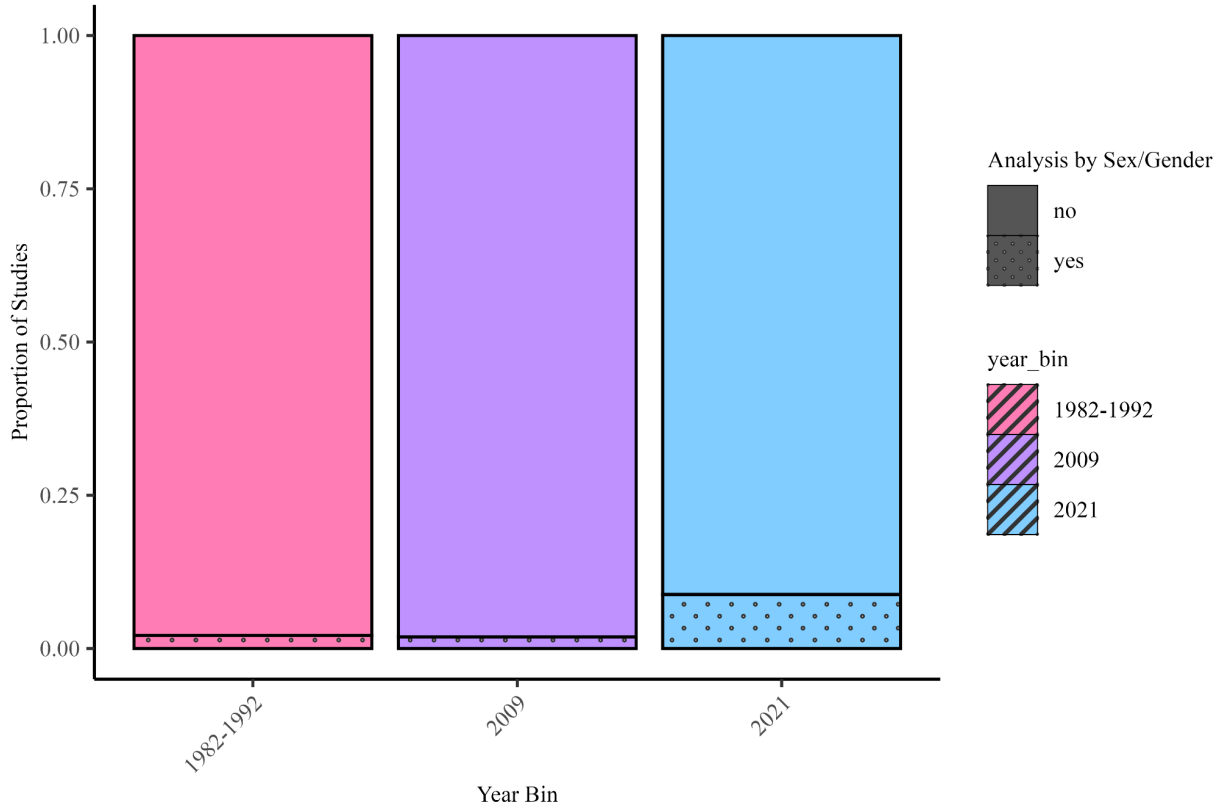
Temporal Trends in Sex/Gender Data Disaggregation in U.S.-Based, NIH-funded PD Clinical Research (1982-2021)



Notes. This line chart depicts the trend in sex/gender data disaggregation in NIH-funded, U.S.-based PD clinical studies from 1982 to 2021. It displays the percentage of studies in which results were disaggregated by sex/gender. Each type of line represents a category of data treatment (disaggregated or not), with color denoting different time periods. The Fisher's Exact Test ($p = 0.165$) indicated no significant shifts in the practice of sex/gender data disaggregation in PD clinical research across these periods.

Figure 11

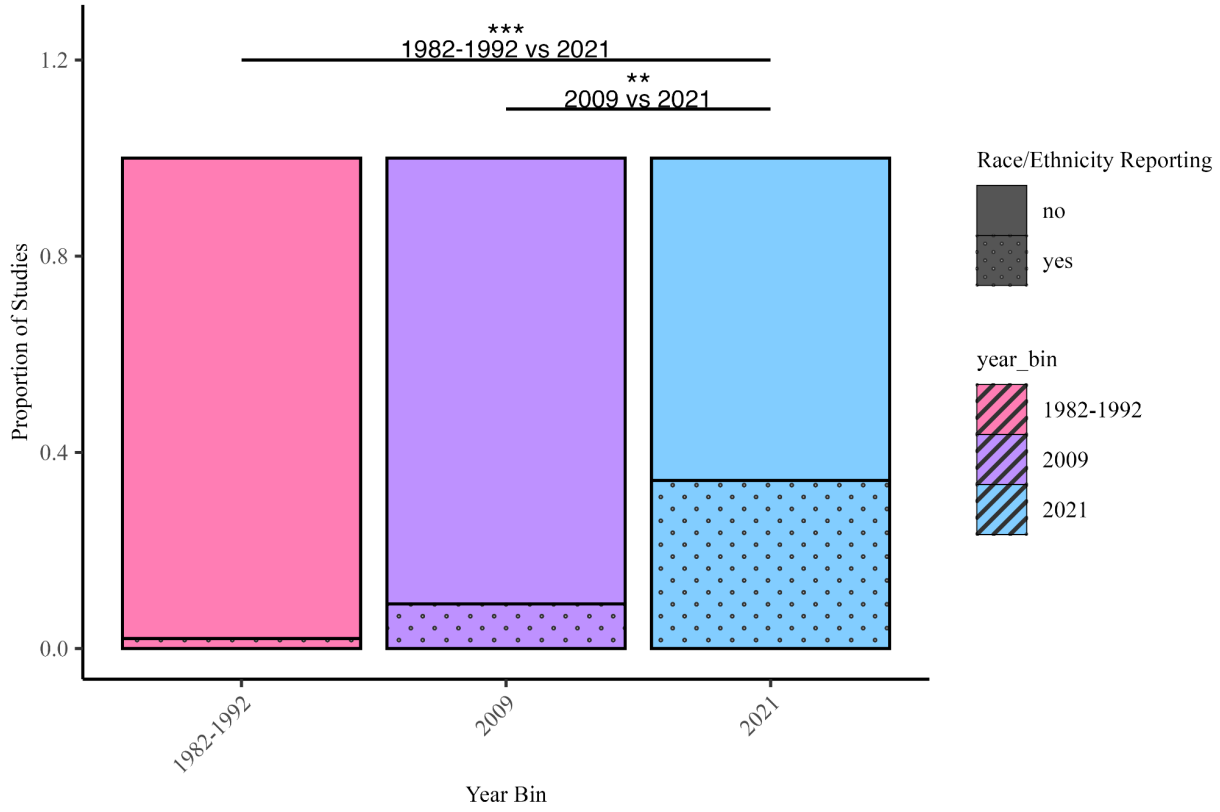
Trends in Sex/Gender-Based Analyses in US PD Research (1982-2021)



Note. This bar chart depicts the proportion of U.S.-based PD studies, from 1982 to 2021, in which sex/gender-based analyses were conducted with sex/gender as a main factor. The bars, color-coded by year bin, demonstrate the percentage of studies incorporating these analyses. We found that there was no significant association between year bin and proportion of sex/gender-based analyses ($p = 0.109$).

Figure 12

Evolution of Race/Ethnicity Reporting in U.S.-Based PD Clinical Research (1982-2021)



Notes. This line chart, based on an analysis of 223 U.S.-based studies, reveals the percentage of papers reporting race/ethnicity data. Color-coded lines and points denote the time periods, while line type represents the presence or absence of race/ethnicity reporting. Significant changes in reporting practices are marked: there was a significant increase between 1982-1992 and 2021 ($p_{adj} < .001$) and between 2009 and 2021 ($p_{adj} = .003$), as indicated by Fisher's Exact Test. These findings highlight a significant shift towards more inclusive demographic reporting in recent PD research.

References

Beery, A. K., & Zucker, I. (2011). Sex bias in neuroscience and biomedical research. *Neuroscience & Biobehavioral Reviews*, *35*(3), 565–572. https://doi.org/10.1016/j.neubiorev.2010.07.002

Blakeman, J. R., & Fillman, V. M. (2021). An analysis of the use of the terms sex and gender in research reported in nursing journals. *Advances in Nursing Science*, *45*(3), E110–E126. https://doi.org/10.1097/ANS.0000000000000402

Bloem, B. R., Okun, M. S., & Klein, C. (2021). Parkinson's disease. *The Lancet*, *397*(10291), 2284–2303. https://doi.org/10.1016/S0140-6736(21)00218-X

Cerri, S., Mus, L., & Blandini, F. (2019). Parkinson's disease in women and men: What's the difference? *Journal of Parkinson's Disease*, *9*(3), 501–515. https://doi.org/10.3233/jpd-191683

Chen, J., Stover, D. G., Ballinger, T. J., Bazan, J. G., Schneider, B. P., Andersen, B. L., Carson, W. E., & Obeng-Gyasi, S. (2023). Racial disparities in breast cancer: From detection to treatment. *Current Oncology Reports*, *26*(1), 10–20. https://doi.org/10.1007/s11912-023-01472-8

Crizzle, A. M., Myers, A. M., & Almeida, Q. J. (2012). Drivers with Parkinson's disease: Who participates in research studies? *Parkinsonism & Related Disorders*, *18*(7), 833–836. https://doi.org/10.1016/j.parkreldis.2012.04.003

Dorsey, E. R., Elbaz, A., Nichols, E., Abbasi, N., Abd-Allah, F., Abdelalim, A., Adsuar, J. C., Ansha, M. G., Brayne, C., Choi, J.-Y. J., Collado-Mateo, D., Dahodwala, N., Do, H. P., Edessa, D., Endres, M., Fereshtehnejad, S.-M., Foreman, K. J., Gankpe, F. G., Gupta, R., … Murray, C. J. L. (2018). Global, regional, and national burden of Parkinson's disease,

1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, *17*(11), 939–953. https://doi.org/10.1016/S1474-4422(18)30295-3

Geller, S. E., Koch, A. R., Roesch, P., Filut, A., Hallgren, E., & Carnes, M. (2018). The more things change, the more they stay the same: A study to evaluate compliance with inclusion and assessment of women and minorities in randomized controlled trials. *Academic Medicine*, *93*(4), 630–635. https://doi.org/10.1097/ACM.0000000000002027

Helmuth, L. (2000). Reports see progress, problems, in trials. *Science*, *288*(5471), 1562–1563. https://doi.org/10.1126/science.288.5471.1562

Kelty, M., Bates, A., & Pinn, V. W. (2012). National Institutes of Health policy on the inclusion of women and minorities as subjects in clinical research. In *Principles and Practice of Clinical Research* (pp. 147–159). Academic Press. https://doi.org/10.1016/B978-0-12-382167-6.00013-8

Maney, D. L., & Rich-Edwards, J. W. (2023). Sex-inclusive biomedicine: Are new policies increasing rigor and reproducibility? *Women's Health Issues*, *33*(5), 461–464. https://doi.org/10.1016/j.whi.2023.03.004

National Academies of Sciences, Engineering, and Medicine. (2022). *Improving representation in clinical trials and research: Building research equity for women and underrepresented groups* (K. Bibbins-Domingo & A. Helman, Eds.; p. 26479). National Academies Press. https://doi.org/10.17226/26479

Parkinson study group. (1989). DATATOP: A Multicenter controlled clinical trial in early Parkinson's disease. *Archives of Neurology*, *46*(10), 1052. https://doi.org/10.1001/archneur.1989.00520460028009

Rajith, R. K., & Angiel, P. R. (2023). Life experiences of informal caregivers of Parkinson

disease patients: A qualitative study. *Indian Journal of Health Studies*, *05*(01), 30–56. https://doi.org/10.56490/IJHS.2023.5103

Tosserams, A., Araújo, R., Pringsheim, T., Post, B., Darweesh, S. K. L., IntHout, J., & Bloem, B. R. (2018). Underrepresentation of women in Parkinson's disease trials. *Movement Disorders*, *33*(11), 1825–1826. https://doi.org/10.1002/mds.27505

Appendix A: Database Search Strategy

Database	Search	Search Date
PubMed	Search: ("Parkinson Disease/drug therapy"[Mesh] OR "Parkinson Disease/therapy"[Mesh]) Filters: Clinical Trial, Randomized Controlled Trial, from 1985; 2009; 2021	9-15-22
PubMed	1. ("Parkinson Disease/diet therapy"[Mesh] OR "Parkinson Disease/drug therapy"[Mesh] OR "Parkinson Disease/prevention and control"[Mesh] OR "Parkinson Disease/rehabilitation"[Mesh] OR "Parkinson Disease/therapy"[MeSH]) NOT Cognition[MESH] 2. ("parkinson disease"[MESH] OR parkinson*[TW]) NOT parkinsonism[TW] 3. therapeutics[MESH:noexp] OR "acoustic stimulation"[MESH] OR "Climatotherapy"[MESH] OR "complementary therapies"[MESH] OR "Electric Stimulation Therapy"[MESH] OR "phototherapy"[MESH] OR "radiotherapy"[MESH] OR "Physical Therapy Modalities"[MESH] OR "rehabilitation"[MESH] OR "drug therapy"[MESH] OR "animal assisted therapy"[MESH] OR "Musculoskeletal Manipulations"[MESH] OR "Kinesiology, Applied"[MESH] OR "therapy, soft tissue"[MESH] OR "Manipulation, Osteopathic"[MESH] OR "Manipulation, Chiropractic"[MESH] OR "musculoskeletal manipulation*"[TW] OR "chiropractic manipulation*"[TW] OR "orthopedic manipulation*"[TW] OR "osteopathic manipulation*"[TW] OR "spine manipulation*"[TW] OR "acoustic stimulation*"[TW] OR climatotherap*[TW] OR "complementary therap*"[TW] OR "drug therap*"[TW] OR phototherap*[TW] OR radiotherap*[TW] OR "applied kinesiology"[TW] OR "soft tissue therap*"[TW] OR "physical therap*"[TW] OR "electric stimulation*"[TW] OR "exercise therap*"[TW] OR "exercise therapeut*"[TW] OR "shockwave therap*"[TW] OR hydrotherap*[TW] OR "manual therap*"[TW] OR "myofunctional therap*"[TW] OR "art therap*"[TW] OR "music therap*"[TW] OR bibliotherap*[TW] OR "dance therap*"[TW] OR "occupational therap*"[TW] OR "recreation* therap*"[TW] OR rehabilitation[TW] OR telerehabilitation[TW] OR "therapeutic exercise*"[TW] OR "alternative therap*"[TW] OR "pet therap*"[TW] OR "Therapy animal*"[TW] OR "therapeutic animal*"[TW] 4. ("Clinical Studies as Topic"[MESH] NOT "Observational Studies, Veterinary as Topic"[MESH]) OR "clinic stud*"[TW] OR "clinical stud*"[TW] OR "clinical trial*"[TW] OR "control trial*"[TW] OR "controlled* trial*"[TW] OR "observation stud*"[TW] OR	2-8-23

	"observational stud*"[TW] OR stud*[TI] OR trial*[TI] OR placebo[TI] OR "double blind*"[TI] 5. ((#1 OR (#2 AND #3)) AND #4) NOT (cognition[TI] OR "case reports"[PT] OR "case stud*"[TI] OR "case report*"[TI] OR "case series"[TI] OR "systematic review"[PT] OR review[PT] OR "meta-analysis"[PT] OR "systematic review"[TI] OR "literature review"[TI] OR "scoping review"[TI] OR "narrative review"[TI] OR "meta analysis"[TI])(#5 AND English[LA] AND ((1982/1/2:1992/12/31[pdat]) OR (2009/1/1:2009/12/31[pdat]) OR (2021/1/1:2021/12/31[pdat]))	
Embase (Elsevier)	1. ('parkinson disease'/exp AND ('drug therapy'/lnk OR 'prevention'/lnk OR 'radiotherapy'/lnk OR 'rehabilitation'/lnk OR 'therapy'/lnk)) NOT 'cognition'/exp 2. ('parkinson disease'/exp OR parkinson*:ti,ab,kw) NOT parkinsonism:ti,ab,kw 3. ('therapy'/de OR 'auditory stimulation'/exp OR 'climatotherapy'/exp OR 'electrotherapy'/exp OR 'phototherapy'/exp OR 'radiotherapy'/exp OR 'physiotherapy'/exp OR 'rehabilitation'/exp OR 'drug therapy'/exp OR 'animal assisted therapy'/exp OR 'musculoskeletal manipulation'/exp OR 'kinesiology'/exp OR 'soft tissue therapy'/exp OR 'musculoskeletal manipulation*':ti,ab,kw OR 'chiropractic manipulation*':ti,ab,kw OR 'orthopedic manipulation*':ti,ab,kw OR 'osteopathic manipulation*':ti,ab,kw OR 'spine manipulation*':ti,ab,kw OR 'acoustic stimulation*':ti,ab,kw OR climatotherap*:ti,ab,kw OR 'complementary therap*':ti,ab,kw OR phototherap*:ti,ab,kw OR radiotherap*:ti,ab,kw OR 'applied kinesiology':ti,ab,kw OR 'soft tissue therap*':ti,ab,kw OR 'physical therap*':ti,ab,kw OR 'drug therap*':ti,ab,kw OR 'electric stimulation*':ti,ab,kw OR 'exercise therap*':ti,ab,kw OR 'exercise therapeut*':ti,ab,kw OR 'shockwave therap*':ti,ab,kw OR hydrotherap*:ti,ab,kw OR 'manual therap*':ti,ab,kw OR 'myofunctional therap*':ti,ab,kw OR 'art therap*':ti,ab,kw OR 'music therap*':ti,ab,kw OR bibliotherap*:ti,ab,kw OR 'dance therap*':ti,ab,kw OR 'occupational therap*':ti,ab,kw OR 'recreation* therap*':ti,ab,kw OR rehabilitation:ti,ab,kw OR telerehabilitation:ti,ab,kw OR 'therapeutic exercise*':ti,ab,kw OR 'alternative therap*':ti,ab,kw OR 'therapeutic exercise*':ti,ab,kw OR 'pet therap*':ti,ab,kw OR 'therap* animal*':ti,ab,kw 4. ('clinical trial (topic)'/exp OR 'clinical trial*':ti,ab,kw OR 'clinic* stud*':ti,ab,kw OR 'clinic* trial*':ti,ab,kw OR 'control* trial*':ti,ab,kw OR 'observation* stud*':ti,ab,kw OR placebo:ti OR	2-8-23

	'double blind*':ti OR 'cross section*':ti OR stud*:ti OR trial*:ti) NOT 'veterinar* stud*':ti,ab,kw 5. ((#1 OR (#2 AND #3)) AND #4 AND ('article'/it OR 'article in press'/it)) NOT (cognition:ti OR 'abstract report'/it OR 'book'/it OR 'chapter'/it OR 'conference paper'/it OR 'conference review'/it OR 'data papers'/it OR 'editorial'/it OR 'erratum'/it OR 'letter'/it OR 'note'/it OR 'preprint'/it OR 'report'/it OR 'review'/it OR 'short survey'/it OR 'tombstone'/it OR 'case reports':ti OR 'case study':ti OR 'case report*':ti OR 'case series':ti OR 'systematic review':ti OR 'literature review':ti OR 'scoping review':ti OR 'narrative review':ti OR 'meta analysis':ti) 6. #5 AND [english]/lim AND ([1982-1992]/py OR 2009:py OR 2021:py)	
Cochrane (Cochrane Library)	1. ([mh "parkinson disease"/dh] OR [mh "parkinson disease"/dt] OR [mh "parkinson disease"/pc] OR [mh "parkinson disease"/rh] OR [mh "parkinson disease"/TH]) NOT [mh cognition] 2. ([mh "parkinson disease"] OR parkinson:ti,ab,kw) NOT parkinsonism:ti,ab,kw 3. [mh ^therapeutics] OR [mh "acoustic stimulation"] OR [mh Climatotherapy] OR [mh "complementary therapies"] OR [mh "Electric Stimulation Therapy"] OR [mh phototherapy] OR [mh radiotherapy] OR [mh "Physical Therapy Modalities"] OR [mh rehabilitation] OR [mh "drug therapy"] OR [mh "animal assisted therapy"] OR [mh "Musculoskeletal Manipulations"] OR [mh "Kinesiology, Applied"] OR [mh "therapy, soft tissue"] OR [mh "Manipulation, Osteopathic"] OR [mh "Manipulation, Chiropractic"] OR "musculoskeletal manipulation*":ti,ab,kw OR "chiropractic manipulation*":ti,ab,kw OR "orthopedic manipulation*":ti,ab,kw OR "osteopathic manipulation*":ti,ab,kw OR "spine manipulation*":ti,ab,kw OR "acoustic stimulation*":ti,ab,kw OR climatotherapy:ti,ab,kw OR "complementary therap*":ti,ab,kw OR phototherapy:ti,ab,kw OR radiotherapy:ti,ab,kw OR "applied kinesiology":ti,ab,kw OR "soft tissue therap*":ti,ab,kw OR "physical therap*":ti,ab,kw OR "drug therap*":ti,ab,kw OR "electric stimulation*":ti,ab,kw OR "exercise therap*":ti,ab,kw OR "therapeutic exercise*":ti,ab,kw OR "shockwave therap*":ti,ab,kw OR hydrotherap*:ti,ab,kw OR "manual therap*":ti,ab,kw OR "myofunctional therap*":ti,ab,kw OR "art therap*":ti,ab,kw OR "music therap*":ti,ab,kw OR bibliotherap*:ti,ab,kw OR "dance therap*":ti,ab,kw OR "occupational therap*":ti,ab,kw OR "recreation* therap*":ti,ab,kw OR rehabilitation:ti,ab,kw OR telerehabilitation:ti,ab,kw OR	2-8-23

	"alternative therap*":ti,ab,kw OR "therapeutic exercise":ti,ab,kw OR "therap* animal*":ti,ab,kw OR "pet therap*":ti,ab,kw 4. ([mh "Clinical Studies as Topic"] NOT [mh "Observational Studies, Veterinary as Topic"]) OR "clinic* stud*":ti,ab,kw OR "clinic* trial*":ti,ab,kw OR "observation* stud*":ti,ab,kw OR "control trial":ti,ab,kw OR "controlled* trial":ti,ab,kw OR placebo:ti OR "double blind*":ti OR "cross section*":ti OR stud*:ti OR trial*:ti 5. ((#1 OR (#2 AND #3)) AND #4) NOT (cognition:ti OR "case report*":ti OR "case stud*":ti OR "case report*":ti OR "case series":ti OR "systematic review":ti OR "literature review":ti OR "scoping review":ti OR "narrative review":ti OR "meta analysis":ti) with Publication Year from 1982-1992, in Trials 6. ((#1 OR (#2 AND #3)) AND #4) NOT (cognition:ti OR "case report*":ti OR "case stud*":ti OR "case report*":ti OR "case series":ti OR "systematic review":ti OR "literature review":ti OR "scoping review":ti OR "narrative review":ti OR "meta analysis":ti) with Publication Year from 2009 to 2009, in Trials 7. ((#1 OR (#2 AND #3)) AND #4) NOT (cognition:ti OR "case report*":ti OR "case stud*":ti OR "case report*":ti OR "case series":ti OR "systematic review":ti OR "literature review":ti OR "scoping review":ti OR "narrative review":ti OR "meta analysis":ti) with Publication Year from 2021 to 2021, in Trials 8. #5 OR #6 OR #7	
CINAHL (Ebscohost)	S1. (MH "Parkinson Disease/DH/DT/PP/PC/RT/RH/TH") NOT (MH Cognition+) S2. (MH "Parkinson Disease" OR TX parkinson*) NOT TX parkinsonism S3. MH "Therapeutics" OR MH "Acoustic Stimulation"OR MH "Climatotherapy" OR MH "Alternative Therapies+" OR MH "Electrotherapy+" OR MH "Phototherapy+" OR MH "radiotherapy"+ OR MH “Physical Therapy+" OR MH "Rehabilitation+" OR MH "Drug Therapy+" OR MH "Therapy Animals+" OR MH "Pet Therapy+" OR MH "Manual Therapy+" OR MH "Manipulation, Orthopedic" OR MH "Manipulation, Osteopathic" OR MH "Manipulation, Chiropractic" OR TX "applied kinesiology" OR TX "soft tissue therap*" OR TX "musculoskeletal manipulation*" OR TX "chiropractic manipulation*" OR TX "orthopedic manipulation*" OR TX "osteopathic manipulation*" OR TX "spine manipulation*" OR TX "acoustic stimulation*" OR TX climatotherap* OR TX "complementary therap*" OR TX phototherap* OR TX	2-8-23

	radiotherap* OR TX "soft tissue therap*" OR TX "physical therap*" OR TX "electric stimulation*" OR TX "exercise therap*" OR TX "shockwave therap*" OR TX hydrotherap* OR TX "manual therap*" OR TX "myofunctional therap*" OR TX "art therap*" OR TX "music therap*" OR TX bibliotherap* OR TX "dance therap*" OR TX "occupational therap*" OR TX "recreation* therap*" OR TX rehabilitation OR TX telerehabilitation OR TX "therap* animal*" S4. (MH "Administrative Research+" OR MH "Clinical Research+" OR MH "Comparative Studies+" OR MH "Descriptive Research" OR MH "Ecological Research" OR MH "Education Research" OR MH "Evaluation Research+" OR MH "Exploratory Research" OR MH "Health Services Research+" OR MH "Historical Research" OR MH "Methodological Research" OR MH "Multicenter Studies" OR MH "Multimethod Studies" OR MH "Physiological Studies" OR MH "Pilot Studies" OR TX "observational stud*" OR TX "observation stud*"OR TX "clinic* stud*" OR TX "control trial" OR TX "controlled* trial" OR TI placebo OR TI "double blind*" OR TI "cross section*" OR TI stud* OR TI trial*) S5. S2 AND S3 S6. (S1 OR S5) AND S4 AND PT ("Journal Article") S7. S6 NOT (TI Cognition OR PT "Meta Analysis" OR PT "Meta Synthesis" OR PT "Review" OR PT "Systematic Review" OR PT "case study" OR TI "Meta analysis" OR TI "Systematic Review OR TI "case study" OR TI "case report*" OR TI "case series" OR TI "systematic review" OR TI "literature review" OR TI "scoping review" OR TI "narrative review") Limiters: Published Date: 19820101-19921231; English Language; Peer Reviewed [Search modes - Boolean/Phrase] S8. S6 NOT (TI Cognition OR PT "Meta Analysis" OR PT "Meta Synthesis" OR PT "Review" OR PT "Systematic Review" OR PT "case study" OR TI "Meta analysis" OR TI "Systematic Review OR TI "case study" OR TI "case report*" OR TI "case series" OR TI "systematic review" OR TI "literature review" OR TI "scoping review" OR TI "narrative review") Limiters: Published Date: 20090101-20091231; English Language; Peer Reviewed [Search modes - Boolean/Phrase] S9. S6 NOT (TI Cognition OR PT "Meta Analysis" OR PT "Meta Synthesis" OR PT "Review" OR PT "Systematic Review" OR PT "case study" OR TI "Meta analysis" OR TI "Systematic Review OR TI "case study" OR TI "case report*" OR TI "case series" OR TI "systematic review" OR TI "literature review" OR TI "scoping	

	review" OR TI "narrative review") Limiters - Published Date: 2021101-20211231; English Language; Peer Reviewed [Search modes - Boolean/Phrase] S10. S7 OR S8 OR S9	
Web of Science (Clarivate)	1. (TS=(parkinson*) NOT TS=(parkinsonims)) NOT TI=(cognition) 2. TS=(*therap* OR *stimulation* OR *Manipulation* OR "applied kinesiology" OR rehabilitation OR telerehabilitation) 3. TS=("clinic* stud*" OR "observation* stud*" OR "clinic* trial*" OR "control* trial*") OR TI=(placebo OR "double blind*" OR "cross section*" OR *stud* OR *trial*) 4. DT=(Article) NOT (DT=("Review Article") OR DT=("Book Chapter") OR TI=("case report*" OR "case stud*"OR "case series"[TI] OR "systematic review" OR "literature review" OR "scoping review" OR "narrative review" OR "meta analysis")) 5. ((#1 AND #2 AND #3 AND #4) AND LA=(English) AND PY=(1982-1992 OR 2009 OR 2021)	2-8-23

Appendix B: Data Extraction Instructions

In what year was the study published?

Possible values: 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 2009, 2021

Publication year was defined as the year in which an article was officially published, as it would appear on the article's citation. This year is not necessarily the same year that the article was accepted or became available online.

For example, the article below was accepted in December 2020, but officially published in 2021 and would therefore be included in our study:

> *Yun SJ, Lee HH, Lee WH, Lee SH, Oh B-M, Seo HG. Effect of robot-assisted gait training on gait automaticity in Parkinson disease: a prospective, open-label, single-arm, pilot study. Medicine 2021;100:5(e24348).*
>
> *Received: 10 September 2019 / Received in final form: 17 November 2020 / Accepted: 27 December 2020*

This information is often found on the first page of a paper.

Is the first and/or last author of this study affiliated with a U.S.-based institution?

Possible values: Yes, No

The criterion can be fulfilled even if the author is not exclusively affiliated with U.S.-based institutions. Author affiliations are usually found at the beginning of or end of a paper.

Do any of the authors have an affiliation with the National Institutes of Health (NIH)?

Possible values: Yes, No

The criterion can be fulfilled even if the author is not exclusively affiliated with the NIH. Further, unlike the U.S.-based criterion above, this applies to any author in the study, not just the first and last authors.

Does the paper have a funding statement?

Possible values: Yes, No

Authors may declare what funding they received or that they did not receive any funding, in numerous places within a manuscript. This includes at the bottom or side of the first page, in the acknowledgements section, and in explicitly labeled funding sections.

Was the study funded, even partially, by the NIH?

Possible values: Yes, No

A study was considered to be funded by the NIH if NIH grants (e.g., R01, F31, R33, etc.) were listed in the funding statement or if any of the authors are affiliated with the NIH or any of its sub-branches. The latter decision was made given that intramural researchers and their research are automatically supported by the NIH. If there is no funding statement nor are any authors affiliated with the NIH, select no.

What was the primary intervention of the study?

Possible values: Physiotherapeutic, Pharmacological

Physiotherapeutic interventions included physical therapy, occupational therapy, speech therapy, and exercise/movement. Pharmacological interventions included small molecule/drug therapy, functional neurosurgery, non-invasive neurostimulation, and other biological therapies like fecal matter transplants or gene therapy.

Was the race/ethnicity of study participants reported?

Possible values: Yes, No

If a statement such as "the majority of our participants were white" was made, we considered the authors to have reported race/ethnicity, even if not quantitatively. If authors said they collected information on race/ethnicity but did not report it, quantitatively or qualitatively, we did not consider the authors to have reported on race/ethnicity.

Was sex/gender referenced in any way in the title?

Possible values: Yes, No

Was sex/gender referenced in any way in the abstract?

Possible values: Yes - Claim, Yes - Statistics, Yes - Demographics, No

If authors make a claim about a sex/gender difference, select 'Yes - Claim'. If authors reference that they performed analyses comparing sexes/genders, but do not report results and/or make a claim, select 'Yes - Statistics'. If authors report the number of men or women in the study or say treatment groups were balanced by sex/gender, select 'Yes - Demographics'. Else, select 'No'.

How many men and women were enrolled in the study, disaggregated by disease status?

Possible values: numeric, Not Reported

Collect the following information from each study:

1. Number of participants, total; Number of men; Number of women
2. Number of participants, total; Number of men with PD; Number of women with PD
3. Number of participants with conditions other than PD, total; Number of men with conditions other than PD; Number of women with conditions other than PD
4. Number of healthy control participants, total; Number of healthy control, male participants; Number of healthy control, female participants

Were data analyzed by sex/gender?

Possible values: Yes - Main factor, Yes - Covariate, Yes - Separately, Yes - Not Described, No, Non-applicable

- *Yes - Main factor*: This applies if the study explicitly tests for sex/gender effects within the same model, reporting specific results (e.g., beta values) for sex/gender.
- *Yes - Covariate*: If the study includes sex/gender in its analysis model but does not specifically report outcomes by sex/gender, it is considered a covariate. This means the study accounted for sex/gender to isolate the effects of other variables under investigation.
- *Yes - Separately*: This designation is used when the study conducts separate analyses for men and women, without directly comparing the results between sexes/genders. DISS errors (Maney & Rich-Edwards, 2023) are possible here and noted where found.

- *Yes - Not Described*: This category is for studies that mention sex/gender analysis but lack detailed reporting on the findings or methodology related to sex/gender.
- *No*: This label is for studies that do not include any sex/gender-based analysis.
- *Not applicable*: This is used for studies that do not involve participants of more than one sex/gender, making sex/gender analysis irrelevant.

Did the authors make any claims about sex/gender differences in their study?

Possible values: Yes, Yes - No difference, No, Not applicable

If authors claim a difference between men and women, select 'Yes'. If authors report and discuss that their analyses yielded no significant differences between sexes/genders, select ' Yes - No difference'. It is not sufficient if non-significant statistical results are displayed if authors make no comment on them. Select 'No' if authors make no claim. Lastly, select 'Not applicable' if authors only researched one sex/gender.

Were the study results presented in a manner disaggregated by sex/gender that would allow for subsequent meta-analyses?

Possible values: Yes - Descriptive, Yes- Statistical Output, Yes - Raw, No, Not applicable

- *Yes - Descriptive*: This category is used when authors report findings summarized by sex/gender, such as average weight loss differences between men and women. These descriptive statistics can be presented in text, tabular format, or in figures.
- *Yes - Statistical Output*: This applies when authors present model outputs (e.g., beta values, correlation coefficients) by sex/gender. To be classified here, the report must clearly indicate or allow inference of how sex/gender was coded in the analysis.

- *Yes - Raw*: If the study provides raw data that enables an independent sex/gender analysis, such as listing each participant's sex/gender along with their individual data points, it falls under this category.
- *No*: This is selected when the study fails to provide sufficient detail for conducting a sex/gender-based meta-analysis. This might occur if sex/gender is factored into the model but the coding or interpretation is unclear, or if no sex/gender-based analysis was performed.
- *Not applicable*: Used for studies that do not include participants of men and women, rendering a sex/gender-based analysis irrelevant.

Did the authors provide a rationale for using a single sex/gender or acknowledge the limitations of this approach?

Possible values: Yes - Rationale, Yes - Limitation, No, Unknown, Nonapplicable

If authors provide explicit reasons for using a single sex (e.g., we already know this in men but not women), select ‘Yes - Rationale’. If authors acknowledge that the use of a single sex/gender is a limitation of their study, select ‘Yes - Limitation’. If authors do not provide a rationale or acknowledgement, select ‘No’. If authors did include men and women, select ‘Not applicable’. If authors did not specify inclusion by sex/gender, select ‘Unknown’.

Appendix C: PRISMA

PRISMA 2020 Checklists

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Title
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Intro
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses	Intro
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Methods
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Appendix B
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Appendix B
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in	Methods

		the process.	
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Methods
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Methods, Appendix C
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Appendix C
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Methods - NA
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Methods
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Methods
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Methods
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Methods

	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Methods - NA
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Methods - NA
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Methods - NA
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Methods - NA
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Methods - NA
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Results
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Methods
Study characteristics	17	Cite each included study and present its characteristics.	Data availability statement
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	NA
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Data availability statement
Results of synthesis	20a	For each synthesis, briefly summarize the characteristics and risk of bias among contributing studies.	Methods - NA

	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Methods - NA
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Methods - NA
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Methods - NA
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Methods - NA
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Methods - NA
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Discussion
	23b	Discuss any limitations of the evidence included in the review.	Discussion
	23c	Discuss any limitations of the review processes used.	Discussion
	23d	Discuss implications of the results for practice, policy, and future research.	Discussion
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Methods - None
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Methods - Not prepared
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	N/A
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Funding Sources

Competing interests	26	Declare any competing interests of review authors.	Competing interest statement
Availability of data, code, and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Data availability statement

PRISMA 2020 for Abstracts Checklist

Section and Topic	**Item #**	**Checklist item**	**Reported (Yes/No)**
TITLE			
Title	1	Identify the report as a systematic review.	Yes
Background			
Objectives	2	Provide an explicit statement of the main objective(s) or question(s) the review addresses.	Yes
METHODS			
Eligibility criteria	3	Specify the inclusion and exclusion criteria for the review	Yes
Information sources	4	Specify the information sources (e.g., databases, registers) used to identify studies and the date when each was last searched.	Yes
Risk of bias	5	Specify the methods used to assess risk of bias in the included studies	NA
Synthesis of results	6	Specify the methods used to present and synthesize results	Yes
RESULTS			
Included studies	7	Give the total number of included studies and participants and summarize relevant characteristics of studies.	Yes
Synthesis of	8	Present results for main outcomes, preferably	Yes

results		indicating the number of included studies and participants for each. If meta-analysis was done, report the summary estimate and confidence/credible interval. If comparing groups, indicate the direction of the effect (i.e., which group is favored).	
DISCUSSION			
Limitations of evidence	9	Provide a brief summary of the limitations of the evidence included in the review (e.g., study risk of bias, inconsistency and imprecision).	Yes
Interpretation	10	Provide a general interpretation of the results and important implications	Yes
OTHER			
Funding	11	Specify the primary source of funding for the review.	Yes - none
Registration	12	Provide the register name and registration number	Yes - not registered

Chapter 3: Monitoring Inclusion Practices by Sex in Scientific Publications using Natural Language Processing

Jeanne M. Powell[1], Abeed Sarker[1], and J. Lucas McKay[1, 2]

[1] Emory University, Atlanta, GA 30322, USA

[2] Georgia Institute of Technology, Atlanta, GA 30332, USA

Abstract

The historical underrepresentation of women and non-human female animals in research, coupled with policies promoting their inclusion, has given rise to a field focused on monitoring sex inclusion practices in research, traditionally through manual data extraction. Advancements in natural language processing (NLP), especially with large language models (LLMs) like GPT-4-Turbo, present opportunities for automating these analyses, potentially increasing efficiency and scalability. This study evaluates an automated pipeline using GPT-4-Turbo to extract sex inclusion data and participant counts from 465 Parkinson's disease (PD) clinical research articles spanning three distinct periods (1982–1992, 2009, 2021), surrounding sex-related NIH policy changes. We converted PDFs into plain text, prompted the GPT-4-Turbo model to extract sex-related data from these texts, and assessed model performance against manual annotations. Despite challenges with scanned PDFs due to parsing errors, the GPT-4-Turbo model achieved high accuracy (0.96 (95% CI: 0.94-0.97) for sex inclusion, 0.83 (95% CI: 0.79-0.86) for both male and female subject counts), demonstrating its capability to automate the extraction of sex-related data effectively. Our findings suggest LLMs like GPT-4-Turbo can streamline the monitoring of sex inclusion in scientific research, offering a scalable and resource-efficient method.

Monitoring Inclusion Practices by Sex in Scientific Publications using Natural Language Processing

In the preceding chapter, we conducted a systematic review to evaluate the categorical inclusion and proportion of women in Parkinson's disease (PD) clinical trials within three distinct periods (1982–1992, 2009, 2021) surrounding changes in NIH policies regarding the consideration of sex in research. Despite an observed overall increase in the categorical inclusion of both men and women, our analysis uncovered a consistent underrepresentation of women in PD studies. Our systematic review joins a number of studies that have aimed to monitor inclusion by sex/gender across research fields (e.g., Avery & Clark, 2016; Geller et al., 2018; Tosserams et al., 2018; Wheeler et al., 2020; Woitowich et al., 2020). To our knowledge, studies of this nature, to date, have all been performed manually, using one to multiple data extractors per publication. As such, these pursuits can be quite resource intensive to perform.

Recent advancements in natural language processing (NLP), specifically large language models (LLMs), hold promise to automate this task, making large scale monitoring of sex/gender inclusion practices tenable. LLMs are a powerful class of language models that are trained on massive amounts of unlabeled text data using self-supervised learning. This allows them to develop a knowledge base of language and generate human language in response to a written prompt (as reviewed by Fan et al., 2023). As such, these models hold exciting potential as a tool for scientific research because of their ability to transform unstructured text data into a structured output (Liu et al., 2023).

In November 2023, OpenAI released a preview of the GPT-4-Turbo model, which can process inputs that are up to 128,000 tokens in length (~96,000 words). This input context length is longer than that of its predecessor models (GPT-4: 8,192 tokens, ~6,144 words and

GPT-4-32K: 32,000 tokens, ~24,000 words), and operates at a fraction of the cost (https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo). As such, it is ideally suited for mining and extracting information from scientific publications. To that end, Foppiano and colleagues evaluated the abilities of a handful of LLMs, including the GPT-4-Turbo model, to extract structured information from scientific documents within the field of materials science (Foppiano et al., 2024). They found that the GPT-4-Turbo model performed beyond baseline when provided with a few task examples.

In this chapter, we aimed to develop an automated pipeline for extracting data on inclusion practices and subject counts from clinical research on PD delineated by sex. Our larger goal is to establish a pipeline that could automatically extract sex-related information across various fields, including research on non-human animal models. Therefore, in contrast to the terminology used in the previous chapter, we crafted the study prompt using the terms 'sex,' 'male,' and 'female' to be applicable to both human and non-human animal studies. The choice of "subjects" over "participants" was also deliberate to maintain relevance across species.

Despite our intentional use of these terms, the GPT-4-Turbo model appears to use 'gender', 'men', and 'women' interchangeably with 'sex', 'male', and 'female' for this task. While OpenAI has not fully disclosed its data sources, it is likely that most texts used sex and gender interchangeably, with only a minority distinguishing between them, leading to this behavior. Lastly, it is crucial to note that the pipeline is designed exclusively to identify information about subjects classified within the traditional binary framework of sex, male and female. It does not recognize data concerning individuals or subjects who do not conform to this binary classification.

Methods

Labeled Corpus

This corpus comprises 465 primary research articles published between 1982-1992, in 2009, or in 2021, focusing on pharmacological or physiotherapeutic clinical research related to PD. These articles were obtained either through direct download or by scanning physical copies accessed via interlibrary loan, resulting in both text- and image-based PDFs.

For each study, we conducted a manual annotation process to determine the sex composition of study participants. Specifically, we labeled studies based on whether they included male participants, female participants, male and female participants, or did not specify. It is important to note that the inclusion of male and female subjects was not always explicitly stated; in some instances, it was inferred from gendered descriptions of patient experiences within the text. Additionally, where available, we manually extracted the reported total number of male and female participants for each study.

Extracting Text from Study PDFs

First, we removed extraneous content from the PDFs. This included deleting cover and end pages provided by interlibrary loans and redacting text from additional articles that were scanned adjacent to target publications.

Next, we categorized the PDFs based on their format: regular (text-based) or scanned (image-based). For each PDF, its format was determined by extracting text from each page using the PyMuPDF library; if more than 75% of the pages contain fewer than 50 detectable words, the PDF was classified as scanned (image-based). Otherwise, it was considered regular (text-based).

For regular PDFs, we employed the VILA (Visual Layout Structures for Scientific Text Classification) model to extract the text (Shen et al., 2022). The VILA model enhances the

extraction process by recognizing and utilizing the visual layout structures of the document, facilitating a more accurate classification of the text into various blocks, such as headers, lists, authors, footers, tables, captions, figures, paragraphs, algorithms, titles, abstracts, footnotes, keywords, equations, bibliographies, and sections. We used Shen et al.'s (2022) End-to-end Paper Parsing tool, which deploys the pre-trained VILA model to extract text organized by section in the provided PDFs.

The extracted content was initially saved in a CSV file, which we used to reconstruct the study and save as a plain text file. The reconstruction of the document followed a specific order to maintain coherence and structure, which sometimes differed from the order it appeared on the actual PDF. We reconstructed the document starting with the title and abstract, followed by the body text (which includes sections, paragraphs, lists, equations, and algorithms), and finally, tables and figures (comprising captions, tables, and figures). Certain elements like headers, authors, footers, footnotes, keywords, and bibliographies were disregarded in the final text file. Finally, each section of the reconstructed document—namely the title, abstract, body text, and tables and figures—was explicitly labeled to ensure clarity and ease of navigation for LLMs, enabling a more effective and targeted analysis of the text.

Two papers identified as being stored in a 'regular' PDF could not be parsed using the VILA model and required processing using the protocol designated for scanned documents.

For scanned PDFs, we converted images of text into strings using the PyTesseract library, which parses text from images using Optical Character Recognition (OCR). For each scanned PDF and the two regular PDFs that could not be parsed by the VILA model, we extracted text from each page using PyTesseract. The text file was constructed in page order, with no further refinement.

GPT-4-Turbo-Preview Model

Our experiments were conducted using the OpenAI API and the "GPT-4-Turbo-Preview" model (equivalent to “GPT-4-0125-Preview” model). As of February 2024, this model can handle a context window of up to 128,000 tokens and has been trained on information through April 2023. It costs $0.01 / 1K input tokens and $0.03 / 1K output tokens.

For the prompt development experiments, we configured a Chat Completions request, setting the temperature to 0.2 and top_p to 0.1, with a response count (n) of 1, and without specifying a stop condition. These settings, especially the conservative values for temperature and top_p, were chosen to enhance the predictability and consistency of the model's outputs.

For the temperature experiments, we experimented with temperature values of 0, 0.2, and 0.4 while keeping the top_p, other parameters, and prompt constant.

For each paper within our corpus, we embedded the text of the respective paper into the prompt at the designated placeholder {{paper}}. We prompted the model to extract the categorical inclusion by sex (male subjects/female subjects/both/not reported) and the total number of male and female subjects at baseline. Responses from each request were saved and then parsed into individual variables for analysis.

Prompt and Temperature Selection

We experimented with different prompts and temperatures using a convenience sample of 52 papers selected from our corpus, selected alphabetically by the first authors’ last names. Before deployment on all 52 papers, we tested each prompt on a few papers using the OpenAI Playground and ChatGPT4. We measured model performance using the accuracy metric for each variable.

Each prompt was sequentially tested and evaluated, beginning with Prompt 1. Adjustments were made to each subsequent prompt based on inaccuracies identified from its predecessor. Discrepancies indicating incorrect ground truth labels—stemming from manual screening errors, failure to parse information from the PDF, or the absence of data from a study's supplemental materials that was not provided to the model—necessitated adjustments to the ground-truth labels to align with data actually received by the model. Performance metrics were calculated using these revised labels. For every response deemed inaccurate, a manual investigation was conducted to ascertain the cause of erroneous information extraction by the model. This led to the refinement of subsequent prompts to incorporate more precise and explicit directives.

From the four prompts, we selected the prompt with the highest average accuracy across the three target variables (sex inclusion, male subject count, and female subject count) and experimented with temperature values.

Prompt 1

You are an AI assistant that helps people find information in scientific publications.

Read the paper carefully. The information needed will likely be in either the Body Text or Tables and Figure sections:

{{paper}}

Please analyze the text to determine the inclusion of subjects by sex in the study. Specifically, identify whether the study included only male subjects, only female subjects, both male and female subjects, or did not report inclusion by sex.

In cases where the study did report inclusion by sex, please extract the exact number of male and female subjects participating in the study. Use 'N' to denote the total number of subjects.

Format your response as follows to ensure the information is clearly structured and easily extractable:

{
"Sexes": "male/female/both/not reported",
"N_male": "number or NA",
"N_female": "number or NA"
}

If the information regarding the number of male or female subjects is not available or not reported in the study, please provide "NA" for the respective field.

Prompt 2

You are an AI assistant that helps people find information in scientific publications.

Read the paper carefully. The information needed will likely be in either the Body Text or Tables and Figure sections:

{{paper}}

Please analyze the text to determine the inclusion of subjects by sex in the study. Specifically, identify whether the study included only male subjects, only female subjects, both male and female subjects, or did not report inclusion by sex.

In cases where the study did report inclusion by sex, please extract the exact number of male and female subjects participating in the study. Use 'N' to denote the total number of subjects. To do this, you may need to calculate the numbers yourself using information in the tables, including taking the total number of participants and multiplying it by the proportion of women reported in the table.

Format your response as follows to ensure the information is clearly structured and easily extractable:

```
{
  "Sexes": "male/female/both/not reported",
  "N_male": "number or NA",
  "N_female": "number or NA"
}
```

If the information regarding the number of male or female subjects is not available or not reported in the study, please provide "NA" for the respective field.

Prompt 3

You are an AI assistant that helps people find information in scientific publications.

Read the paper carefully. The information needed will likely be in either the Body Text or Tables and Figure sections:

{{paper}}

Please analyze the text to determine the inclusion of subjects by sex in the study. Specifically, identify whether the study included only male subjects, only female subjects, both male and female subjects, or did not report inclusion by sex.

In cases where the study did report inclusion by sex, please extract the exact number of male and female subjects participating in the study. Use 'N' to denote the total number of subjects. To do this, you may need to calculate the numbers yourself using information provided, including taking the total number of participants and multiplying it by the proportion of women reported in the text. Also, you may need to infer missing information based on context, like when the total number of subjects and total number of male subjects are provided with no mention of female subjects. You will need to calculate the number of female subjects based on the number of total and male subjects reported.

Further, you may need to combine participant numbers across treatment groups within a single study (like patients and healthy controls) and/or combine counts from multiple studies reported in the same paper. Pay attention to whether participants are distinct or overlap across groups and/or studies. Next, if the study is longitudinal, calculate the number of subjects at baseline. Finally, only count subjects analyzed in the current study. For example, a study may be a long term follow-up study that first reports the original cohort but then only analyzes data from a subset of that cohort.

Double check your calculations and inferences before reporting the final numbers.

Format your response as follows to ensure the information is clearly structured and easily extractable:

```
{
  "Sexes": "male/female/both/not reported",
  "N_male": "number or NA",
  "N_female": "number or NA"
}
```

If the information regarding the number of male or female subjects is not available or not reported in the study, please provide "NA" for the respective field.

Prompt 4

You are an AI assistant that helps people find information in scientific publications.

Read the paper carefully. The information needed will likely be in either the methods, results, and/or tables:

{{paper}}

Analyze the text to determine the inclusion of subjects by sex in the study. Specifically, identify whether the study included only male subjects, only female subjects, both male and female subjects, or did not report inclusion by sex.

Next, extract the total number of male and female subjects participating in the study at baseline. Use 'N' to denote the total number of subjects. To do this, you may need to extract and/or infer information from multiple sources, combine that information, and perform calculations. Further, only count subjects analyzed in the current study. For example, a study may be a long term follow-up study that first reports the original cohort but then only analyzes data from a subset of that cohort.

Here are a few examples for this task that require inference, calculations, and combining information across groups/experiments. Note, these examples are displayed using markup for clarity. You will not encounter a markup table in the paper. The tables will be parsed into plain text.

Example 1:

Text:

Table 1. Demographics Information

| | Parkinson's (n=30) | Controls (n = 25) |

| ---- | ---- | ---- |

| Sex, female % | 50% | 20% |

Response:

{

"Sexes": "both"

"N_male": "35"

"N_female": "20"

}

Example 2:

Text:

Table 1. Experiment 1 Demographics Information

| | PD Participants, n=10 | Controls, n = 10 |

| ---- | ---- | ---- |

| Male, n (%) | 6 (60%) | 8 (80%) |

Table 2. Experiment 2 Demographics Information

| | placebo, treatment n = 20 | treatment, placebo n = 15 |

| ---- | ---- | ---- |

| Male, n (%) | 10 (50%) | 10 (67%) |

Response:

{

"Sexes": "both"

"N_male": "34"

"N_female": "21"

}

Example 3:

Text:

Table 1. Clinical and demographic features over time

| | Ctrl-0 yr | PD-0 yr | Ctrl-1 yr | PD-1 yr | Ctrl-2 yr | PD-2 yr |

| ---- | ---- | ---- | ---- | ---- | ---- | ---- |

| N | 50 | 50 | 45 | 40 | 20 | 10 |

| Male:female ratio | 50:50 | 56:44 | 58:42 | 67:33 | 55:45 | 50:50 |

Response:

{

"Sexes": "both"

"N_male": "53"

"N_female": "47"

}

Double check your calculations and inferences before reporting the final numbers.

Format your response as follows to ensure the information is clearly structured and easily extractable:

{

"Sexes": "male/female/both/not reported",

"N_male": "number or NA",

"N_female": "number or NA"

}

If the information regarding the number of male or female subjects is not available or not reported in the study, please provide "NA" for the respective field.

Analysis

We assessed model performance via accuracy for each variable: sex inclusion, male participant count, and female participant count. Predictions were compared against two annotated datasets. The first dataset underwent correction solely for identified human errors, as

evidenced by mismatches between initial annotations and model outputs. The second dataset received adjustments for both human errors and discrepancies in information availability to the model, addressing instances where the manual reviewer had access to data not available to the model, whether it was not parsed correctly or located in supplementary files.

In our categorical analyses, we employed Fisher's Exact tests for group comparisons. We did post hoc analyses using pairwise comparisons and the Bonferroni correction. For our numerical analyses comparing manual annotations to predicted subject counts, we employed paired t-tests and ANOVAs.

Results

Prompt and Temperature Selection

We conducted experiments with four different prompts to enhance the performance of the GPT-4-Turbo model in extracting sex inclusion data and counts of male and female subjects from scientific articles (**Table 1**). Prompt 4 had the highest average accuracy across variables at 0.91 (95% CI: 0.79-0.96), but it was not significantly better than the other prompts based on the 95% confidence intervals. Prompt four achieved an accuracy of 0.98 (95% CI: 0.90-1.00) in identifying sex inclusion, 0.85 (95% CI: 0.72-0.92) in determining male subject counts, and 0.90 (95% CI: 0.79-0.96) for female subject counts.

Subsequently, we tested three temperature settings using Prompt 4 to further refine the model's accuracy. The highest results were obtained at a temperature setting of 0.2 (**Table 2**). Despite these improvements, the differences were not significant based on 95% confidence intervals. We used Prompt 4 with a temperature setting of 0.2 for full extraction from the corpus.

Model Performance

Structured Output

The model always returned the target variables and their predicted values in JSON format. In 7.53% of instances, it also generated unstructured text before the JSON entry, yet the relevant information remained extractable with minimal coding.

Annotations Corrected for Human Error and Adjusted for Missing Information

Our examination of the discrepancies between manual annotations and predicted values identified 15 manual errors, 23 instances of incorrect parsing of target information from PDFs, and 8 cases where the target information was omitted from the model input because it resided in supplementary study materials (n = 7) or a previous report (n = 1). Following corrections and adjustments to annotations, the model achieved an accuracy of 0.99 (95% CI: 0.97-0.99) for sex inclusion data extraction, 0.87 (95% CI: 0.84-0.9) for male subject count, and 0.88 (95% CI: 0.84-0.9) for female subject count.

Annotations Corrected for Human Error

When only correcting for human error, the model achieved an accuracy of 0.96 (95% CI: 0.94-0.97) for sex inclusion data extraction, 0.83 (95% CI: 0.79-0.86) for male subject count, and 0.83 (95% CI: 0.79-0.86) for female subject count. This model performance reflects expected performance on a new dataset, as identifying and rectifying parsing and omission errors would necessitate manual annotation and comparison of discrepancies. This approach contradicts the tool's objective of automating the process and reducing manual labor.

Scanned Versus Regular PDFs

We categorized each main paper file as either a scanned PDF (image-based) or a regular PDF (text-based). For scanned PDFs, text extraction was performed using OCR, while a

visual-layout aware extraction model was applied to regular PDFs. In our dataset, we identified 64 scanned PDFs and 401 regular PDFs (**Table 3**). All scanned PDFs, except for two, originated from studies published before 1992. The exceptions were two PDFs from studies in 2021, which, despite being text-based, encountered issues with the VILA model processing and were consequently analyzed using OCR, akin to the image-based PDFs.

PDF Format and Model Accuracy. We conducted Fisher's Exact Tests to determine the association between PDF format and model accuracy for each study variable (**Figure 1**). The results of the Fisher's Exact Tests for all study variables indicate a significant association between PDF format and model accuracy (sex inclusion: $p < .001$, OR: 12.86; male subject count: $p = .007$, OR: 2.35, female subject count: $p = .004$, OR: 2.58). These results indicate that the model is almost 13 times more likely to accurately extract sex inclusion information for regular compared to scanned PDFs and over two times more likely to accurately extract male and female subject counts from regular articles compared to scanned articles.

PDF Format and Information Omission Errors. Our objective was to ascertain how the PDF format—regular versus scanned—impacts the occurrence of information availability errors, which encompass both parsing errors and omissions (**Figure 2**). We provided the model only with text within the main paper file, excluding any supplementary files or previous reports where the target information might reside. We considered cases where the information was found in an external document to be an omission error. Through Fisher's Exact Tests, we identified a significantly higher incidence of parsing errors in scanned PDFs when compared to regular PDFs ($p < .001$, OR = 23.49), with parsing errors being over 23 times more likely to occur for scanned PDFs compared to regular PDFs. Conversely, the frequency of errors related to information only

being present in external documents did not significantly differ between the two PDF formats (p = .61, Haldane corrected OR = 0.36).

PDF Format and Model Errors. We performed Fisher's Exact Tests to assess whether the occurrence rates of model errors varied between regular and scanned PDF documents (**Figure 3**). A model error is defined as an instance when the target information is present in the text file but is not accurately extracted. There was no significant difference in the occurrence of model errors between regular and scanned PDFs (sex inclusion data: p = .19, OR = 0.31; male subject counts: p = .11, OR = 2.38; female subject counts: p = .31, OR = 1.80).

Comparison of Manual Annotations and Model Predictions

Sex Inclusion. We performed Fisher's Exact Tests to assess whether model predictions of sex inclusion differed significantly from ground-truth annotations. We found that the overall counts in each sex inclusion category did not differ between the manually- and model-derived datasets (p = .59, **Table 4**).

Subject Counts. To assess the accuracy of predicted subject counts against actual counts, we conducted paired sample t-tests. For these analyses, we excluded studies lacking numerical values for both predicted and actual subject counts. For the analysis comparing male subject counts, this exclusion process removed five studies where male subject counts were reported but not predicted by the model, and 20 studies where the model made predictions in the absence of reported subject counts. For the analysis comparing female subject counts, this exclusion process removed four studies where female subject counts were reported but not predicted by the model, and 23 studies where the model made predictions in the absence of reported counts.

Using paired t-tests, we found no significant difference between predicted and actual male subject counts ($t(405) = 0.87$, $p = .39$, $d = 0.004$, **Figure 4**) or predicted and actual female subject counts ($t(403) = 1.12$, $p = .26$, $d = 0.007$, **Figure 5**).

Comparison of Conclusions Derived from Manual Annotations and Model Predictions

Here, we aim to compare answers to sex-based research questions generated from manual annotations and model predictions.

Sex Inclusion Practices Over Time. In the previous chapter, we found a statistically significant difference ($p = .049$) in the reporting of sex inclusion over time. Post-hoc analysis revealed a difference in the reporting of sex inclusion between studies published during the earliest (1982-1992) and the latest (2021) time frames ($p_{adj} = .0497$). Specifically, studies from 1982 to 1992 were significantly less likely to report the inclusion of male and female participants and instead not specify sex-based inclusion compared to those published in 2021 ($p_{adj} = .0268$, OR = 0.0928). No significant differences were observed in sex inclusion reporting between the periods of 1982-1992 and 2009 ($p_{adj} = 1$) or between 2009 and 2021 ($p_{adj} = 1$).

From model predictions, we found a statistically significant difference ($p < 0.01$) in the reporting of sex/gender inclusion over time. Post-hoc analysis revealed a significant difference in the reporting of sex/gender inclusion between studies published during the earliest (1982-1992) and the latest (2021) time frames ($p_{adj} < .001$). Specifically, studies from 1982 to 1992 were significantly less likely to report the inclusion of both sexes and instead not specify sex/gender-based inclusion compared to those published in 2021 ($p_{adj} < .001$, OR = 0.1203). No significant differences were observed in sex inclusion reporting between the periods of 1982-1992 and 2009 ($p_{adj} = .167$) or between 2009 and 2021 ($p_{adj} = 1$).

Collectively, the data from manual annotations and model predictions come to the same conclusion: there has been a notable shift in sex inclusion reporting between studies on PD clinical research published in the period of 1982-1992 and those from 2021. This change is characterized by an increased tendency to report the inclusion of male and female subjects, moving away from the practice of not specifying the sex of study participants.

Proportion of Female Subjects Over Time. This analysis represents a slight departure from the approach detailed in the previous chapter. Previously, we focused on comparing the proportion of women with PD among all individuals with PD who participated in the studies. In contrast, for this analysis, we directed the model to extract the total counts of male and female subjects, which could include both healthy control subjects as well as individuals with other disorders. Consequently, we re-computed and compared the annotated and predicted logit-transformed proportions of total female subjects in each study over time (**Figure 6**).

To determine the proportion of female subjects based on manual annotations, we focused on the 200 papers from our dataset that specified the number of male and female subjects included in the study and were authored by researchers affiliated with US institutions. To assess whether the proportion of total female subjects in these studies has increased over time, we conducted a logit-transformed ANOVA. Our analysis revealed that the proportion of female subjects, when logit-transformed, showed non-significant differences across different time periods ($F(2, 197) = 0.495$, $p = .61$).

To determine the proportion of female subjects based on model predictions, we focused on the 182 papers from our dataset that were authored by researchers affiliated with US institutions and from which the model predicted male and female subject counts. To assess whether the proportion of total female subjects in these studies has increased over time, we

conducted a logit-transformed ANOVA. Our analysis revealed that the proportion of female subjects, when logit-transformed, showed non-significant differences across different time periods ($F(2, 178) = 0.336$, $p = .715$).

Taken together, both manually annotated and model predicted data point to the same conclusion, the proportion of female subjects in PD clinical research studies has not significantly changed over time.

Cost

The papers in our corpus were on average 7,515.16 (SD = 3,595.06) tokens long, ranging from 1,031.25 to 36,184.25 tokens. Our final prompt was 677 tokens long. On average, model responses were 20.88 (SD = 23.34) tokens long, ranging from 14.25 to 193 tokens. In terms of cost, this translates to an average processing cost of $0.08 (SD = $0.04) per paper, ranging from $0.02 to $0.37.

Leveraging a LLM as an adjunct reviewer would significantly reduce the person-hours required for this research, approximately halving the time compared to conducting a traditional systematic review with dual reviewers. Extracting data on sex inclusion and male and female subject counts from 465 papers took each person around 30 hours. Additionally, we allocated time to reconcile discrepancies in our data—a process comparable in duration to resolving discrepancies between ground truth labels and model predictions, including prompt refinement. Consequently, this methodology resulted in a saving of 30 person-hours (equating to $450 at $15/hour). With testing and deployment costs totaling around $100, this approach netted a saving of $300, alongside substantial time savings. This efficiency gain is expected to increase with the scaling of the number of papers analyzed.

Discussion

In this investigation, we utilized the GPT4-Turbo-Preview model to systematically extract data on sex inclusion, specifically participant sex reporting and the counts of male and female subjects, from scientific publications on clinical PD research dated between 1982-1992, 2009, and 2021. The model demonstrated high accuracy in retrieving the specified information across the majority of the analyzed articles. Comparisons between model predictions and manual annotations for these variables revealed no significant discrepancies. Consequently, insights regarding sex inclusion practices derived from model analyses aligned with those obtained through manual data extraction.

Importance of Structured Data Extraction

One of the challenges of working with text data for research is that it is unstructured, and insights need to be extracted in a structured format to allow for quantitative analyses (Foppiano et al., 2024). This study has demonstrated the GPT4-Turbo-Preview model's capability to not only process vast quantities of information but also to generate structured, machine-readable outputs that facilitate further analysis.

The Role of Few-shot Prompting to Improve Model Performance

Consistent with other reports using GPT models for data extraction (e.g., Brown et al., 2020; Gao et al., 2021; Srivastava et al., 2022), our investigation into the model's performance underscored the critical role of few-shot prompting in enhancing the model's ability to navigate complex data sets. By providing the model with explicit instructions and examples, including demographic table structures and calculations, we improved its capacity to extract accurate counts of male and female subjects. This was particularly evident in the handling of complex scenarios requiring calculations and inferences across experiment groups and arms.

Challenges in Data Extraction: Regular vs. Scanned PDFs

The limitations of our image-to-text extraction method became apparent in our analysis of scanned PDFs. Errors in text conversion led to incomplete or incomprehensible data, affecting the model's accuracy. We exclusively used the PyTesseract library to extract text from image-based PDFs, and did not experiment with other tools. Roopesh and colleagues (2021) experimented with multiple OCR libraries to extract text from images on the web. They found that PyTesseract outperformed other libraries, achieving higher precision, speed, and accuracy. Importantly, they found PyTesseract accurately extracted only 83.45% of text from images. Further research is needed to improve OCR tools for this task. Practically, our results indicate that our pipeline may be less effective at extracting accurate information from historical texts within a field given they are more likely to be stored as a scanned PDF.

Despite these challenges, the model occasionally demonstrated an ability to reconcile conflicting information, highlighting its potential for sophisticated inference. For example, in one study (Allen et al., 1989), the authors stated that there were 18 study participants, and their individual demographic data were available in Table 1. Due to parsing errors, only 15 of the 18 participants were presented in Table 1. This study was part of our convenience sample of 52 papers used to experiment with different prompts and model temperatures. Therefore, we observed model behavior on extracting information from this document numerous times. The model always predicted male and female subject counts that summed to 18, and the individual male and female subject counts were never less than what was included in Table 1.

Utility of Automated Pipelines For Sex-Related Data Extraction

Our analysis of discrepancies between model predictions and manual annotations highlighted instances of human error, emphasizing the utility of automated pipelines as adjunct

tools for systematic reviews. For instance, a notable discrepancy was observed in the male and female subject count predictions versus the labels from Patel et al. (2021). In this paper, Table 1 outlines patient characteristics, including sex. Initially, our labels were derived solely from this table. However, further examination revealed the inclusion of a control group, with demographics detailed in the narrative. The model adeptly aggregated counts across groups—a detail manual annotators overlooked. Reading entire papers is impractically time-consuming, leading annotators to search for information in expected locations, such as demographic tables. This approach, while time-saving, is prone to oversights. The model's ability to analyze the complete text allows it to integrate scattered demographic data. Automation of the data extraction process could significantly lessen the workload and duration needed, potentially acting as a complementary reviewer and reducing the person-hours required by half.

Cost analysis revealed the pipeline to be highly cost-effective, with the entire experiment conducted at a fraction of the cost of manual labor. Conducted on a standard iMac with a 3 GHz 6-Core Intel Core i5 processor, this study further demonstrates the accessibility of the OpenAI platform for leveraging advanced computational models without the necessity for high-performance computing infrastructure.

Conclusions and Future Directions

In conclusion, this study illustrates the potential of the GPT-4-Turbo model to automate the extraction of structured data from scientific literature, offering a scalable, efficient, and cost-effective solution for analyzing sex inclusion in PD research. Future work should focus on refining the model's accuracy, particularly in handling complex data presentations and optimizing OCR technology, to enhance the reliability and applicability of automated data extraction in

scientific research. Further, future work should test the performance of this pipeline on a corpus of papers from a different field to test whether it does indeed generalize.

Data Availability Statement

All data, code, and statistical analyses is available and can be found at https://github.com/jeannempowell/ExtractSexPDStudies.

Acknowledgements

We would like to thank Yuting Guo for providing sample code used to query the OpenAI API.

Tables and Figures

Table 1

Comparative Accuracy of Model Responses Across Different Prompts for Extracting Sex-Related Data from Scientific Literature

	Prompt 1	Prompt 2	Prompt 3	Prompt 4
Sex Inclusion	**1 (.93-1)**	**1 (.93-1)**	.98 (.9-1)	.98 (.9-1)
Male subject count	.79 (.66-.88)	.83 (.7-.91)	**.88 (.77-.95)**	.85 (.72-.92)
Female subject count	.77 (.64-.86)	.83 (.7-.91)	.85 (.72-.92)	**.9 (.79-.96)**
Average across variables	.85 (.72-.92)	.89 (.77-.95)	.9 (.79-.96)	**.91 (.79-.96)**

Note. This table presents the accuracy of the model in identifying and extracting sex-related variables (sex inclusion, male subject count, female subject count) from scientific literature, benchmarked against manual annotations. Confidence intervals at 95% are provided in parentheses. The highest accuracy value for each variable is bolded.

Table 2

Impact of Sampling Temperature on Model Accuracy for Sex-Related Data Extraction Using Prompt 4

	Temperature = 0	Temperature = 0.2	Temperature = 0.4
Sex Inclusion	**1 (.93-1)**	.98 (.9-1)	**1 (.93-1)**
Male subject count	.83 (.7-.91)	**.85 (.72-.92)**	.83 (.7-.91)
Female subject count	.85 (.72-.92)	**.9 (.79-.96)**	.81 (.68-.89)
Average across variables	.89 (.77-.95)	**.91 (.79-.96)**	.88 (.77-.95)

Note. This table delineates the accuracy of extracting sex-specific information (sex inclusion, male subject count, female subject count) from scientific literature using Prompt 4 at different sampling temperatures (0, 0.2, and 0.4). Confidence intervals at 95% are provided in parentheses. The highest accuracy value for each variable is bolded.

Table 3

PDF Formats in our Corpus, Disaggregated by Publication Year

	1982-1992	2009	2021
Regular	36	55	310
Scanned	62	0	2

Note. This table categorizes the PDF documents in our dataset based on their format—regular or scanned—and disaggregates them by publication year. Scanned PDFs predominantly come from earlier studies, mainly between 1982 and 1992. Notably, two scanned PDFs from 2021, despite being saved in a text-based format, were processed using OCR due to compatibility issues with the VILA model, and were therefore labeled as "scanned" because text was extracted using the pipeline used on image-based PDFs.

Table 4

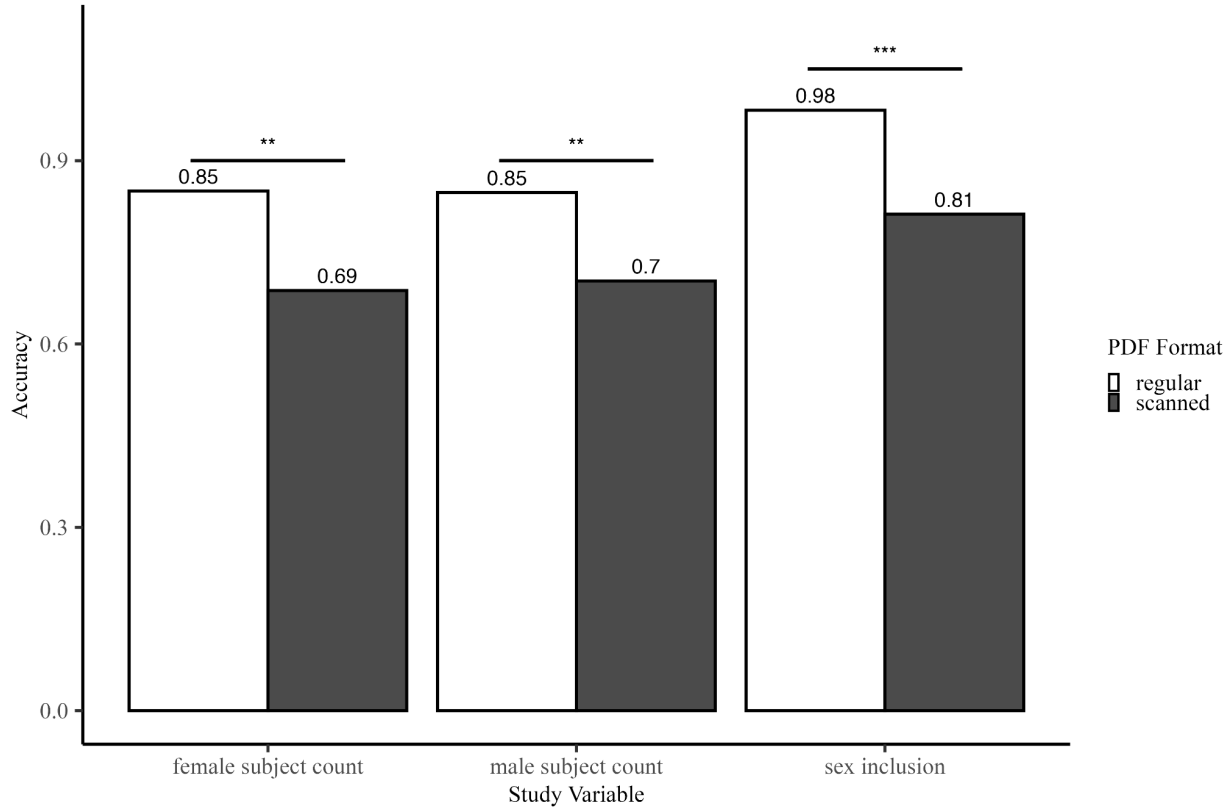
Confusion Matrix for Sex Inclusion Labels and Predictions

		Predicted				
		Both	Female	Male	Not Reported	Total
Actual	Both	407	0	0	15	422
	Female	0	1	0	0	1
	Male	0	0	12	0	12
	Not Reported	4	0	0	26	30
	Total	411	1	12	41	

Note. This table presents a detailed breakdown of frequencies for sex inclusion categories (both, female, male, not reported) as predicted by the model compared to the manually annotated ground-truth labels. Each row represents the actual category from ground-truth annotations, while each column represents the model's predictions. The diagonal cells (407 for both, 1 for female, 12 for male, and 26 for not reported) indicate accurate model predictions. Off-diagonal cells represent discrepancies between predictions and actual annotations. Fisher's Exact Test suggests no significant difference in the distribution of sex inclusion categories between manually and model-derived datasets ($p = .59$), indicating overall model accuracy in predicting sex inclusion categories.

Figure 1

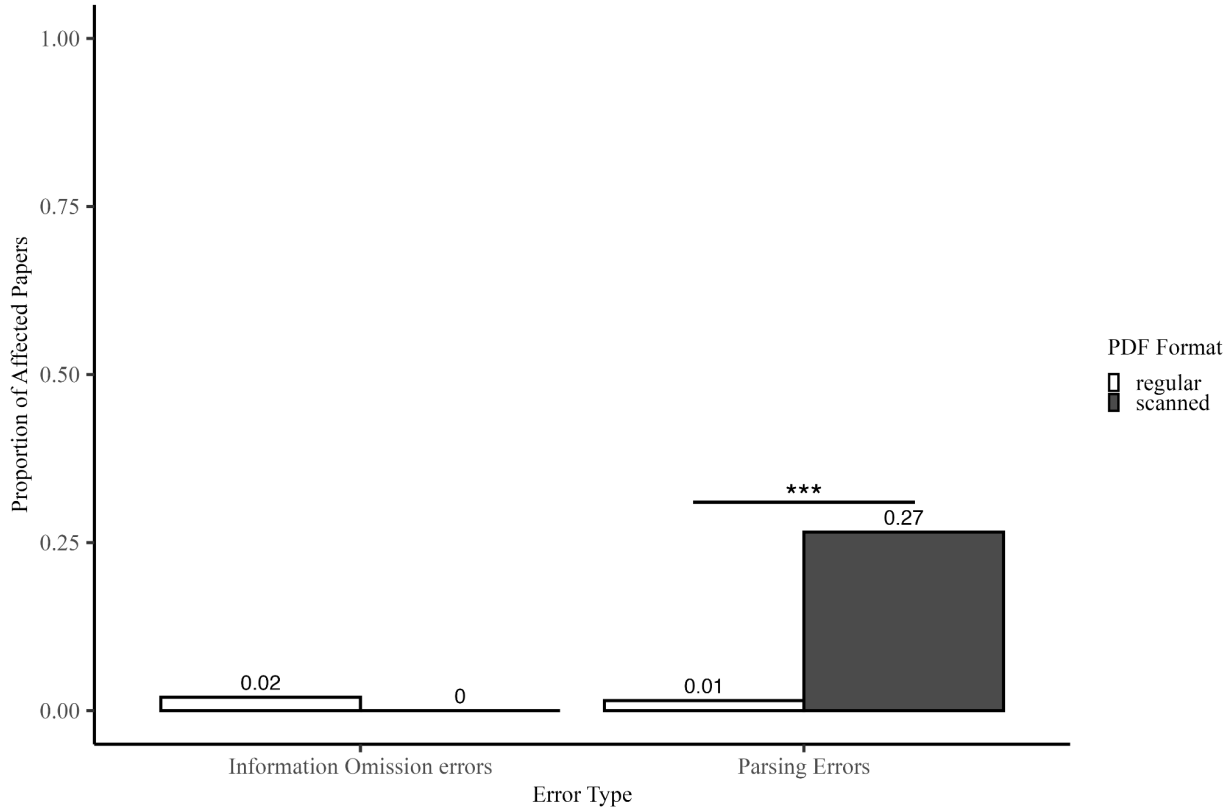
Accuracy of Model Outputs Stratified by PDF Type and Variable



Note. This figure illustrates the accuracy of the GPT-4-Turbo Model in extracting sex-related information, stratified by PDF format (scanned vs. regular) and study variable. Fisher's Exact Tests reveal a significant association between PDF format and model accuracy across variables: sex inclusion ($p < .001$, OR: 12.86), male subject count ($p = .007$, OR = 2.35), and female subject count ($p = .004$, OR = 2.58). These results demonstrate that the model's accuracy in automatic data extraction is higher with regular PDFs compared to scanned PDFs.

Figure 2

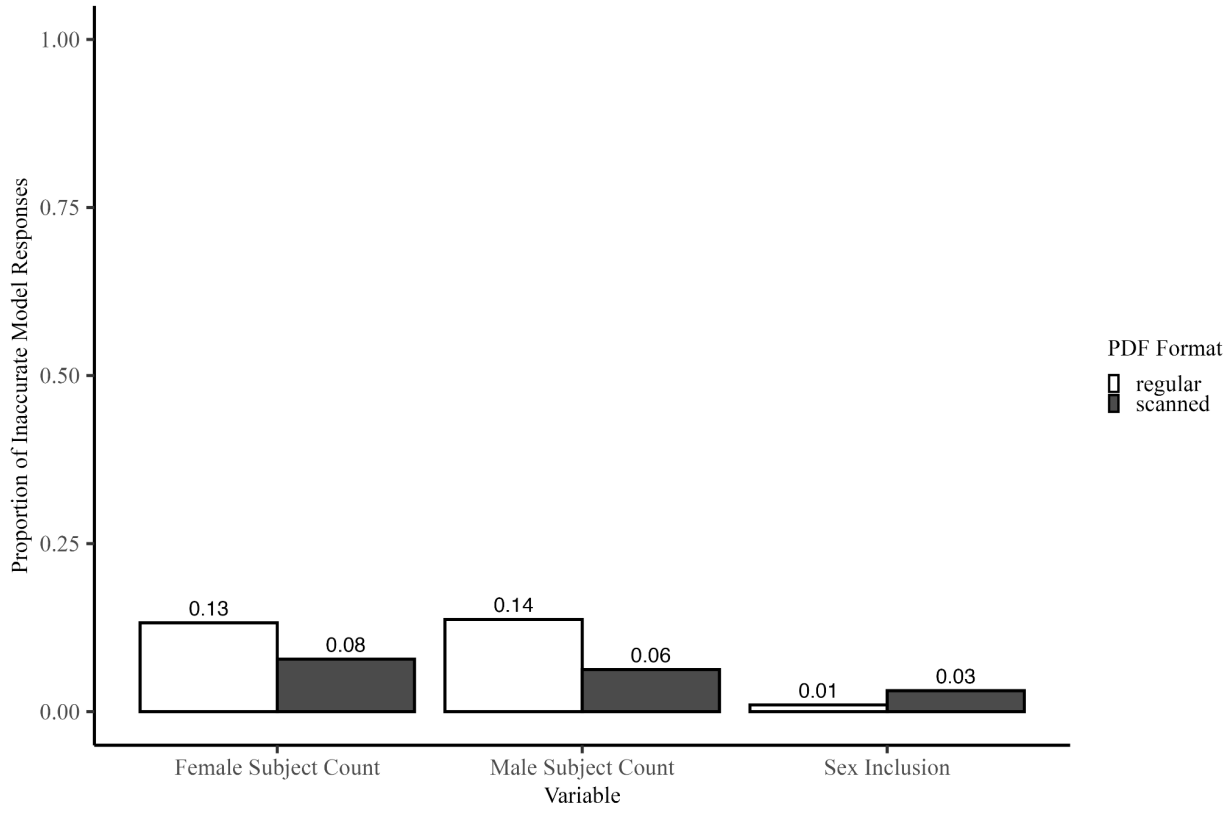
Comparison of Information Availability Errors Between Regular and Scanned PDFs



Note. This figure illustrates the rate of occurrence of different information availability errors (parsing errors and omissions) between scanned and regular PDFs. Fisher's Exact Tests revealed a significantly higher occurrence of parsing errors in scanned PDFs compared to regular PDFs ($p < .001$, Odds Ratio (OR) = 23.49), suggesting a substantial impact of PDF format on the accuracy of text extraction. Advancements in optical character recognition may help improve scanned PDF parsing and subsequently enhance model accuracy. The rate of errors related to information found in external documents was not significantly different between the two formats ($p = .61$, Haldane corrected OR = 0.36).

Figure 3

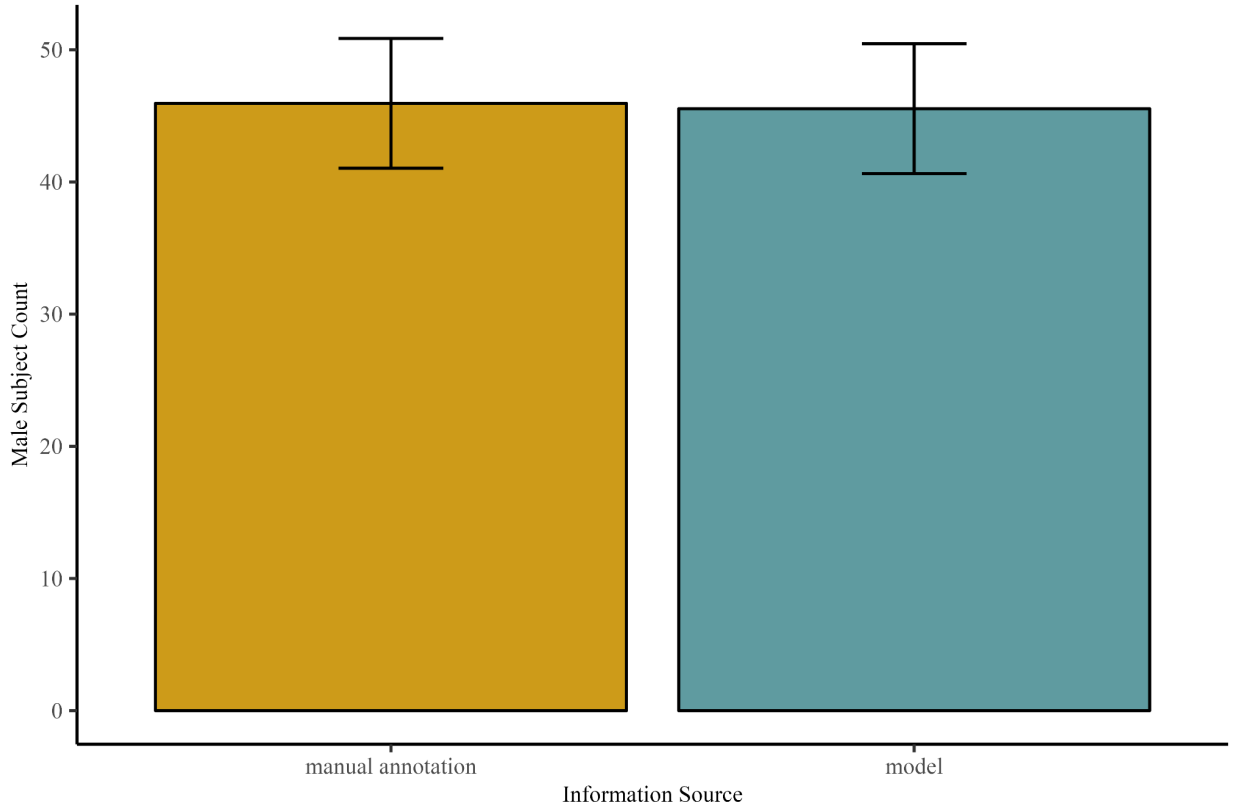
Proportion of Inaccurate Model Responses When Target Information was Provided, Stratified by Study Variable and PDF Format



Note. This figure summarizes the outcomes of Fisher's Exact Tests designed to evaluate the variation in error rates made by the model in extracting information from regular versus scanned PDF documents. A model error is defined as an instance when the target information is present in the text file but is not accurately extracted. There was no significant difference in the occurrence of model errors between regular and scanned PDFs (sex inclusion data: $p = .19$, OR = 0.31; male subject counts: $p = .11$, OR = 2.38; female subject counts: $p = .31$, OR = 1.8).

Figure 4

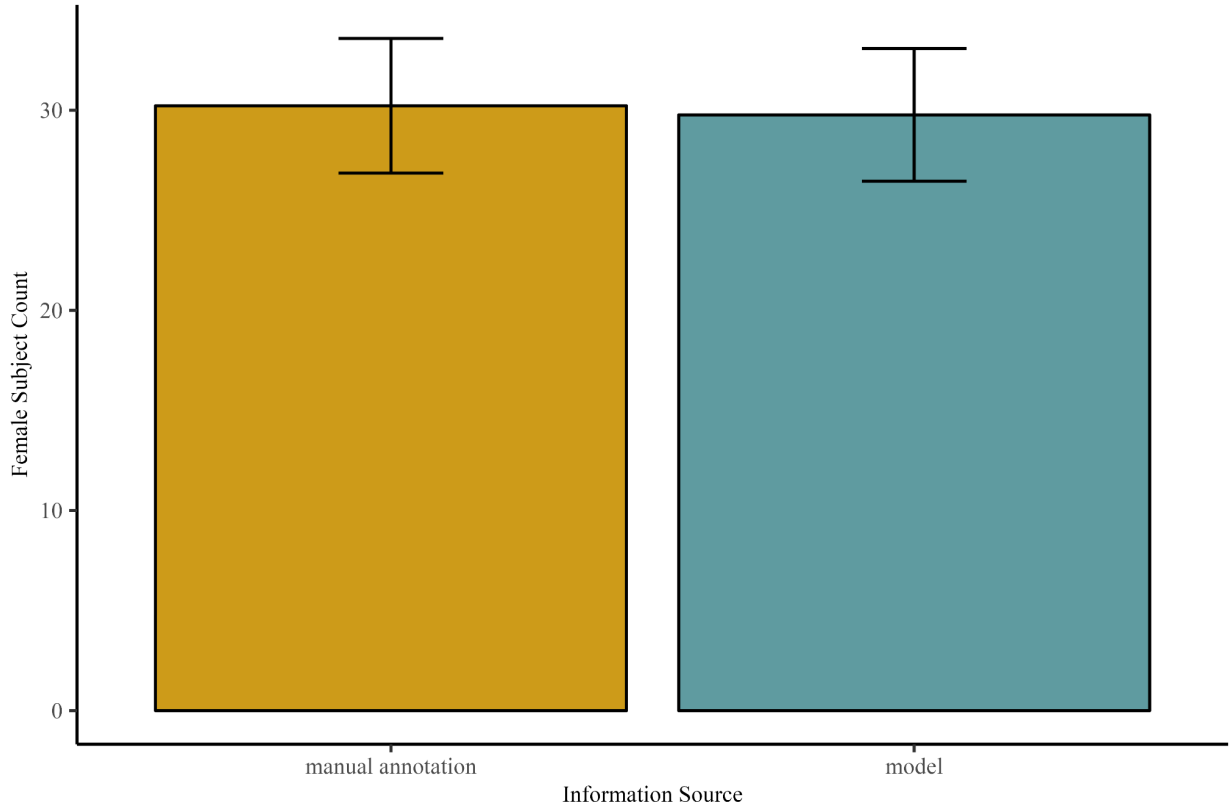
Comparison Between Actual and Predicted Male Subject Counts



Note. This figure visualizes the comparison of actual versus predicted male subject counts within our dataset. Using paired t-tests, we found no significant difference between predicted and actual male subject counts ($t(405) = 0.87$, $p = .39$, $d = 0.004$). These results indicate that there is no statistically significant discrepancy between the predicted and actual counts of male subjects, suggesting that the prediction model performs well in estimating male subject numbers.

Figure 5

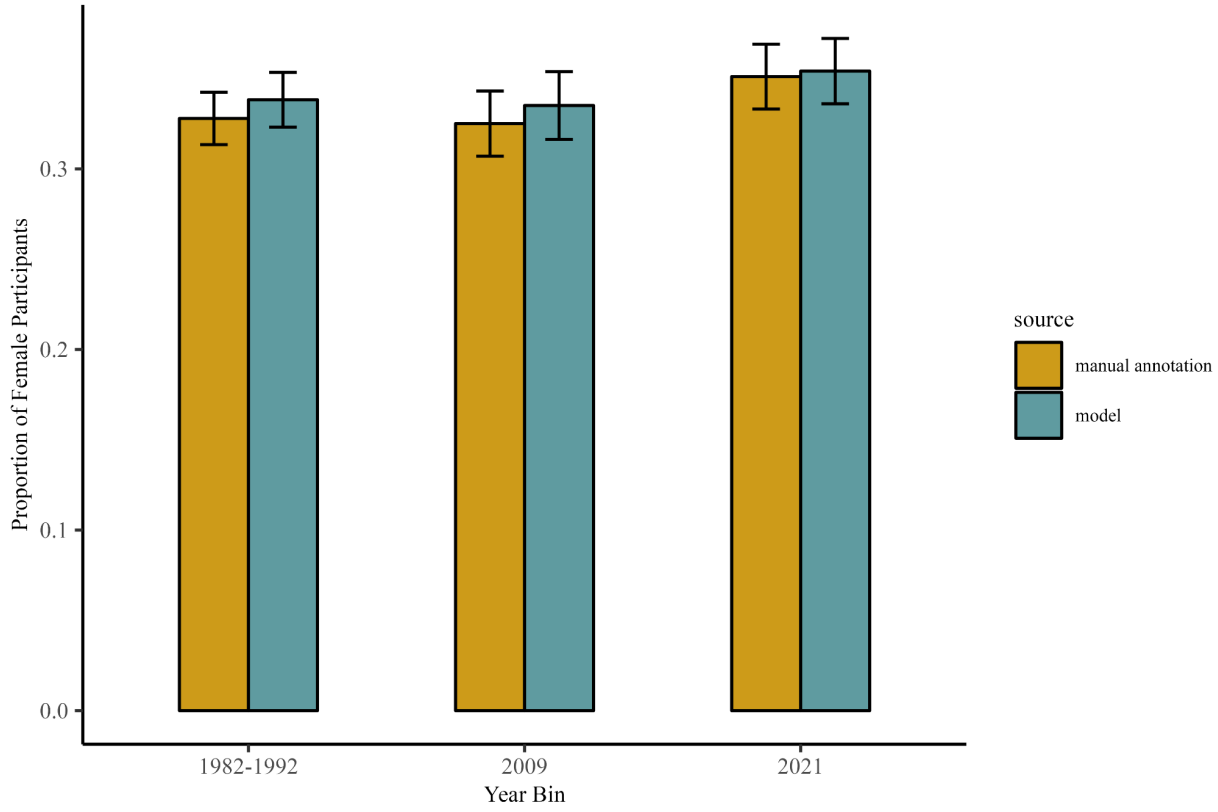
Comparison Between Actual and Predicted Female Subject Counts



Note. This figure illustrates the comparison between actual and predicted counts of female subjects. Utilizing a paired t-test for this analysis, we found no significant difference between predicted and female subject counts (t(403) = 1.12, p = .26, d = 0.007). These findings demonstrate the model's accuracy in predicting the number of female subjects, paralleling the results observed for male subject predictions.

Figure 6

Manually Annotated and Model Predicted Proportion of Female Subjects Over Time



Note. This figure displays the proportion of female subjects in PD clinical research studies over time, as determined via manual annotations (gold) and the GPT4-Turbo-Preview model (blue). We found that both data sources pointed to the same conclusion: there has been no significant change in female participation in clinical research on PD over time.

References

Allen, G. S., Burns, R. S., Tulipan, N. B., & Parker, R. A. (1989). Adrenal medullary transplantation to the caudate nucleus in Parkinson's disease. Initial clinical results in 18 patients. *Archives of Neurology*, *46*(5), 487–491. https://doi.org/10.1001/archneur.1989.00520410021016

Avery, E., & Clark, J. (2016). Sex-related reporting in randomised controlled trials in medical journals. *The Lancet*, *388*(10062), 2839–2840. https://doi.org/10.1016/S0140-6736(16)32393-5

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners* (No. arXiv:2005.14165). arXiv. http://arxiv.org/abs/2005.14165

Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., & Hemphill, L. (2023). A bibliometric review of large language models research from 2017 to 2023. *arXiv*. https://doi.org/10.48550/arXiv.2304.02020

Foppiano, L., Lambard, G., Amagasa, T., & Ishii, M. (2024). *Mining experimental data from materials science literature with large language models* (No. arXiv:2401.11052). arXiv. http://arxiv.org/abs/2401.11052

Gao, T., Fisch, A., & Chen, D. (2021). Making pre-trained language models better few-shot learners. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3816–3830. https://doi.org/10.18653/v1/2021.acl-long.295

Geller, S. E., Koch, A. R., Roesch, P., Filut, A., Hallgren, E., & Carnes, M. (2018). The more things change, the more they stay the same: A study to evaluate compliance with inclusion and assessment of women and minorities in randomized controlled trials. *Academic Medicine*, *93*(4), 630–635. https://doi.org/10.1097/ACM.0000000000002027

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). *Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models* (No. arXiv:2304.01852). arXiv. http://arxiv.org/abs/2304.01852

Patel, M., Nilsson, M. H., Rehncrona, S., Tjernström, F., Magnusson, M., Johansson, R., & Fransson, P.-A. (2021). Spectral analysis of body movement during deep brain stimulation in Parkinson's disease. *Gait & Posture*, *86*, 217–225. https://doi.org/10.1016/j.gaitpost.2021.03.023

Roopesh, N., Akarsh, S., & Babu, C. N. (2021). An optimal data entry method, using web scraping and text recognition. *2021 International Conference on Information Technology (ICIT)*, 92–97. https://doi.org/10.1109/ICIT52682.2021.9491643

Shen, Z., Lo, K., Wang, L. L., Kuehl, B., Weld, D. S., & Downey, D. (2022). VILA: Improving structured content extraction from scientific PDFs using visual layout groups. *Transactions of the Association for Computational Linguistics*, *10*, 376–392. https://doi.org/10.1162/tacl_a_00466

Srivastava, P., Ganu, T., & Guha, S. (2022). *Towards zero-shot and few-shot table question answering using GPT-3* (No. arXiv:2210.17284). arXiv. http://arxiv.org/abs/2210.17284

Tosserams, A., Araújo, R., Pringsheim, T., Post, B., Darweesh, S. K. L., IntHout, J., & Bloem, B. R. (2018). Underrepresentation of women in Parkinson's disease trials. *Movement*

Disorders, *33*(11), 1825–1826. https://doi.org/10.1002/mds.27505

Wheeler, J. J., Allen-Moyer, K. N., Davis, J. M., & Mishra, S. K. (2020). A systematic review of animal models and sex as a variable in itch research. *Itch*, *5*(3), e40–e40. https://doi.org/10.1097/itx.0000000000000040

Woitowich, N. C., Beery, A., & Woodruff, T. (2020). A 10-year follow-up study of sex inclusion in the biological sciences. *eLife*, *9*, e56344. https://doi.org/10.7554/eLife.56344

Chapter 4: Classification of Fall Types in Parkinson Disease From Self-report Data Using Natural Language Processing

Jeanne M. Powell[1], Yuting Guo[1], Abeed Sarker[1], and J. Lucas McKay[1,2]

[1] Emory University, Atlanta, GA 30322, USA

[2] Georgia Institute of Technology, Atlanta, GA 30332, USA

*Published as conference paper at 21st International Conference of Artificial Intelligence in Medicine (2023):

Powell, J.M., Guo, Y., Sarker, A., McKay, J.L. (2023). Classification of Fall Types in Parkinson's Disease from Self-report Data Using Natural Language Processing. In: Juarez, J.M., Marcos, M., Stiglic, G., Tucker, A. (eds) Artificial Intelligence in Medicine. AIME 2023. Lecture Notes in Computer Science(), vol 13897. Springer, Cham. https://doi.org/10.1007/978-3-031-34344-5_20

Abstract

Falls are a leading cause of injury globally, and people with Parkinson's disease are particularly at risk. An important step in reducing the probability of falls is to identify their causes, but manually classifying fall types is laborious and requires expertise. Natural language processing (NLP) approaches hold potential to automate fall type identification from descriptions. The aim of this study was to develop and evaluate NLP–based methods to classify fall types from Parkinson's disease patient self-report data. We trained supervised NLP classifiers using an existing dataset consisting of both structured and unstructured data, including the age, sex/gender, and duration of Parkinson's disease of the faller, as well as the fall location, free-text fall description, and fall class of each fall. We trained supervised classification models to predict fall class based on these attributes, and then performed an ablation study to determine the most important factors influencing the model. The best performing classifier was a hard voting ensemble model that combined the Adaboost, unweighted decision tree, weighted k-nearest neighbor, naïve Bayes, random forest, and support vector machine classifiers. On the testing set, this ensemble classifier achieved an F_1-macro of 0.89. We also experimented with a transformer-based model, but its performance was subpar compared to that of the other models. Our study demonstrated that automatic fall type classification in Parkinson's disease patients is possible via NLP and supervised classification.

Classification of Fall Types in Parkinson Disease From Self-report Data Using Natural Language Processing

Falls are unintentional events where a person lands on a lower level (Bloem et al., 2001; Maki et al., 1994), which can result in significant personal, financial, and health costs (Burns & Kakara, 2018; Florence et al., 2018; Haddad et al., 2019; Stack & Ashburn, 1999). For example, falls were the leading cause of injury-related death in the United States between 2007 and 2016 (Burns & Kakara, 2018). These high costs could be minimized with a better understanding of the causes of falls and subsequent implementation of preventative measures.

Many studies report overall fall frequency without accounting for the circumstances surrounding a fall, which limits our understanding of their etiology (Ross et al., 2017). Falls are heterogeneous and can result from multiple types of biomechanical perturbations, including perturbations to an individual's base of support (BoS; e.g., trips) or center of mass (CoM; e.g., overextension during bending) (Maki et al., 1994). BoS falls are more common in healthy older adults compared to CoM falls (Maki et al., 1994). However, the opposite holds true in subpopulations of people, such as people with Parkinson's disease, where disease-related postural instability results in more CoM falls (Bloem et al., 2001).

People with Parkinson's disease are more likely to fall and be frequent fallers than healthy older adults (Bloem et al., 2001; Stack & Ashburn, 1999). Falling in this population can be incapacitating, often resulting in soft tissue injuries, and disabling even early in disease progression (Bloem et al., 2001). Therefore, it is of particular importance to predict and prevent falls in this population.

A necessary step in this pursuit is to track falls and fall circumstances because risk factors for trips and slips might differ from those for falls due to impaired self-motion or other causes.

To determine the cause of a fall, one must collect free-text information about the circumstances of the fall from the faller (Stack & Ashburn, 1999). Historically, fall classes have been manually coded from these free-text descriptions (Ashburn et al., 2008; Magnani et al., 2020; Pelicioni et al., 2019; Stack & Ashburn, 1999), but this practice is subjective, resource intensive, and difficult to scale. Recent advances in the field of natural language processing (NLP) hold exciting promise to automate processes such as fall type classification from free-text fall descriptions.

NLP techniques have been used in many biomedical domains, including mining unstructured electronic health records (Houssein et al., 2021). For example, Tohira and colleagues trained support vector machine (SVM) and random forest (RF) classifiers to detect falls from ambulance services provider reports (Tohira et al., 2022). Electronic health records and patient self-reports provide rich data that can capture nuances that structured medical data may miss (Guetterman et al., 2018). NLP techniques may aid in Parkinson's disease diagnosis, given its impact on language production. Pérez-Toro and colleagues demonstrated that NLP techniques could be leveraged to distinguish people with Parkinson's disease from healthy older adults based on differences transcribed in speech patterns (Pérez-Toro et al., 2019). Given that Parkinson's disease is the second most common neurodegenerative disease worldwide, affecting over six million people globally (Dorsey et al., 2018), it is extremely important to gain a nuanced understanding of the disease.

Here, we aimed to develop an NLP classification model to distinguish CoM falls from other fall types in people with Parkinson's disease based on patient-provided free-text descriptions. Our particular focus is on CoM falls as they are more common in people with Parkinson's disease (Bloem et al., 2001).

Methods

Fall Self-report Dataset

In a recent study (Bloem et al., 2001), we followed patients with Parkinson's disease for 12-months to correlate fall risk with biomarkers of balance control. Participants tracked falls on monthly "fall calendars" and missing data and fall details were acquired over telephone interviews. Our dataset consisted of 124 fall self-reports collected from 23 individuals. Demographic information about those individuals is provided in **Table 1**.

Each fall self-report included structured data (i.e., age, self-reported sex/gender, and time since Parkinson's disease diagnosis) and unstructured data (free-text description of the fall and its location). Dr. McKay, an expert in the biomechanics of falls, classified fall causes (i.e, CoM, BoS, or 'Other') based on fall descriptions, and Powell classified falls as occurring inside the home or not using location descriptions. Three fall descriptions were modified because the patient described the fall as "same as fall #1". In these cases, the full description used for #1 was copied. The fall descriptions were also manually checked for spelling errors.

The average length of the fall descriptions was 38 words, with the shortest consisting of only 3 words and the longest of 170 words. There were 922 unique words in the dataset before processing and 731 unique words after processing (i.e., stemming and removal of stop words).

Data Pre-processing

All binary categorical variables (i.e., sex/gender, fall class, and fall location) were one-hot encoded. Our numerical factors, age and disease duration, were scaled. Free-text fall descriptions were pre-processed by lowercasing all text, removing English stop words and punctuations, and tokenizing the remaining text. Each word token was stemmed using the Porter stemmer (Porter, 1980). Fall descriptions were then vectorized as follows. We generated two sets

of features from the pre-processed description texts—n-grams and word clusters. A word n-gram is a sequence of contiguous n words in a text segment. This feature enabled us to represent a document using the union of its terms. We used 1-, 2-, and 3-grams as features with the max number of features set to 150. The n-grams were vectorized so that each n-gram was represented by a numeric value in the feature vector indicating its frequency within a given instance. To enable a more generalized representation of the terms, we used the CMU word clusters (Owoputi et al., 2013). The word clusters were generated via a two-step process—dense vector representations of words were first learned from large unlabeled data using the method described in Owoputi et al. (2013) so that similar terms were close together in vector space, and then the words were grouped via hierarchical clustering. The word clusters were represented as unigrams during training, and were also capped at a maximum of 150 features.

Fall Class Classification

We modeled the discrimination between CoM- and Other-class falls as a binary classification problem, using both structured and unstructured features. Because the dataset was relatively small, we applied a predefined 3-fold cross validation for (80% of data) training and (20% of data) evaluation. We experimented with multiple classifiers, specifically: naïve Bayes (NB), K-Nearest Neighbors (KNN), SVM, RF, Adaboost with single split trees as base classifiers, a Decision Tree (DT) classifier, and a hard-voting ensemble classifier with contributions from each of the previously mentioned classifiers. We experimented with both weighted and unweighted KNN and DT classifiers to account for our unbalanced classes. The performances of the classifiers were compared using the F_1-macro score on the test data because that metric is more appropriate when classes are unbalanced. We then performed an ablation

study to determine the individual impact of each factor on model performance, as well as model performance when only trained on the text.

We also modeled the discrimination between CoM- and Other-class falls using the RoBERTa transformer model (Liu et al., 2019). For this experiment, we utilized the unprocessed, free-text fall descriptions to predict fall-class labels and did not include other factors in the model. We applied a 3-fold cross validation for (80% of data) training and (20% of data) evaluation. The model was trained for 2, 5, and 10 epochs. All model parameters were fine-tuned during training. Performance was measured by taking the median of the F_1-macro score.

Confidence intervals were calculated using bootstrapping with samples taken from the test prediction and ground truth datasets with replacement over 1000 iterations.

Terminology Choice: Sex/Gender

In this study, 'sex/gender' was categorized as 'male', 'female', and 'unknown/not reported' within our data corpus. These categorizations were input by researchers following discussion with patients, with missing values supplemented from medical charts. The numerous variables correlated with these categories, for example social norms that might influence the circumstances of falls (e.g., yard work, taking out the trash), and the potential biomechanical differences in gait variability, etc., span what could be considered “gender” and “sex” (for NIH definitions of these terms, see https://orwh.od.nih.gov/sex-gender). To acknowledge this complexity and the inseparability of these two concepts, particularly in the context of fall patterns, we use the term 'sex/gender' in this manuscript.

Results

Demographic Information

Our fall dataset included 23 individuals with an average age of 67.3 years and an average Parkinson's disease duration of 8.9 years (**Table 1**). There was no significant difference between the overall fall frequency by sex/gender (p = .203), nor in fall frequency between sex/gender within each fall class (CoM: p = .267; BoS: p = .662; Other: p = .614).

Of the 124 unique falls, 88 falls occurred due to perturbations to the individuals' center of mass (CoM), 25 falls occurred due to perturbations to the individuals' base of support (BoS), and the remaining 11 falls occurred for other reasons, including falling during exercise (n = 9), low blood pressure (n =1), and rolling out of bed (n = 1). Because of the relatively low number of BoS- and Other-class falls in our dataset, we collapsed the data into CoM- and Other-class falls. There were 88 CoM-class falls and 36 Other-class falls.

Binary Classification Hyperparameter Selection

For each classifier, we iterated through a wide range of hyper-parameter values and calculated the mean squared error between the actual fall class and predicted fall class at each hyperparameter value. Mean squared error is inversely proportional to model performance. Therefore, the hyperparameter value that resulted in the lowest mean squared error for each model was chosen (**Table 2**).

Binary Classifier Performance

Each classifier was trained on the same 80% of the dataset using 3-fold cross validation with the same folds. Then, each classifier was tested on the same 20% of data that had been excluded from the training phase, and we report the resulting F_1-macro score and its 95%

confidence interval (CI). The ensemble classifier had the highest performance out of all trained classifiers (F_1-macro = 0.89, 95% CI: [0.67-1]; **Table 2**).

Adaboost.

We determined the optimal number of estimators for the Adaboost model to be 27 by training on all values of n between 1 and 100, inclusive, and selecting the value of n with the lowest mean squared error between the true and predicted fall classes. The Adaboost model achieved an F_1-macro of 0.80 (95% CI: [0.55-0.96]; **Table 2**).

Decision Tree.

We trained two DT classifiers, one with the weight set to the inverse frequency of each class in the training set (DTa) and the other with equal weights for both classes (DTb). We determined the optimal maximum depth of DTa to be 23 and the optimal maximum depth of DTb to be 18. The weighted model DTa achieved an F_1-macro of 0.67 (95% CI: [0.45-0.85]; **Table 2**). The unweighted model DTb performed slightly better with an F_1-macro of 0.72 (95% CI: [0.50-0.90]; **Table 2**).

K-Nearest Neighbor.

We trained two KNN classifiers, one with the weight function equal to ‘distance’ (KNNa) and the other with the weight function equal to ‘uniform’ (KNNb). For the weighted model, KNNa, the optimal value of k was 6 and for the unweighted model, KNNb, the optimal value of k was 10. The weighted model KNNa achieved an F_1-macro of 0.84 (95% CI: [0.66-1; **Table 2**). The unweighted model KNNb achieved an F_1-macro of 0.78 (95% CI: [0.53-0.95]; **Table 2**).

Naïve Bayes.

We trained a Gaussian Naïve bayes classifier. The model achieved an F_1-macro of 0.84 (95% CI: [0.63-1]; **Table 2**).

Random Forest.

We determined the optimal number of estimators for the RF model to be 25 by training each model on all values of n between 1 and 100, inclusive, and selecting the value of n that resulted in the lowest mean squared error between the true fall classes and predicted fall classes. The RF model achieved an F_1-macro of 0.78 (95% CI: [0.53-0.95]; **Table 2**).

Support Vector Machine.

We trained the SVM model with gamma set to 'scale' and kernel set to 'rbf' on all values of C between 1 and 100 inclusive. We found that the mean squared error between the true and predicted fall class was at its lowest when C was 4. The model achieved an F_1-macro of 0.78 (95% CI: [0.53-0.95]; **Table 2**).

Ensemble.

The ensemble was composed of one of each type of the classifiers described above. The weighted KNN model KNNa and unweighted DT model DTa were included in the ensemble because they outperformed their complementary model. We set the voting for the ensemble equal to hard. The ensemble achieved an F_1-macro of 0.89 (95% CI: [0.67-1]; **Table 2**).

Ablation Study of Features in the Ensemble Model

We chose to perform an ablation study on the ensemble model because it achieved the highest F_1-macro score out of all trained classifiers. The F_1-macro score of the model decreased with the removal of each factor, except age (**Table 3**).

Binary Classification Using RoBERTa

The RoBERTa transformer model was trained on 80% of the dataset using 3-fold cross validation. Classifier performance was measured using the F_1-macro metric. The RoBERTa model had equal performance across epochs (F_1-macro = 0.42; **Table 4**).

We also trained the machine learning models on the vector representations generated by RoBERTa using the same hyperparameters in **Table 2**. The results show that using RoBERTa as a feature generator underperformed compared to using the n-grams and word clusters as features (**Table 5**).

Discussion

We trained multiple machine learning classifiers to perform a binary classification task to categorize falls as CoM- or Other-type falls based on patient-provided free-text descriptions of the circumstances surrounding each fall. We found that a hard-voting ensemble classifier performed better than individual classifiers and transformer-based models on this task, achieving an F_1-macro of 0.89 (95% CI: [0.67-1]). Importantly, our ensemble approach obtained high performance despite the relatively small size of the annotated dataset, which is often the limiting factor for supervised classification tasks. This serves as a proof of concept that the historically resource intensive task of manual fall type classification can be automated using NLP models, allowing for large-scale fall classification for research purposes.

Current clinical best practices for tracking falls in Parkinson's disease involve retrospective patient reports during regular Neurologist visits. Although standardized instruments exist (Harris et al., 2021), they are very uncommon in clinical practice due to the burden on patients and providers. A technology for classification of falls circumstances and causes based on patient reports may enable future "online trials" and reduce misclassification errors in research studies, which might contribute to the high variability across studies applying exercise on fall risk in Parkinson's disease (Allen et al., 2022).

Better Performance From ML Classifiers

An unexpected finding from our study was that traditional machine learning classifiers and their ensemble outperformed the RoBERTa transformer model on this classification task. Initially, we hypothesized that a transformer-based architecture would outperform machine learning classifiers because transformer-based models rely on pretrained models built from exceptionally large corpora and typically perform better even with limited training data. RoBERTa, which is a robustly optimized model, has been reported to perform extremely well on text classification and other NLP tasks. However, this was not the case in our study. The RoBERTa model categorized all falls as CoM falls, suggesting that both the small size and class imbalance of our data was not ideal for this model, and it failed to pick up the linguistic markers associated with each type of fall. Given the results of our ablation study revealing that non-text factors are important contributors to model performance, we suspect that a key reason for the relatively poor performance of the transformer model stemmed from only using the free-text descriptions. In the future, we will attempt to train a biomedical-specific transformer, such as BioBERT (Lee et al., 2020), to investigate whether it performs better given its specialized biomedical vocabulary. We will also explore how structured data elements may be effectively incorporated with transformer-based models.

Inability to Detect Sex/Gender Differences in Fall Frequency and Type

The incidence of fall events in women is generally higher than in men (Duckham et al., 2013). However, the difference in frequency of falls in individuals with PD by sex/gender is unclear. Some studies suggest that women with PD are at higher risk of experiencing frequent falls compared to men with PD (Cerri et al., 2019). Conversely, other research indicates no significant sex/gender difference in fall frequency among PD patients, a contrast to patterns

observed in healthy older adults (Bloem et al., 2001). Our study found no statistically significant difference in falls between sexes/genders. This lack of significance may be attributed to our small sample size, limiting our study's power to detect sex/gender differences.

Future Work

To improve our study's generalizability, we plan to expand our small, unbalanced dataset with limited vocabulary by extracting more reports from our database. This approach will allow us to test and retrain our ensemble model for better performance. We also plan to explore other transformer-based models and strategies for integrating structured data.

Conclusion

Our study demonstrated that it is possible to automate the laborious process of fall type identification by using supervised classification methods that integrate structured and unstructured data. An ensemble classification approach produced excellent results, outperforming a state-of-the-art transformer model despite the small size of annotated data. Find the dataset and code at the following repository: https://github.com/jeannempowell/PD_Falls_NLP.

Acknowledgements

We would like to thank our study participants without whom this study would not be possible.

Tables

Table 1

Demographic Information and Fall Frequency of Patients

	Overall	**Women**	**Men**	**P-Value**
n	23	8	15	
Age, mean (SD)	67.3 (7.1)	65.4 (6.4)	68.3 (7.5)	0.333
Duration, mean (SD)	8.9 (5.1)	8.6 (5.5)	9.1 (5.1)	0.838
All Falls, mean (SD)	5.4 (4.9)	7.5 (6.1)	4.3 (4.0)	0.203
CoM, mean (SD)	3.8 (4.5)	5.5 (5.5)	2.9 (3.8)	0.267
BoS, mean (SD)	1.1 (1.0)	1.2 (1.5)	1.0 (0.7)	0.662
Other, mean (SD)	0.5 (1.5)	0.8 (2.1)	0.3 (1.0)	0.614

Note. This table presents the demographic characteristics and fall frequencies among 23 PD patients (8 women and 15 men) from our dataset, which included 124 self-reported falls. It summarizes the age, duration of PD, and fall frequencies (total falls, center of mass (CoM) falls, base of support (BoS) falls, and other falls), providing mean values with standard deviations (SD). Statistical comparisons between genders were conducted using t-tests, with P-values indicating the level of statistical significance for differences observed between men and women.

Table 2

Classifier Performance at Predicting Fall Type

Classifier	Hyperparameters	F_1-macro	95% CI
Adaboost	n_estimators = 27	0.80	0.55-0.96
DTa	max_depth = 23, class_weight = {0:70.0, 1:29.0)	0.67	0.45-0.85
DTb	max_depth = 18	0.72	0.50-0.90
KNNa	k = 6, weights = 'distance'	0.84	0.66-1
KNNb	k = 10, weights = 'uniform'	0.78	0.53-0.95
NB	N/A	0.84	0.63-1
RF	n_estimators = 25	0.78	0.53-0.95
SVM	Gamma = 'scale', kernel = 'rbf', C = 4	0.78	0.53-0.95
Ensemble	{NB, KNNa, SVM, RF, Adaboost, DTb} voting = 'hard'	**0.89**	**0.67-1**

Note. This table outlines the performance of different classifiers used to predict fall types among Parkinson's disease (PD) patients, evaluated through the F_1-macro score along with their respective 95% confidence intervals (CIs). The classifiers include AdaBoost, Decision Trees (DTa and DTb), K-Nearest Neighbors (KNNa and KNNb), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and an Ensemble method combining NB, KNNa, SVM, RF, AdaBoost, and DTb with hard voting. The table also details the hyperparameters optimized for each classifier, highlighting their impact on the prediction accuracy of fall types, with the Ensemble method exhibiting the highest F_1-macro score.

Table 3

Results of Ablation Study on the Ensemble Model

Dropped Feature(s)	**F_1-macro**	**95% CI**
None	**0.89**	0.67-1
Sex/Gender	0.80	0.59-0.96
Location	0.84	0.66-1
PD Duration	0.80	0.59-0.95
Age	**0.89**	**0.70-1**
Word Clusters	0.82	0.63-0.96
1,2,3-grams	0.78	0.55-0.95
Sex/gender, Location, PD Duration, & Age	0.83	0.58-1

Note. This table presents the results of an ablation study conducted to assess the impact of omitting various features on the F_1-macro score of an ensemble model designed for classifying fall types in Parkinson's disease (PD) patients. Features tested for their importance include Sex/Gender, Location, PD Duration, Age, Word Clusters, and 1,2,3-grams, in addition to a combination of Sex/Gender, Location, PD Duration, and Age. The table shows the performance of the ensemble model with none of the features omitted (baseline) and the performance after each feature or combination of features is excluded, alongside the 95% confidence intervals (CIs) for each scenario. The comparison highlights the relative contribution of each feature to the model's predictive accuracy, demonstrating the varying degrees of impact their removal has on the model's effectiveness in accurately classifying fall types among PD patients.

Table 4

Performance of RoBERTa Model

Epochs	F_1-macro	95% CI
2	0.42	0.37-0.46
5	0.42	0.37-0.45
10	0.42	0.37-0.45

Note. This table provides a detailed overview of the RoBERTa model's performance in terms of the F_1-macro score across different epochs during the training phase. It compares the model's effectiveness after being trained for 2, 5, and 10 epochs, respectively. Additionally, the table includes the 95% confidence intervals (CIs) for the F_1-macro scores at each epoch level, offering insights into the stability and variability of the model's performance across these training intervals.

Table 5

Classifier Performance using RoBERTa-generated Vector Representations

Model	F_1-macro	95% CI
Adaboost	0.61	0.40-0.81
DTa	0.70	0.49-0.88
DTb	0.58	0.38-0.77
KNNa	0.71	0.44-0.90
KNNb	0.47	0.44-1
NB	0.48	0.29-0.66
RF	0.61	0.39-0.81
SVM	0.41	0.34-0.46

Note. This table summarizes the performance of various classifiers utilizing vector representations generated by the RoBERTa model, as measured by the F_1-macro score. For each classifier, including Adaboost, decision trees (DTa and DTb), k-nearest neighbors (KNNa and KNNb), Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM), the table lists the corresponding F_1-macro scores along with the 95% confidence intervals (CIs).

References

Allen, N. E., Canning, C. G., Almeida, L. R. S., Bloem, B. R., Keus, S. H., Löfgren, N., Nieuwboer, A., Verheyden, G. S., Yamato, T. P., & Sherrington, C. (2022). Interventions for preventing falls in Parkinson's disease. *Cochrane Database of Systematic Reviews*, *2022*(8). https://doi.org/10.1002/14651858.CD011574.pub2

Ashburn, A., Stack, E., Ballinger, C., Fazakarley, L., & Fitton, C. (2008). The circumstances of falls among people with Parkinson's disease and the use of falls diaries to facilitate reporting. *Disability and Rehabilitation*, *30*(16), 1205–1212. https://doi.org/10.1080/09638280701828930

Bloem, B. R., Grimbergen, Y. A. M., Cramer, M., Willemsen, M., & Zwinderman, A. H. (2001). Prospective assessment of falls in Parkinson's disease. *Journal of Neurology*, *248*(11), 950–958. https://doi.org/10.1007/s004150170047

Burns, E., & Kakara, R. (2018). Deaths from falls among persons aged ≥65 years—United States, 2007–2016. *Morbidity and Mortality Weekly Report*, *67*(18), 509–514. https://doi.org/10.15585/mmwr.mm6718a1

Cerri, S., Mus, L., & Blandini, F. (2019). Parkinson's disease in women and men: What's the difference? *Journal of Parkinson's Disease*, *9*(3), 501–515. https://doi.org/10.3233/jpd-191683

Dorsey, E. R., Elbaz, A., Nichols, E., Abbasi, N., Abd-Allah, F., Abdelalim, A., Adsuar, J. C., Ansha, M. G., Brayne, C., Choi, J.-Y. J., Collado-Mateo, D., Dahodwala, N., Do, H. P., Edessa, D., Endres, M., Fereshtehnejad, S.-M., Foreman, K. J., Gankpe, F. G., Gupta, R., … Murray, C. J. L. (2018). Global, regional, and national burden of Parkinson's disease, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The*

Lancet Neurology, *17*(11), 939–953. https://doi.org/10.1016/S1474-4422(18)30295-3

Duckham, R. L., Procter-Gray, E., Hannan, M. T., Leveille, S. G., Lipsitz, L. A., & Li, W. (2013). Sex differences in circumstances and consequences of outdoor and indoor falls in older adults in the MOBILIZE Boston cohort study. *BMC Geriatrics*, *13*(1), 133. https://doi.org/10.1186/1471-2318-13-133

Florence, C. S., Bergen, G., Atherly, A., Burns, E., Stevens, J., & Drake, C. (2018). Medical costs of fatal and nonfatal falls in older adults: Medical costs of falls. *Journal of the American Geriatrics Society*, *66*(4), 693–698. https://doi.org/10.1111/jgs.15304

Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. V. (2018). Augmenting qualitative text analysis with natural language processing: Methodological study. *Journal of Medical Internet Research*, *20*(6), e231. https://doi.org/10.2196/jmir.9702

Haddad, Y. K., Bergen, G., & Florence, C. S. (2019). Estimating the economic burden related to older adult falls by state. *Journal of Public Health Management and Practice*, *25*(2), E17–E24. https://doi.org/10.1097/PHH.0000000000000816

Harris, D. M., Duckham, R. L., Daly, R. M., Abbott, G., Johnson, L., Rantalainen, T., & Teo, W.-P. (2021). Development of a Parkinson's disease specific falls questionnaire. *BMC Geriatrics*, *21*(1), 614. https://doi.org/10.1186/s12877-021-02555-6

Houssein, E. H., Mohamed, R. E., & Ali, A. A. (2021). Machine learning techniques for biomedical natural language processing: A comprehensive review. *IEEE Access*, *9*, 140628–140653. https://doi.org/10.1109/ACCESS.2021.3119621

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining.

Bioinformatics, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach* (No. arXiv:1907.11692). arXiv. http://arxiv.org/abs/1907.11692

Magnani, P. E., Genovez, M. B., Porto, J. M., Zanellato, N. F. G., Alvarenga, I. C., Freire, R. C., & de Abreu, D. C. C. (2020). Use of the BESTest and the Mini-BESTest for fall risk prediction in community-dwelling older adults between 60 and 102 years of age. *Journal of Geriatric Physical Therapy*, *43*(4), 179–184. https://doi.org/10.1519/JPT.0000000000000236

Maki, B. E., Holliday, P. J., & Topper, A. K. (1994). A prospective study of postural balance and risk of falling in an ambulatory and independent elderly population. *Journal of Gerontology*, *49*(2), M72–M84. https://doi.org/10.1093/geronj/49.2.M72

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). *Improved part-of-speech tagging for online conversational text with word clusters*. 380–390. https://aclanthology.org/N13-1039

Pelicioni, P. H. S., Menant, J. C., Latt, M. D., & Lord, S. R. (2019). Falls in Parkinson's disease subtypes: Risk factors, locations and circumstances. *International Journal of Environmental Research and Public Health*, *16*, 2216. https://doi.org/10.3390/ijerph16122216

Pérez-Toro, P. A., Vásquez-Correa, J. C., Strauss, M., Orozco-Arroyave, J. R., & Nöth, E. (2019). Natural Language Analysis to Detect Parkinson's Disease. In K. Ekštein (Ed.), *Text, Speech, and Dialogue* (Vol. 11697, pp. 82–90). Springer International Publishing. https://doi.org/10.1007/978-3-030-27947-9_7

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Ross, A., Yarnall, A. J., Rochester, L., & Lord, S. (2017). A novel approach to falls classification in Parkinson's disease: Development of the Fall-Related Activity Classification (FRAC). *Physiotherapy*, *103*(4), 459–464. https://doi.org/10.1016/j.physio.2016.08.002

Stack, E., & Ashburn, A. (1999). Fall events described by people with Parkinson's disease: Implications for clinical interviewing and the research agenda. *Physiotherapy Research International*, *4*(3), 190–200. https://doi.org/10.1002/pri.165

Tohira, H., Finn, J., Ball, S., Brink, D., & Buzzacott, P. (2022). Machine learning and natural language processing to identify falls in electronic patient care records from ambulance attendances. *Informatics for Health and Social Care*, *47*(4), 403–413. https://doi.org/10.1080/17538157.2021.2019038

Chapter 5: Analyzing Facebook Use in Parkinson's disease Patients and Caregivers: A preliminary Analysis of an Exploratory Study

Jeanne M. Powell[1], Sahithi Lakamana[1], Abeed Sarker[1,2], and J. Lucas McKay[1,2]

[1] Emory University, Atlanta, GA 30322, USA

[2] Georgia Institute of Technology, Atlanta, GA 30332, USA

Abstract

Parkinson's disease (PD) poses a multifaceted challenge to patients and caregivers due to its variable symptoms and progression. Traditional research often overlooks the subtle yet critical aspects of patients' experiences and the markers of disease progression, which are essential for tailoring management approaches. Our research employs natural language processing (NLP) and text mining techniques to dissect Facebook interactions of people with PD and their caregivers. Our objectives were to: 1) assess their Facebook activity, particularly in PD-focused groups and pages, and their sharing of PD-related information; 2) identify mentions of early markers of disease progression and symptoms present before diagnosis; and 3) discern any behavioral variances between patients and caregivers. We crafted a binary classifier to sift through PD-related content and conducted both qualitative and quantitative content analysis. Our findings show that 81% of the participants we reached had Facebook accounts, with 87% consenting to data sharing. Both groups actively engaged in PD-specific online communities, without notable differences in their levels of activity. Our analysis uncovered references to progression indicators (e.g., falls, gait freezing) and symptoms occurring prior to diagnosis (e.g., rigidity, fatigue). This underscores social media's utility in capturing the nuanced experiences of those affected by PD and their caregivers.

Analyzing Facebook Use in Parkinson's disease Patients and Caregivers: A Preliminary Analysis of an Exploratory Study

> *I am grateful for this group because there is probably no other place in the world where we could find 300 kind people who understand the challenges we each are facing and will face and are so willing to share personal experience with so little inhibition. This pool of firsthand knowledge and experience we are creating—and the speed with which we share new developments—is a tool few if any medical practices could replicate. Very few days go by when I don't learn something new or get needed reassurance, and I expect to share my experience as it grows and may become useful to those journeying on the path behind me. There's an occasional ebb in the collective mood here, but someone always seems to generate positive energy when the vibe is headed south—a redirecting force that would be seriously missed if we each faced our challenges alone.*
>
> - Comment in a Parkinson's Disease Facebook Group

Parkinson's disease (PD) is a progressive neurodegenerative disorder that affects over six million people worldwide, imposing a significant socioeconomic burden on healthcare systems (Armstrong & Okun, 2020; Bloem et al., 2021; Dorsey et al., 2018). Characterized by a wide range of motor and non-motor symptoms such as tremors, rigidity, and fatigue, PD's highly heterogeneous nature means that patients experience varied symptoms, progression rates, and treatment responses (Bloem et al., 2021). This diversity complicates PD management for clinicians and poses unique challenges for both patients and their caregivers during their healthcare journey.

While there is a wealth of research devoted to understanding PD, relatively little of it has been on the lived experiences of PD patients and their caregivers. Caregivers experience a lot of stress and strain as a result of their loved one's disease. (Macchi et al., 2020). While patients navigate these complex clinical landscapes, caregivers find themselves entwined in a network of emotional and physical challenges, which have been notably underrepresented in the literature (Rajith & Angiel, 2023). Monitoring the disease from a caregiver's point of view could provide invaluable insights into better symptom management and quality of life. Moreover, traditional research methods often fail to capture real-time markers of disease progression such as falls and freezing of gait (FOG), critical indicators that could alert healthcare providers to emerging issues and help tailor more personalized treatment strategies (Deane et al., 2014; Okuma, 2014).

The ubiquity of social media platforms, such as Facebook, offers an unprecedented opportunity for capturing the rich tapestry of lived experiences shared over the course of years. People often share their thoughts, emotions, and even medical experiences openly (Dudina et al., 2019), providing a treasure trove of data waiting to be tapped into. Natural language processing (NLP)—a multidisciplinary field combining linguistics and computer science—makes analyzing vast amounts of unstructured text data with nuance efficient and therefore tenable. By harnessing NLP algorithms, we can delve into the layers of conversations, posts, and comments to extract meaningful patterns related to PD patient and caregiver experiences.

The present study aimed to leverage technological advancements to fill the gap in PD research by qualitatively and quantitatively exploring PD-related content shared on Facebook by patients and caregivers. We aspired to understand the intricacies of the lived experiences of PD patients and their caregivers by mining Facebook posts and comments. Our research focused on examining the extent to which people with PD and their caregivers use Facebook, join or follow

PD-specific groups and pages, and post about PD-related content. Additionally, we aimed to detect early markers of disease progression and identify premorbid symptoms based on Facebook narratives, while also investigating differences in Facebook activity between patients and caregivers.

This study introduces the use of NLP algorithms for the monitoring of PD markers, considering both patients and caregivers. This approach facilitates automated, real-time healthcare analytics and contributes to a broader understanding of PD. Our research highlights the potential of using social media as a data source in healthcare.

Methods

Study Population

The inclusion criteria for the study were as follows: participants must either be diagnosed with PD or be caregivers of someone with the condition, possess an active Facebook account, and have the ability to communicate and make posts in English. Individuals with atypical parkinsonisms were excluded from the study. No additional exclusion criteria were specified. All participants provided written informed consent, which was obtained either in person or through teleconferencing, in accordance with protocols approved by Emory University's Institutional Review Board (STUDY00005722).

Patient Recruitment

We recruited participants for our study through multiple channels: recontacting previous PD study participants, distributing flyers at the Emory Brain Health Center Movement Clinic, informing affiliated neurologists, advertising at events like the Parkinson’s Disease Foundation Movement Day 2023, and leveraging social media, including a post on the Georgia Clinical and Translational Science Alliance (CTSA) Facebook account.

Interview

Participants provided written consent and were interviewed on Zoom for one hour, during which data were gathered on demographics data (age, location, education, race/ethnicity) and clinical variables (PD diagnosis, disease duration, symptom onset, treatments). The Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) parts I, II, and IV (Goetz et al., 2008) were used to assess clinical status. Study staff helped participants download their Facebook data (posts, comments, pages, groups) as JSON files, which were securely uploaded to REDCap for analysis. Of note, an individual's Facebook data download only includes Facebook activities performed by that user, not their friends. For example, a person's Facebook data includes a comment written on one of their friend's posts. However, it does not include any information about that friend's post.

Participant Account Ownership

We calculated descriptive statistics on participant recruitment success and Facebook account ownership, using point estimates and applying the plus four method to calculate averages and confidence intervals.

Overall Facebook Activity

Each record in the exported Facebook JSON files represents an activity by a user, encompassing actions such as sharing, reacting, commenting on content, and group and page interactions. Overlap was noted in 'your_posts_check_ins_photos_and_videos', 'uncategorized_photos', and 'album' files; we exclusively utilized 'your_posts_check_ins_photos_and_videos' to prevent data duplication. Additionally, we used only 'pages_you_follow' data due to overlap with 'pages_you've_liked'. See **Appendix A** for more information about these files.

We analyzed Facebook usage by normalizing activity to annual averages according to account duration and calculated two 95% confidence intervals for activity, employing the add four method, both with and without outliers. We also compared activity between PD patients and caregivers using a Wilcoxon rank sum exact test, including outliers.

PD-relevant Groups and Pages

We analyzed the extent to which participants joined/followed PD-relevant groups and pages. We manually identified PD-relevant groups/pages, which included those dedicated to notable individuals with PD like Michael Fox and Muhammad Ali, as well as those about exercise, research, and government health agencies. We compared the number of PD-relevant groups and pages joined/followed between PD patients and caregivers using nonparametric tests, also noting whether participants were admins of any PD-related group or page.

PD-Relevant Posts Identification

Activity records containing participant-written text data, such as posts, comments, and captions, were extracted from multiple JSON files. For simplicity, all text data types are collectively termed 'Posts'. Briefly, we developed a pipeline to identify PD-relevant posts by creating a term dictionary and training a text classifier on manually labeled data. More information on this process is described below.

Participant-Written Text Corpus

We processed each participant's JSON files, extracting all text records and rectifying Unicode abnormalities. We removed Facebook-specific formatting (e.g., tags) and hyperlinks. Duplicate texts were removed.

PD-Related Term Dictionary and Search Strategy

We created a keyword dictionary encompassing PD-related terms including symptoms, treatments, and related discussions, sourced from academic literature (Bloem et al., 2021 [motor, non-motor, and premorbid symptoms, genes]; Powell et al., 2023 [fall-related terms]) and WebMD (https://www.webmd.com/parkinsons-disease/drug-treatments). ChatGPT4 assisted in generating medical and colloquial terms related to these categories. The dictionary was refined iteratively, incorporating new terms identified during manual screening. See **Appendix B** for our full list of terms.

Texts were pre-processed using lowercasing, punctuation removal, and the Porter stemming algorithm (Porter, 1980). A keyword search flagged texts containing any stemmed term, long unstemmed terms, or exact short unstemmed terms for manual review.

Binary Classifier Development for PD-Related Content Detection

Given the impracticality of manual post identification on Facebook due to data volume, we developed a binary text classifier using NLP algorithms. This automated approach enhances scalability and consistency in identifying PD-related content.

Manual Screening. Texts identified by the keyword search were manually classified for the Ground Truth PD Corpus as PD-relevant or irrelevant. Relevance was determined by explicit mentions of PD aspects, such as the disease, medications, motor and non-motor symptoms, treatment complications, falls, medical interventions, socio-political issues, beneficial exercises, events, or research related to PD. Irrelevant texts included discussions of symptoms in non-PD individuals, unrelated medical treatments, non-impactful socio-political discourse, and general content like viral messages or TV show discussions. Ambiguous or contextually insufficient posts were also deemed irrelevant.

Binary Classifier Experiments. The corpus was anonymized using the spaCy library's named entity recognition NLP tool, and features for model development included n-grams, participant demographics, and text perspective. The data was split into 80/20 for training and testing. We used Scikit-learn's CountVectorizer for feature extraction of 1000 tokens. Classifier experiments used a variety of algorithms and an ensemble model, optimized for recall using 10-fold cross validation. Post-optimization, models were evaluated for accuracy, recall, precision, and F_1-score. The optimized model was re-fit to the full manually labeled dataset.

PD-relevant Posts

PD-relevant posts were extracted from the participant-written text corpus by flagging terms from our PD-related dictionary and classifying them as relevant or irrelevant using our binary classifier. Relevant posts were manually reviewed, excluding those related to caregiver activities not involving PD. We calculated the frequency of PD-related posts per participant, distinguishing between pre- and post-diagnosis posts for PD individuals. We analyzed the average, standard deviation, and range of posts for both PD individuals and caregivers, categorizing posts based on their sharing context (broadly on Facebook or in PD-related groups). A Wilcoxon rank-sum test compared the post numbers and group-specific post proportions between PD individuals and caregivers, reporting test statistics and p-values.

Qualitative analysis involved categorizing PD-related posts, creating word clouds from the top 100 words used, stratified by pre- or post-diagnosis for PD individuals and separately for caregivers, only for those with more than five posts. These word clouds and manual screening themes informed the construction of personal narratives and a detailed case study on an individual's medical history.

Facebook-derived Medical History Case Study

This case study analyzed the Facebook-derived medical data of an individual with PD, focusing on posts manually identified as relevant. We extracted key information, including family history, PD diagnosis details, symptoms, treatments, side effects, disease progression, exercise routines, and quality of life impacts. The data was organized chronologically based on post timestamps to construct a detailed medical history. This timeline was supplemented with retroactive information found in later posts to ensure completeness, capturing the individual's evolving condition, treatment responses, and the impact of PD on their daily life as depicted in their social media activity.

Results

Participants

In our study, we successfully contacted and either enrolled or obtained refusal reasons from 54 individuals. Among them, 37 had a Facebook account and consented to participate. Of these 37, 23 completed the study and provided their Facebook data.

Regarding refusals, out of 17 individuals who declined participation, 4 reported having a Facebook account, 8 did not, and for 5 the Facebook account status was unknown. One person who has a Facebook account was willing to participate, but his diagnosis had changed to an atypical parkinsonism and therefore no longer qualified for the study.

In summary, we found that 81% (71-92%) of potential participants we approached possessed a Facebook account. Of these, 87% (77-97%) were willing to share their data, inclusive of those who later withdrew or did not schedule a study meeting. See **Table 1** for detailed participant demographics.

Overall Facebook Activity

The average length of Facebook account ownership in our sample was 13 ± 3 years, with participants engaging in an average of 2,396 (95% CI: [-24, 4,816]) activities annually (median: 850, range: 9 to 26,982). For people with PD, the average length of Facebook account ownership was 13 ± 3 years, with an annual average of 2,338 (95% CI: [-1,184, 5,861]) activities performed on Facebook (median: 666, range: 9 to 26,982). Caregivers had an average account duration of 14 ± 3 years, with an annual average of 2,528 (95% CI: [344, 4,712]) activities performed on Facebook (median: 1,888, range: 422 to 6,666).

With outliers removed, on average, participants engaged in 835 (95% CI: [463, 1,207]) activities on Facebook per year. Participants with PD engaged in an average of 695 (95% CI: [288, 1,103]) activities per year. Caregivers engaged in an average of 1,254 (95% CI: [134, 2374]) activities per year.

A Wilcoxon rank sum test comparing PD individuals and caregivers showed no significant difference in Facebook activity ($W = 80$, $p = .12$, $d = 0.03$). This analysis has a statistical power of 0.55, which indicates a moderate probability of correctly identifying a true effect of size $d = 0.2$. This level of power suggests that the study may not be adequately equipped to reliably detect small effect sizes, potentially leading to a failure to identify meaningful differences when they do exist.

PD Groups and Pages

All participants in our study joined at least one Facebook group. Overall, 56% of people with PD and 71% of caregivers joined at least one PD-related Facebook group. Three people with PD were admins of one PD-related group each. People with PD joined an average of 5 (SD = 11, Range: 0-39), and caregivers joined an average of 5 (SD = 8, Range: 0-22) PD-related

groups. These groups were mostly related to fitness, PD support, specific demographic communities, DBS support/information, and PD advocacy, with some groups specific to caregivers (See **Table 2**). We found no difference in the number of PD-related groups joined between people with PD and caregivers (W=66.5, p = .49, d = .01). This analysis has a statistical power of 0.07 to correctly identify a true effect of size d = 0.2.

All participants in our study followed/liked at least one Facebook page. Overall, 81% of people with PD and 100% of caregivers followed at least one page related to PD, which were commonly associated with fitness/exercise, PD-specific foundations, and healthcare services (see **Table 3**). People with PD followed an average of 12 (SD = 18, Range: 0-72) PD-related pages, and caregivers followed an average of 15 (SD = 20, Range: 1-59) PD-related pages. We found no difference in the number of PD-related pages followed between people with PD and caregivers (W=61.5, p = .74, d = .15). This analysis has a statistical power of 0.07 to correctly identify a true effect of size d = 0.2.

PD-Relevant Post Identification

We developed a binary classifier using 6,732 labeled text posts (2,274 relevant) from 14 people with PD and 5 caregivers. The corpus was split into 80% for training and 20% for testing. Optimized models included Support Vector Machine (SVM) with RBF kernel, K-Nearest Neighbor (KNN) in two versions (KNN-balanced: balanced class weight and KNN-uniform: uniform class weights), two Decision Tree (DT) models based on entropy (DT-balanced: balanced class weights; DT-non: no class weights), Naive Bayes (NB), two Random Forest (RF) variants (RF-balanced: balanced class weights; RF-none: no class weights), Adaboost Classifier, Gaussian Process, and Multi-Layer Perceptron (MLP). We performed 10-fold cross-validation and trained an ensemble model that combined all individual models. RF-balanced was the

highest-performing classifier with 50 trees, achieving a recall of 0.88 (95% CI: 0.87-0.89) (see **Table 4**).

An ablation study on RF-balanced showed improved performance by excluding the post perspective feature and data from a LRRK2+ caregiver (see **Table 5**). The optimized RF-balanced classifier achieved a recall of 0.89 (accuracy = 0.89 (95% CI: 0.88-0.90), precision = 0.88 (95% CI: 0.88-0.89), F_1-score = 0.88 (95% CI: 0.88-0.89)) on the test data.

Following the ablation study, we conducted an ROC curve analysis on the test data to further evaluate the optimized RF-balanced classifier's discriminative power. The ROC curve, detailed in **Figure 1**, illustrates the trade-off between sensitivity and specificity across different thresholds, with an AUC score of 0.94, indicating a highly effective classifier with strong predictive performance.

This classifier was refit with the full dataset for deployment.

PD-relevant Posts

Out of 88,786 Facebook posts analyzed, 2,233 (2.5%) were PD-relevant. Over 91% of participants (14 people with PD and 6 caregivers) discussed PD-specific topics. Nearly 70% (11 people with PD and 5 caregivers) explicitly referenced PD in at least one post. On average, people with PD posted about PD 108 times (SD = 199, range 1-667), which increased from an average of 10 posts pre-diagnosis to 98 posts post-diagnosis. Caregivers posted about PD topics an average of 120 times (SD = 177, range 3-427). There was no significant difference in the number of posts between people with PD and caregivers (W=45, p = .84, d = 0.06). Approximately 32% of posts by people with PD and 29% by caregivers were shared in specific Facebook groups, with no significant difference between people with PD and caregivers (W =

46.5, p = .74, d = 0.06). These analyses have a statistical power of 0.07 to correctly identify a true effect of size d = 0.2.

Qualitative Analysis of Facebook Activity Among Patients and Caregivers

The qualitative analysis of Facebook activities, as shown in a word cloud (**Figure 2**), depicts varied experiences of people with PD and their caregivers. People with PD share experiences about symptoms, treatments, and disease impact, including falls, dyskinesia, FOG, and sometimes premorbid symptoms. Caregivers post about their experiences in supporting loved ones, discussing caregiving challenges like fatigue and patient needs, as well as their roles in community engagement and research. These narratives from patients and caregivers provide a comprehensive view of the lived experiences with PD on social media.

Patient Narratives

> *I had DBS surgery in 2020 and suddenly developed freezing of gait in 2022. I take Rytary but I don't feel any better. Has anyone else experienced this and what did the doctor add?*
>
> - Post by a woman with PD in a DBS support group

Some individuals with PD actively shared detailed accounts of their experiences, including symptoms, treatment regimens, and the impact of PD on their daily lives. This sharing extended to participation in PD-related events and clinical research, underscoring a robust online community that supports information exchange and mutual encouragement.

The posts analyzed revealed clear mentions of disease progression indicators, such as falls, dyskinesia, and FOG. These narratives not only highlight the personal challenges faced by individuals with PD but also offer a window into the reality of disease progression as experienced by patients.

In several cases, premorbid symptoms were identifiable, offering potential insights into the early stages of PD. The mention of symptoms like sleep issues, dysarthria, and postural instability prior to diagnosis suggests that social media data could play a role in recognizing early signs of PD.

Here is a summary of the insights gained from each patient in the study. Notably, the volume and variety of information disclosed differed significantly across patients.

Patient 1. A woman diagnosed with PD documented some of her experiences through 27 Facebook posts, with 4 pre-diagnosis posts focused on exercise. Post-diagnosis, her engagement shifted towards a broader PD narrative, including her Deep Brain Stimulation (DBS) surgery, challenges with balance and FOG, and participation in PD community events. Notably, she experienced at least one fall resulting in injury and seeks physical therapy for pain management. Her word cloud highlights a transition from a focus on physical activity to a broader engagement with PD, including both the struggles and improvements encountered. Key terms such as 'gait', 'balance', and 'tumble' indicate fall experiences, while 'improvement' and 'delight' reflect positive outcomes. Conversely, 'pain' and 'symptom' denote the expression of PD-related difficulties (see **Figure 3a**).

Patient 2. This individual, diagnosed with PD, shared his journey through 6 posts, all after his diagnosis. The content and thematic analysis, illustrated by his word cloud (**Figure 3b**), underscore a deep engagement in PD advocacy, particularly with groups such as the Parkinson's Foundation. The prominence of 'donation' within the word cloud, along with a manual review of the posts, reveals a strong commitment to fundraising and advocacy efforts. Further insights from his postings highlight the significant role his wife plays within these organizations, including her contributions as a speaker at a symposium. This narrative showcases the patient's

active role in PD awareness and the collaborative support within his family for the PD community.

Patient 3. This individual's extensive interaction with PD, documented through 181 posts, offers a comprehensive view into his life with the condition. A notable pre-diagnosis post described an unusual gait, hinting at early movement symptoms. Post-diagnosis, the narrative broadens to encompass a variety of experiences and topics: personal reflections on diagnosis, challenges with falls, exercise, insomnia, balance issues, dystonia, rigidity, and medication management. His contributions extend beyond personal narratives to active participation in and admin roles in PD-specific Facebook groups, offering support to others.

His engagement is multifaceted, involving advocacy, participation in PD-related organizations and events, and participation in clinical trials and research. He even documents his daily experiences within trials. The word cloud (**Figure 3c**) and thematic analysis reveal a balance of positive and negative experiences, characterized by terms like 'gratitude', 'love', and 'hope', alongside 'suffer', indicating the complexity of living with PD. His involvement with organizations such as the American Parkinson Disease Association (APDA), discussions about PD, and references to the FDA underscore a commitment to community building, awareness, and navigating the landscape of PD management and support.

Patient 4. This individual shared two posts concerning his experience with PD, both occurring after his diagnosis. In the initial post, they convey optimism, sharing a piece of information they hope could contribute to others' healing journeys. The subsequent post is a heartfelt expression of appreciation for his caregiver, highlighting the supportive role his daughter plays in his life. This narrative underscores the patient's hopeful outlook and the significance of familial support in managing PD.

Patient 5. This individual, diagnosed with PD prior to creating his Facebook account, has shared 24 posts focused on his PD journey. His word cloud (**Figure 3d**) and post analysis reveal a unique aspect of his treatment and personal empowerment: boxing. Notably, boxing emerges as a significant element of his therapy, markedly improving his condition and self-confidence. Further insights from his postings indicate successful outcomes with DBS therapy, notably in reducing tremors and the need for medications. Additionally, engaging in art has served as a form of physical therapy, contributing to his overall well-being. This account reflects a multifaceted approach to PD management, emphasizing the importance of physical activity, innovative treatments, and creative expression in fostering a positive outlook and improving quality of life.

Patient 6. This individual with PD shared a single post after her diagnosis where they discuss the challenges faced with eating after incorporating amantadine into her existing carbidopa/levodopa medication regimen. This narrative sheds light on the complexities of managing PD symptoms and the side effects associated with medication adjustments, highlighting the patient's ongoing journey to navigate treatment efficacy and quality of life.

Patient 7. This individual with PD has shared a total of 116 posts concerning her PD journey. Notably, 83 of these were made before her diagnosis, as illustrated in **Figure 4a**, with the remaining 33 posts, depicted in **Figure 3e**, occurring post-diagnosis. The analysis of these posts, informed by word cloud visualizations, reveals distinct thematic focuses before and after diagnosis.

Pre-diagnosis posts primarily engage with topics such as fall events, confusion, exercise, migraines, and sleep issues like fatigue. The presence of 'parkinsons_disease' in the pre-diagnosis word cloud stems from a post dedicated to general healthcare advocacy.

Post-diagnosis, the narrative shifts to include expressions of gratitude, anxiety, continued focus on exercise, and new challenges such as pain, fall events, and sleep disturbances, including fatigue and insomnia. Notably, after her diagnosis, there is an absence of explicit mentions of PD in her posts. Despite being the most active user on Facebook within this study, this individual's high volume of activity does not correlate with more detailed disclosures about living with PD, suggesting a nuanced approach to sharing personal health information online.

Patient 8. This person with PD authored a total of 5 posts relevant to her PD experience. Among these, one was shared prior to her diagnosis, focusing on fatigue, while the remaining four were posted after her diagnosis. These post-diagnosis contributions cover a range of topics: they detail personal experiences with pain, actively seek information about PD research, offer support to others within the PD community, and engage in fundraising efforts for PD research.

Patient 9. Patient 9 has shared 431 posts related to her PD journey, with 8 preceding and 423 following her diagnosis. Her narrative encompasses a complex array of experiences including an initial misdiagnosis with dopamine responsive dystonia (DRD), undergoing DBS surgery, and navigating medication side effects.

The pre-diagnosis word cloud (**Figure 4b**) indicates her engagement with exercise amidst challenges with sleep issues and pain. Post-diagnosis, the word cloud (**Figure 3f**) highlights continued commitment to exercise, successful DBS therapy, and feelings of hope. Notably, themes of family, children, and love emerge prominently, reflecting her personal journey of explaining mobility limitations to her child and celebrating post-DBS achievements, such as walking unaided for two days at a theme park.

This account provides a detailed exploration of the lived experience with PD, from initial symptoms and misdiagnosis to therapeutic triumphs and the impact on family life. She is also an

admin for a PD-related Facebook group. For comprehensive medical insights derived from her data, refer to the Facebook-derived medical narrative section.

Patient 10. This individual contributed 18 posts relevant to her PD experience. The majority, 16 posts, were shared before her diagnosis, with the remaining 2 occurring post-diagnosis. The pre-diagnosis word cloud (**Figure 4c**) showcases her active lifestyle, highlighting activities such as hiking, ziplining, gardening, and soccer. Among these, one post specifically mentions fatigue.

Following her diagnosis, the narrative shifts slightly. While one post continues to emphasize exercise, another post explores her initial engagement with the online PD community, questioning the suitability of a group for someone newly diagnosed.

Patient 11. This individual with PD shared 14 posts pertaining to her experience with the condition. Of these, 13 posts predate her diagnosis, all focusing on exercise (**Figure 4d**). Following her diagnosis, she made a single post inquiring about strategies to improve hand mobility issues.

Patient 12. This patient with PD has contributed 21 posts related to his PD journey. Prior to his diagnosis, he shared 14 posts (**Figure 4e**) focused on his experience with shoulder surgery, aimed at enhancing mobility. These posts often sought prayers and acknowledged the support received. After his diagnosis, documented in 7 posts (**Figure 3g**), his narrative broadened to include a variety of themes: humor and interactions with his children regarding PD, and efforts towards better health through increased exercise and weight loss to boost mobility and manage comorbid conditions. This progression reflects a shift from addressing specific physical challenges to embracing a holistic approach to health and well-being post-PD diagnosis.

Patient 13. This patient, diagnosed with PD, has been remarkably active online, sharing a total of 667 posts since his diagnosis, all reflecting his engagement with the PD community. As an administrator of a PD-related Facebook group, he leverages this platform for support and information exchange. His involvement is highlighted in his word cloud (**Figure 3h**), showing deep connections with PD organizations, fitness networks, and discussions on clinical trials, fitness routines, medication, and support groups.

His posts cover a wide range of topics: from navigating insurance complexities for treatment and managing pain to fundraising and providing updates on group members. He actively seeks community support for others, sharing his personal challenges with balance, tremors, sleep disturbances, and anxiety, for which he is medicated. Additionally, he discusses the use of Lee Silverman Voice Treatment (LSVT) for speech issues and shares his experiences with medication "wearing off". Throughout, he expresses profound gratitude for the solidarity and support found within the PD community.

Patient 14. This individual with PD has made two posts following his diagnosis, focusing on expressing gratitude. These posts celebrate and thank family members for their participation in a half marathon to raise funds for PD research. The patient's messages highlight a sense of appreciation for the support and efforts of loved ones in contributing to the fight against PD.

Facebook-derived Medical History Case Study of Patient 9

In this case study, we examine a 20-year medical history of a woman with PD meticulously chronicled through her Facebook posts made over the course of 14 years (some data were extrapolated from before she had a Facebook).

The narrative commences in her early 20s, marked by the onset of muscle spasms in her right foot and a subsequent period of misdiagnoses and ineffective treatments. Her symptoms spread throughout her body, and incapacitated her to the point that she needed to use a wheelchair. She only regained her mobility through vigorous exercise. This period was particularly challenging, characterized by frustration and dissatisfaction with the medical care received.

During her early 30s, she was diagnosed with PD, a diagnosis that was later revised to Dopamine-responsive Dystonia (DRD) during her pregnancy. She primarily used carbidopa/levodopa (Sinemet), augmented by other medications like Cymbalta and Mirapex, to manage her symptoms. The effectiveness of these treatments varied, with noted decreases in efficacy postpartum and during premenstrual periods.

Despite these challenges, she maintained an active lifestyle. This included regular physical activities like Crossfit, cycling, yoga, and running, which she often cited as beneficial for symptom management. The postpartum period brought additional struggles, such as insomnia, fatigue, and increased muscle spasms, necessitating further adjustments in her medication regimen.

Over time, she experienced a decline in treatment effectiveness, leading to symptoms like FOG and dyskinesias. She also had difficulties swallowing. During emergencies, she used Apokyn, despite its side effects, such as severe nausea, to maintain mobility.

A significant turning point in her medical journey was her decision to undergo DBS surgery in her forties. The DBS proved to be a positive intervention, enhancing her mobility and reducing the severity and frequency of 'off periods'. However, the postoperative phase was not

without challenges; she experienced speech side effects, blurred vision, and discomfort from the DBS battery pack. She also experienced a fall.

In the ensuing years, she faced other health issues, including hair loss post-COVID and leg cramps, while continuing to adapt her lifestyle and work to accommodate her condition. Her resilience is further exemplified by her decision to start her own business and engage in part-time work.

Her story is also one of community engagement and advocacy. She co-authored a book and actively sought and shared information within the PD and DRD communities. Furthermore, her family history is notable: her mother and sister were also diagnosed with DRD.

Caregiver Narratives

> *...It is difficult to watch my dad suffer and caregiving is exhausting…Ever since he got out of rehab, he needs 24/7 care and my mom is not strong enough to do it…So, my sister quit her job and now takes care of my dad at my parent's house from 9 am to 8 pm during the weekday…I work full-time, but I only live 3 doors down from my parents. So, I get my dad up in the morning and put him to bed on the weekends…Overall, the most heartbreaking though is my dad's changing personality. He is irritable and angry that he lost his independence and control. The psychological burdens are hard on my dad and hard on us. For now, we are just taking it one day at a time and celebrating the little victories.*
>
> - Post by a caregiver in a caregiver support group

The Facebook posts of caregivers of individuals with PD reveal a multifaceted experience encompassing advocacy, support, and personal involvement in managing the disease. These

caregivers, spouses or children of the patients, share a wide range of experiences: from participating in PD-focused organizations and fundraising events to providing detailed updates on their loved ones' health and treatment regimens. They also express personal challenges and experiences, such as caregiver fatigue and proactive health management for those at genetic risk for PD. Collectively, their narratives underscore the integral role of caregivers in the PD community, highlighting their commitment to both supporting their loved ones and engaging in broader community and research efforts related to PD.

Here is a summary of the insights gained from each caregiver in the study. Notably, the volume and variety of information disclosed differed significantly across caregivers.

Caregiver 1. She, a caregiver and wife of a person with PD (Patient 2), has shared 244 posts related to PD, as depicted in **Figure 4a**. The analysis of her word cloud reveals active engagement with the Parkinson's Foundation, fundraising efforts, and discussions on being a care partner, underscored by expressions of love and support, frequently accompanied by the hashtag #itswhywefight. Her involvement with the Parkinson's Foundation is notable, serving as a leading fundraiser for Movement Day. She provides unwavering support for her husband, sharing updates on his exercise routines, participation in tele-therapy during the pandemic, and challenges on days when his medication wears off. In comparison to her husband, also a subject in this study, she contributes much more to Facebook discussions about PD, offering a comprehensive view of the caregiver's perspective and advocacy within the PD community.

Caregiver 2. This caregiver and wife of an individual with PD (Patient 5) has contributed 5 posts related to PD. In these posts, she expresses gratitude for the support received for her loved one, highlights attendance at PD-specific boxing exercise events—a passion of her husband's—and provides medical updates on his condition. Additionally, she offers support by

sharing insights into her loved one's experiences with PD, underscoring the collaborative and supportive role she plays in navigating his journey with PD.

Caregiver 3. This caregiver and wife of a patient with PD (Patient 12) has shared 21 posts related to PD. Analysis of her word cloud (**Figure 4b**) reveals frequent discussions of her husband's health and wellness, focusing on weight management, cognitive function, and experiences with dementia. These posts detail their joint journey towards weight loss to enhance their quality of life. She provides updates on his exercise routines, his recovery from shoulder surgery, hospital stays, and how his improved health contributes to better PD management. Additionally, she expresses gratitude for the high-quality care her husband has received for his PD, highlighting the positive impact of supportive care on their lives.

Caregiver 4. This caregiver, who is the daughter of Patient 12, has made three PD-related posts. Her contributions encompass general information on speech pathology, a call for writing cards to support individuals with illness, and insights into adaptations for sleep issues, including her parents sleeping in separate beds to better manage these challenges.

Caregiver 5. This caregiver, who is the wife of Patient 14, has shared 19 posts related to PD. Analysis of her word cloud (**Figure 4c**) indicates active involvement with the Parkinson's Foundation, efforts in fundraising through donation solicitations, and themes of gratitude and empowerment. She frequently discusses research developments and community engagement. Her posts predominantly highlight her husband's physical activities, express optimism for finding a PD cure, recount experiences with local rehabilitation centers, and seek community connections for her husband to engage in golf with other individuals with PD. Additionally, it is revealed that their daughter is pursuing an advanced degree focused on PD research, a journey

that fills them with pride. This narrative showcases a comprehensive approach to coping with PD, emphasizing physical well-being, hope, community support, and familial achievements.

Caregiver 6. This caregiver, who has shared 427 PD-related posts, navigates a unique intersection of roles as both a caregiver to her father with PD and as an individual genetically predisposed to the disease due to the LRRK2 mutation. Her engagement spans from sharing her family's PD experiences—encompassing her father, uncle, and the legacy of a grandfather who passed before her birth—to her own proactive health management and research participation.

The narrative weaves through the complexities of caregiver fatigue, the intricacies of managing family members' PD symptoms and treatments, and her personal journey with genetic testing, annual neurologist visits, and early PD indicators like reduced arm swing. Additionally, her commitment to PD research is evident in her decision to register as a brain donor and her active dissemination of PD study information. A word cloud (**Figure 4d**) underscores themes of research, genetic insights, and personal and familial health narratives, encapsulating her comprehensive engagement with the PD community on both a personal and a scientific level.

Discussion

Given Facebook's global reach and its consistent usage across different age groups in North America, our study aimed to understand its relevance for individuals with PD and their caregivers. Facebook has 2.4 billion users as of 2019 (Marengo et al., 2020) and is a significant tool for maintaining social connections, especially for older adults (Gil-Clavel & Zagheni, 2019).

Our findings show that 81% (71-92%) of the successfully contacted participants had Facebook accounts, and 87% (77-97%) were willing to share their data. This substantial presence confirms Facebook's utility as a data collection tool in clinical research, particularly for PD.

However, these figures might be inflated due to our recruitment methods. Initially, we recontacted individuals previously involved in our research, who may already be predisposed to participating. Additionally, recruitment at the Emory Brain Health Center Movement Disorder Clinic involved physicians making the first contact with potential participants, and some refusals were not reported, which could further bias the results.

Facebook Account Activity

Our analysis indicates that both people with PD and their caregivers are actively using Facebook. Excluding outliers, both groups showed consistent Facebook engagement. However, including outliers revealed that while caregivers consistently used Facebook actively, some individuals with PD had varied activity levels.

The least active Facebook user was a woman with PD, averaging 9 activities per year, and a man with PD who averaged 12 activities annually. The woman did not engage in any PD-related activity on Facebook. Despite his low activity levels, the man used Facebook for PD advocacy and fundraising, showcasing diverse usage among individuals with PD. On the other end of the spectrum, the most active participant averaged nearly 27,000 activities per year but never disclosed her diagnosis or discussed PD explicitly. This variation highlights that Facebook activity levels do not uniformly reflect how individuals use the platform to learn and share about PD.

Our findings, particularly when excluding outliers, suggest that both people with PD and their caregivers engage actively on Facebook, with no significant differences in their activity levels. However, due to the small sample size and limited study power, further research with larger samples is needed to validate these conclusions.

PD-specific Groups and Pages

Our study reveals that both individuals with PD and their caregivers actively use Facebook, particularly in PD-specific groups and pages. Facebook Pages primarily serve as channels for unidirectional information dissemination, suitable for businesses, public figures, or healthcare organizations. In contrast, Facebook Groups offer a more interactive space, facilitating discussions and information exchange among members with shared interests.

We found that 56% of people with PD and 71% of caregivers joined at least one PD-related group, with both groups averaging five PD-related groups each. This engagement spans various themes, including fitness, PD support, demographic-specific communities, DBS, support/information, and PD advocacy.

These findings align with broader trends in social media use for health-related purposes. Martínez-Pérez et al. (2015) highlighted the substantial presence of Facebook groups dedicated to neurodegenerative diseases, including PD, emphasizing the platform's role in fostering self-help, advocacy, awareness, and fundraising.

Additionally, our study found that 81% of people with PD and all caregivers follow at least one PD-related Facebook page. The distinct functionalities of Facebook Pages and Groups help cater to the diverse informational and interactive needs of the PD community.

The thematic analysis of group and page titles reveals the PD community's interests, ranging from medical and fitness information to advocacy and social support. However, relying on titles for thematic categorization may miss the depth of interactions and content within these groups. This surface-level analysis overlooks nuanced dynamics and unobservable activities. Thus, while the identified themes provide a snapshot, they represent only part of the complex support and information exchange in online PD-related activities.

Posts Related to PD

We found that individuals with PD and their caregivers use Facebook to share PD-related content, ranging from minimal sharing to detailed medical histories spanning over two decades. Common themes include personal experiences with diagnosis, symptoms, and treatment, as well as participation in PD community events. Caregivers share their unique experiences, discussing the emotional and physical challenges of caregiving.

These findings align with research by Bayen and colleagues (2021), who noted that the main themes of posts in PD forums were symptom management and sharing illness experiences. Similarly, studies summarized by Al-Busaidi and Alamri (2020) across various social media platforms, including YouTube, Twitter, and Facebook, indicate that individuals with PD create online content to discuss their condition. Al-Busaidi (2017) further found that a significant portion of PD-related YouTube content is generated by personal users, suggesting that people with PD engage with and contribute to social media platforms. Patel and colleagues (2015) reviewed studies that reported clinical outcomes from social media use and found that social media can provide social, emotional, and experiential support to individuals with chronic diseases, and that this support may improve patient care.

Our qualitative analysis offers insights into the lived experiences of people with PD and their caregivers, highlighting the diverse impact of PD on daily life and the importance of support systems. The narratives reflect how PD affects patients over time and demonstrate the value of social media for sharing experiences and seeking support.

A limitation of our study is the potential non-exhaustiveness of our keyword Term Dictionary. Despite iterative refinements, some PD-relevant posts may have been missed if they

fell outside the predefined search parameters, leading to false negatives. Future improvements, such as fuzzy word matching for misspelled words, may uncover additional insights.

Additionally, population-level insights from social media must be approached cautiously, as they are limited to individuals with internet access and digital literacy. These factors vary across demographics and are emerging social determinants of health. In individuals with PD, digital literacy may deteriorate due to disease-related cognitive impairments (as reviewed by Esper et al., 2024). Therefore, it is crucial to consider these limitations in the context of the source population.

Markers of Disease Progression on Facebook

Our study explored the feasibility of detecting early markers of disease progression through text data analysis. We found that it is possible to identify mentions of key PD progression markers, such as fall events, medication on-off periods, dyskinesias, and FOG, within the textual data provided by individuals with PD and their caregivers on social media platforms.

This underscores the potential of NLP techniques in monitoring and understanding PD progression from patient-generated content. However, a critical limitation is our inability to determine whether these symptoms were reported to healthcare practitioners or merely shared within online communities. This gap highlights the need for further research to understand the alignment between self-reported symptoms on social media and clinically observed symptoms.

This limitation also suggests an untapped potential for social media as an adjunct tool for healthcare providers. Future studies could explore the co-occurrence of mentions of progression markers in patient-generated content and their electronic health records. Additionally, research

could investigate how healthcare providers might integrate social media insights into clinical practice to enhance patient care and identify early markers of disease progression.

PD, Menstruation, and Pregnancy

Our study identified rare events and complications among individuals with PD. Notably, Patient 9, a young-onset PD patient, reported decreased efficacy of both her medications and DBS therapy during menstruation, which aligns with previous findings (Tolson et al., 2002). However, literature on reduced DBS efficacy during menstruation is scarce.

Patient 9 also experienced diminished medication effectiveness following pregnancy, requiring adjustments during pregnancy. Pregnancy's impact on PD management is complex, with some studies indicating worsened PD symptoms and a persistent change from the pre-pregnancy baseline (Shulman et al., 2000).

Given the rarity of menstruation and pregnancy cases in PD patients, social media monitoring offers a unique opportunity to gather detailed insights into these impacts. This approach can enhance our understanding of PD management in these specific contexts, where traditional research methods may be limited by the scarcity of data.

Premorbid Symptoms on Facebook

Our analysis revealed that it is feasible to identify premorbid symptoms of PD within Facebook narratives shared before an official diagnosis. Many posts, primarily related to exercise, also referenced symptoms such as rigidity, bradykinesia, fatigue, insomnia, and decreased arm swing, which can precede a PD diagnosis by up to a decade (Bloem et al., 2021).

While these isolated references are not diagnostic, they could yield valuable population-level insights into early medical issues that precede a formal PD diagnosis. Aggregated data may provide indicators suggestive of PD.

The duration over which individuals had maintained their Facebook accounts, averaging over 13 years, facilitated our findings, offering a unique opportunity for longitudinal disease monitoring. Notably, one participant had been diagnosed before the advent of Facebook in 1998, and another, a caregiver, shared extensively about her experiences as a LRRK2 carrier, likely to develop PD herself.

As we move forward, most individuals developing conditions later in life, including PD, will have some form of online presence. This era of extensive longitudinal data offers unprecedented opportunities for understanding the early markers and progression of conditions like PD.

No difference in Quantity of Facebook Activity Between Patients and Caregivers

Our analysis revealed no statistically significant differences in Facebook activity between patients with PD and their caregivers. This activity includes general Facebook usage and PD-specific engagement, such as overall volume of Facebook activity, number of PD-related groups joined and pages followed, and frequency of PD-related posts.

It is important to note that our study may be underpowered, indicating that these preliminary findings should be interpreted with caution. To conclusively address the question of behavioral differences between patients and caregivers on Facebook, further research with a larger sample size is necessary.

Conclusion

Our findings indicate that people with PD and their caregivers engage with Facebook by joining groups and following pages that resonate with their experiences of living with PD. This engagement extends to sharing personal narratives related to PD, encompassing aspects such as disease progression, rare symptoms, and premorbid symptoms. Furthermore, Facebook serves as

a pivotal source for community support for both individuals with PD and caregivers. Caregivers share insights into their loved ones' PD journeys alongside their own experiences, challenges, and coping mechanisms as caregivers.

Although the extent of information shared by individuals varied significantly, the aggregate data on Facebook offer valuable insights into the lived experiences of those affected by PD and their caregivers. The successful development and implementation of a text classifier for automatically identifying relevant posts facilitate large-scale monitoring of these data. This approach holds the potential to provide a deeper understanding of the PD community's needs, experiences, and challenges, highlighting the role of social media as a crucial tool in the landscape of healthcare research and patient support.

Acknowledgements

We extend our heartfelt gratitude to all the individuals with Parkinson's disease and their caregivers who generously donated their Facebook data for this research project. Your contributions have been invaluable, and your willingness to share your personal experiences is deeply appreciated. This research could not have been possible without your support. Thank you also to Dr. Kenzie Prichard for providing graphics to accompany instructions for downloading Facebook data from an Android phone.

Tables and Figures

Table 1

Demographic Information of Study Participants

Variable	Caregivers (N=7)	People with PD (N=16)	p value
Age, years			0.07
Mean (SD)	55.3 (15.2)	64.9 (8.8)	
Range	25.7 - 70.0	45.1 - 80.8	
Gender			0.11
Men	1 (14%)	8 (50%)	
Women	6 (86%)	8 (50%)	
Race			0.15
African American/Black	0 (0%)	4 (25%)	
White	7 (100%)	12 (75%)	
Ethnicity			0.33
Not Hispanic or Latino	7 (100%)	14 (88%)	
Unknown / Not Reported	0 (0%)	2 (12%)	
Education			0.78
Completed high school	0 (0%)	2 (12%)	
Completed junior college	1 (14%)	2 (12%)	
Completed college	4 (57%)	7 (44%)	
Completed graduate degree	2 (29%)	5 (31%)	
PD Duration, years			
Mean (SD)		8.7 (5.7)	
Range		1.3 - 25.1	
MDS UPDRS Part 1			
Mean (SD)		16.1 (6.1)	
Range		4.0 - 26.0	
N-Miss		2	

MDS UPDRS Part 2	
Mean (SD)	15.2 (9.0)
Range	4.0 - 31.0
N-Miss	2
MDS UPDRS Part 4	
Mean (SD)	7.6 (3.8)
Range	0.0 - 12.0
N-Miss	2

Note. Statistical comparisons were done with a t-test. Demographic information was self-reported via a REDCap survey. Abbreviations: MDS UPDRS = Movement Disorder Society United Parkinson's disease Rate Scale, PD = Parkinson's disease.

Table 2

Thematic Clusters of Groups joined by PwPD and Caregivers

Group Themes	PwPD	CG
Caregivers of a Specific Demographic	0	1
Caregiver Support	0	1
DBS Support	4	0
DBS Information	2	1
Dystonia Support	4	0
General Fitness/Exercise	14	5
LRRK2	0	1
PD Social Events	3	3
PD Advocacy	3	5
PD Art	1	0
PwPD of a Specific demographic	5	2
PD-specific Fitness	3	1
PD Fundraising	0	1
PD Support	22	10
PD Information	7	0
PD Other	0	1
PD-related Publications	1	1
PD Foundations	1	0
Research	5	3
YOPD DBS Support	1	0
PwYOPD of a Specific Demographic	1	0
YOPD Support	4	0

Note. This table reports the frequency of PD-related group membership stratified by participant role and group theme.

Table 3

Thematic Clusters of Pages Followed/Liked by People with PD and Caregivers

Page Themes	PwPD	CG
Caregiver Resources	0	4
PD-specific Health Publications	6	3
Exercise/Fitness	45	15
Famous People with PD	6	3
Government Health Organizations	9	2
Political Health Advocacy	0	2
Healthcare/Medical Services	13	7
Healthcare Practitioners	13	5
Insurance	2	1
Medical Institutes/Departments	8	6
Medical Research	7	12
Mental Health Awareness	0	1
Other	7	3
PD Art	1	1
PD Community	8	7
PD-specific Fitness	10	3
Information about PD	7	4
PD-specific Page	10	5
PD-specific Foundations	30	15
PD Support Groups	5	3
Senior Services/Accessibility	1	2

Note. This table reports the frequency of followed PD-related pages stratified by participant role and group theme.

Table 4

Classifier Performance on Training Data

Model	Recall (95% CI)	Precision (95% CI)	F_1-Score (95% CI)
RF-balanced	**0.88 (0.87-0.89)**	**0.88 (0.87-0.89)**	**0.88 (0.87-0.88)**
RF-none	0.88 (0.87-0.88)	0.88 (0.87-0.88)	0.87 (0.86-0.88)
Ensemble	0.88 (0.87-0.88)	0.88 (0.87-0.89)	0.87 (0.86-0.88)
SVM	0.85 (0.84-0.86)	0.85 (0.84-0.86)	0.85 (0.84-0.86)
AdaBoost	0.84 (0.83-0.85)	0.85 (0.84-0.86)	0.83 (0.82-0.84)
MLP	0.84 (0.83-0.85)	0.84 (0.83-0.85)	0.84 (0.83-0.85)
DT-none	0.82 (0.82-0.83)	0.83 (0.82-0.84)	0.83 (0.81-0.84)
DT-balanced	0.82 (0.80-0.83)	0.82 (0.81-0.83)	0.82 (0.81-0.83)
KNN-distance	0.76 (0.74-0.77)	0.75 (0.74-0.76)	0.74 (0.73-0.75)
KNN-uniform	0.75 (0.74-0.76)	0.75 (0.73-0.76)	0.73 (0.72-0.75)
Naïve Bayes	0.62 (0.61-0.64)	0.78 (0.77-0.79)	0.62 (0.61-0.64)

Note. This table summarizes the performance of various classifiers used to develop a binary classifier for identifying relevant text posts from people with PD and their caregivers. The metrics shown include Recall, Precision, and F_1-Score, each with their respective 95% confidence intervals (CI). The classifiers evaluated were Random Forest with balanced class weights (RF-balanced) and without class weights (RF-none), an Ensemble model combining all individual models, Support Vector Machine with RBF kernel (SVM), AdaBoost, Multi-Layer Perceptron (MLP), Decision Tree with balanced class weights (DT-balanced) and without class weights (DT-none), K-Nearest Neighbor with distance weighting (KNN-distance) and uniform weighting (KNN-uniform), and Naive Bayes (NB). The RF-balanced model, using 50 trees, was the highest-performing classifier, achieving a recall of 0.88 (95% CI: 0.87-0.89).

Table 5

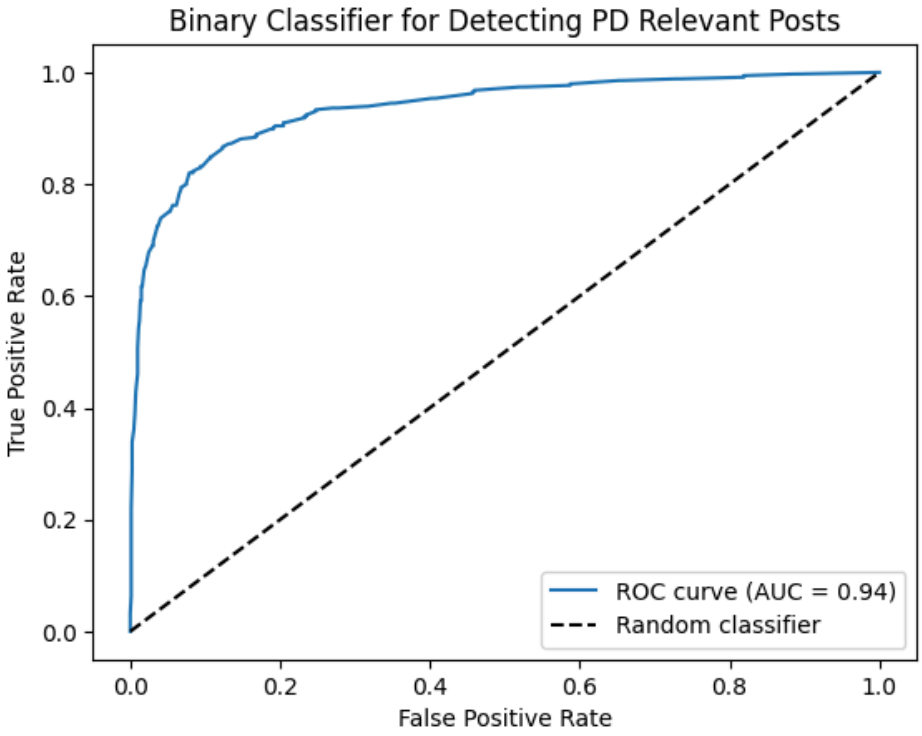
Ablation Study to Determine Feature Importance of Random Forest Classifier

Dropped Feature	Recall (95% CI)	Precision (95% CI)	F_1-Score (95% CI)
none	0.88 (0.87-0.89)	0.88 (0.87-0.89)	0.88 (0.87-0.88)
1,2,3-grams	0.77 (0.76-0.78)	0.79 (0.78-0.80)	0.78 (0.76-0.79)
Disease status	0.87 (0.87-0.88)	0.87 (0.87-0.88)	0.87 (0.86-0.88)
Gender	0.87 (0.86-0.88)	0.87 (0.86-0.88)	0.87 (0.86-0.88)
Age at time of posting	0.85 (0.84-0.86)	0.85 (0.84-0.86)	0.85 (0.84-0.86)
Perspective (first person, second person)	0.88 (0.87-0.89)	0.88 (0.87-0.89)	0.87 (0.86-0.88)
LRRK+ caregiver	**0.89 (0.88-0.90)**	**0.89 (0.88-0.90)**	**0.88 (0.88-0.89)**
LRRK+ caregiver + perspective	**0.89 (0.88-0.90)**	0.89 (0.88-0.89)	**0.89 (0.88-0.89)**

Note. This table presents the performance metrics of the RF-balanced classifier after dropping specific features. The metrics shown include Recall, Precision, and F1-Score, each with their respective 95% confidence intervals (CI). The features tested were 1,2,3-grams, disease status, gender, age at the time of posting, perspective (first person, second person), LRRK2+ caregiver, and a combination of LRRK2+ caregiver and perspective. The optimized RF-balanced classifier, excluding the post perspective feature and data from a LRRK2+ caregiver, achieved a recall of 0.89 (95% CI: 0.88-0.90), precision of 0.89 (95% CI: 0.88-0.90), and an F1-score of 0.88 (95% CI: 0.88-0.89) on the test data.

Figure 1

ROC Curve for Binary Classifier Detecting PD-Relevant Posts



Note. The Receiver Operating Characteristic (ROC) curve illustrates the performance of the random forest binary classifier developed to identify Parkinson's Disease (PD)-relevant posts from Facebook data. The curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings. An Area Under the Curve (AUC) of 0.94 indicates a high degree of accuracy in distinguishing between relevant and irrelevant posts, showcasing the classifier's effectiveness. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The point on the curve closest to the top left corner represents the optimal threshold for maximizing both sensitivity and specificity.

Figure 2

Word Clouds of PD-relevant Text Shared on Facebook

Note. This figure depicts two word clouds: one derived from PD-related posts authored by people with PD (tulip) and another derived from PD-related posts authored by caregivers of people with PD (hands). The tulip is a symbol of hope for people with PD.

Figure 3

Word clouds of PD-related Posts Authored After Diagnosis by People with PD Stratified by Individual.

Note. Each panel in this figure represents the top 100 most frequent words and phrases used by each patient following their PD diagnoses.

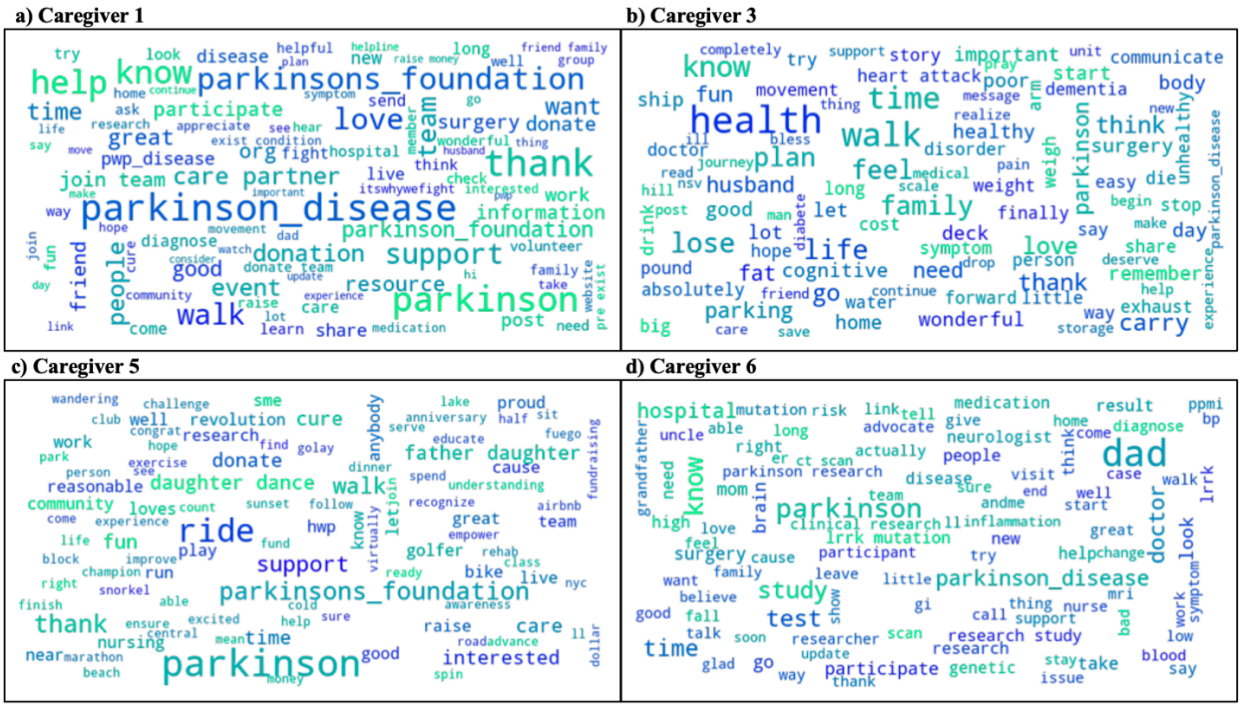
Figure 4

Word clouds of PD-related Posts Authored Before Diagnosis by People with PD Stratified by Individual

Note. Each panel in this figure represents the top 100 most frequent words and phrases used by each patient before their PD diagnoses.

Figure 5

Word clouds of PD-related Posts Authored by Caregivers Stratified by Individual



Note. Each panel in this figure represents the top 100 most frequent words and phrases used by each caregiver.

References

Al-Busaidi, I. S. (2017). Qualitative analysis of Parkinson's disease information on social media: The case of YouTube™. *European Association for Predictive, Preventive and Personalised Medicine*, *8*, 273–277. https://doi.org/10.1007/s13167-017-0113-7

Al-Busaidi, I. S., & Alamri, Y. (2020). Parkinson's disease and social media. In *Diagnosis and Management in Parkinson's Disease* (pp. 125–138). Academic Press. https://doi.org/10.1016/B978-0-12-815946-0.00008-9

Armstrong, M. J., & Okun, M. S. (2020). Diagnosis and treatment of Parkinson disease: A review. *Journal of the American Medical Association*, *323*(6), 548. https://doi.org/10.1001/jama.2019.22360

Bayen, S., Carpentier, C., Baran, J., Cottencin, O., Defebvre, L., Moreau, C., Devos, D., & Messaadi, N. (2021). Parkinson's disease: Content analysis of patient online discussion forums. A prospective observational study using Netnography. *Patient Education and Counseling*, *104*(8), 2060–2066. https://doi.org/10.1016/j.pec.2021.01.028

Bloem, B. R., Okun, M. S., & Klein, C. (2021). Parkinson's disease. *The Lancet*, *397*(10291), 2284–2303. https://doi.org/10.1016/S0140-6736(21)00218-X

Dorsey, E. R., Elbaz, A., Nichols, E., Abbasi, N., Abd-Allah, F., Abdelalim, A., Adsuar, J. C., Ansha, M. G., Brayne, C., Choi, J.-Y. J., Collado-Mateo, D., Dahodwala, N., Do, H. P., Edessa, D., Endres, M., Fereshtehnejad, S.-M., Foreman, K. J., Gankpe, F. G., Gupta, R., … Murray, C. J. L. (2018). Global, regional, and national burden of Parkinson's disease, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, *17*(11), 939–953. https://doi.org/10.1016/S1474-4422(18)30295-3

Dudina, V., Judina, D., & Platonov, K. (2019). Personal illness experience in Russian social

media: Between willingness to share and stigmatization. In S. El Yacoubi, F. Bagnoli, & G. Pacini (Eds.), *Internet Science* (pp. 47–58). Springer International Publishing. https://doi.org/10.1007/978-3-030-34770-3_4

Esper, C. D., Valdovinos, B. Y., & Schneider, R. B. (2024). The importance of digital health literacy in an evolving Parkinson's disease care system. *Journal of Parkinson's Disease*, 1–9. https://doi.org/10.3233/JPD-230229

Gil-Clavel, S., & Zagheni, E. (2019). Demographic differentials in Facebook usage around the world. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 647–650. https://doi.org/10.1609/icwsm.v13i01.3263

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., … LaPelle, N. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders*, *23*(15), 2129–2170. https://doi.org/10.1002/mds.22340

Macchi, Z. A., Koljack, C. E., Miyasaki, J. M., Katz, M., Galifianakis, N., Prizer, L. P., Sillau, S. H., & Kluger, B. M. (2020). Patient and caregiver characteristics associated with caregiver burden in Parkinson's disease: A palliative care approach. *Annals of Palliative Medicine*, *9*(Suppl 1), S24-S2S33. https://doi.org/10.21037/apm.2019.10.01

Marengo, D., Sindermann, C., Elhai, J. D., & Montag, C. (2020). One social media company to rule them all: Associations between use of Facebook-owned social media platforms, sociodemographic characteristics, and the big five personality traits. *Frontiers in*

Psychology, *11*, 936. https://doi.org/10.3389/fpsyg.2020.00936

Martínez-Pérez, B., De La Torre-Díez, I., Bargiela-Flórez, B., López-Coronado, M., & Rodrigues, J. J. (2015). Content analysis of neurodegenerative and mental diseases social groups. *Health Informatics Journal*, *21*(4), 267–283. https://doi.org/10.1177/1460458214525615

Patel, R., Chang, T., Greysen, S. R., & Chopra, V. (2015). Social Media use in chronic disease: A systematic review and novel taxonomy. *The American Journal of Medicine*, *128*(12), 1335–1350. https://doi.org/10.1016/j.amjmed.2015.06.015

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Powell, J. M., Guo, Y., Sarker, A., & McKay, J. L. (2023). Classification of fall types in Parkinson's disease from self-report data using natural language processing. In J. M. Juarez, M. Marcos, G. Stiglic, & A. Tucker (Eds.), *Artificial Intelligence in Medicine* (Vol. 13897, pp. 163–172). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-34344-5_20

Shulman, L. M., Minagar, A., & Weiner, W. J. (2000). The effect of pregnancy in Parkinson's disease. *Movement Disorders*, *15*(1), 132–135. https://doi.org/10.1002/1531-8257(200001)15:1<132::AID-MDS1020>3.0.CO;2-C

Tolson, D., Fleming, V., & Schartau, E. (2002). Coping with menstruation: Understanding the needs of women with Parkinson's disease. *Journal of Advanced Nursing*, *40*(5), 513–521. https://doi.org/10.1046/j.1365-2648.2002.02408.x

Appendix A: JSON Files

Our analysis utilized JSON exports to detail Facebook activities of study participants, capturing a range of interactions from posts and comments to group memberships and page interactions. The structure of these JSON files varied slightly, with each action on the platform documented, including a 'timestamp' for each entry. User-generated text was also captured, albeit under varying keys such as 'name,' 'description,' and 'text,' necessitating individual file review for accurate data extraction.

From the Posts export, several JSON files were acquired, including your_posts_check_ins_photos_and_videos, which records a diverse array of user activities on Facebook. Other notable files include archive, documenting archived posts; your_videos, tracking video uploads; and various files related to photo albums and uncategorized photos, each providing entries for photos posted, with or without albums, and inclusive of timestamps and possible captions.

The Comments and Reactions export yielded a JSON file documenting every user comment outside of groups, alongside files for likes and reactions to posts and comments. This dataset offers a broad view of user engagement and social interaction on the platform.

From the Groups Export, we gathered files documenting activities within Facebook groups. These include anonymous_posts-you've_written for anonymous group posts; group_posts and_comments for posts and comments within groups; your_answers_to_membership questions detailing responses to group membership queries; Your_comments_in_groups for comments within groups; your_group_membership_activity logging group joins and leaves; Your_groups listing user-created groups; and Your_pending_posts_in_groups for posts pending admin approval.

The Pages Export provided insights into page interactions. Pages_and_profiles_you_follow and Pages_you've_liked files detail the pages and profiles followed and liked by users, often overlapping due to automatic liking upon following. Your_pages records pages created by the user.

This collection of JSON files forms a comprehensive dataset for analyzing Facebook user behavior, highlighting the wide-ranging nature of social media activities without delving into the implications or interpretations of this data.

Posts:

Your posts, check-ins, photos, and videos

Sharing on One's Own Timeline: albums, celebrations, collections, episodes, events, groups, life events, links, live videos, locations, marked themselves safe, marketplace posts, memories, moments from their year, notes, pages, photos, posts, products, profiles, question answering, quotes, read books, recommendations, reviews, statuses, trending topics, videos
Sharing on Friends' Timelines: albums, events, fundraisers, links, memories, photos, posts, profiles, videos

Appendix B: Term Dictionary

term_dictionary = {

 # premorbid symptoms

'constipation': {'intestinal sluggishness', 'bowel', 'digestive issues', 'bowel issues', 'struggle in the bathroom', 'irregular bowel movements', 'gastrointestinal issues', 'slow digestion', 'bowel discomfort', 'constipation', 'blocked up', 'irregular bowels', 'lack of fiber', 'infrequent stools', 'stomach pain', 'intestinal discomfort', 'laxative', 'infrequent bowel movements', 'bowel trouble', 'digestive upset', 'fecal impaction', 'poop', 'colon concerns', 'stool issues', 'difficulty passing stool', 'gut health concerns', 'constipated', 'cramping', 'abdominal bloating', 'stool inconsistency', 'slow bowel transit', 'digestive problems', 'poor bowel habits', 'need more fiber', 'stool problems', 'straining to defecate', 'stomach cramps', 'infrequent defecation', 'digestive slowdown', 'laxative use', 'stool softener', 'digestive distress', 'digestive blockage', 'irregularity', 'bowel movement trouble', 'bowel obstruction', 'hard stools', 'bowel difficulty', ‘stool’},

'rem sleep behavior disorder': {'dream acting', 'thrashing in bed', 'bed partner disturbance', 'sleep aggression', 'acting out while asleep', 'nighttime episodes', 'sleep movement', 'dream aggression', 'sleep disruptions', 'nocturnal activity', 'sleep behavior disorder', 'kicking in sleep', 'sleepwalking', 'dream enactment', 'rem sleep behavior disorder', 'punching in sleep', 'sleep walking', 'dream reenactment',
'violent sleep behavior', 'sleep disturbance', 'vivid dreaming', 'sleep activity', 'aggressive nocturnal behavior', 'night terrors', 'restless nighttime behavior', 'abnormal sleep behavior', 'restless

sleep', 'nighttime restlessness', 'nighttime agitation', 'violent dreaming', 'talking in sleep', 'unusual sleep activities', 'agitated sleep', 'disruptive sleep behavior', 'acting out dreams', 'physical dream expression', 'hitting during sleep', 'vocalization in sleep', 'active dreaming', 'physical activity during sleep', 'rbd', 'sleep talking', 'violent dreams', 'sleep-related violence', 'sleep interruptions'},

'hyposmia': {"can't smell", 'faint odors', 'hyposmia', 'loss of smell', 'smell impairment', 'weak sense of smell', 'smelling issues', 'reduced smell'},

'asymmetric vague shoulder pain': {'irregular shoulder pain', 'shoulder pain', 'shoulder discomfort', 'shoulder stiffness', 'uneven shoulder ache', 'occasional shoulder ache', 'intermittent shoulder pain', 'shoulder pain on one side', 'shoulder pain that comes and goes', 'one-sided shoulder pain', 'asymmetrical shoulder pain', 'shoulder hurts', 'one shoulder pain', 'asymmetric vague shoulder pain', 'shoulder tenderness', 'shoulder pain without injury', 'random shoulder pain', 'shoulder ache', 'shoulder soreness', 'shoulder pain without explanation', 'unexplained shoulder soreness', 'unequal shoulder pain', 'one-sided shoulder discomfort', 'shoulder pain without cause', 'mild shoulder ache', 'vague pain in shoulder'},

'depression': {'emotional numbness', 'sadness', 'feeling down', 'mental', 'hopelessness', 'loss of interest', 'chronic unhappiness', 'depression', 'lack of interest', 'depressed', 'hopeless', 'persistent sadness', 'feeling worthless', 'feeling blue', 'lack of pleasure', 'melancholy', 'low spirits', 'low mood', 'sad', 'feeling empty', ‘funk’},

'impaired color vision': {'impaired color vision',"can't see colors", 'trouble with colors', 'color vision deficiency', 'colors seem faded', 'color blindness', 'dull colors'},

'erectile dysfunction': {'erectile dysfunction', 'impotence', 'sexual dysfunction', 'trouble with erection', 'sexual issues', 'intimacy issues', 'ed', "can't get hard"},

'reduced arm swing': {'reduced arm swing', 'less arm movement', "arm doesn't swing", 'arm rigidity', 'arm swing imbalance', 'one arm swing', 'stiff arm walk', 'arm stiffness'},

'increased stride time variability': {'increased stride time variability', 'irregular stride', 'uneven walking', 'gait variability', 'stride irregularities', 'change in walk pattern', 'walking issues'},

'urinary dysfunction': {'urinary dysfunction', 'bladder issues', 'bladder control issues', 'urinary incontinence', 'leaking urine', 'frequent urination', 'urination problems', 'urine', 'pee', 'bladder problems', 'urinary issues', 'bladder dysfunction', 'incontinence', 'bladder control problems', 'frequent urination', 'peeing issues'},

'pain': {'pain treatment', 'pain management', 'body aches', 'pain control', 'constant pain', 'sharp pain', 'debilitating pain', 'dull ache', 'shooting pain', 'muscle pain', 'throbbing pain', 'piercing pain', 'twinge of pain', 'migraine', 'nerve pain', 'post-surgical pain', 'neck pain', 'discomfort', 'ache', 'soreness', 'tender pain', 'muscular discomfort', 'unbearable pain', 'excruciating pain', 'inflammation pain', 'visceral pain', 'stiffness', 'painful sensation', 'mild pain', 'chronic pain', 'moderate pain', 'cramping', 'arthritis pain', 'neuropathic pain',

'tenderness', 'subacute pain', 'breakthrough pain', 'burning sensation', 'unrelenting pain', 'painful symptoms', 'numbness and pain', 'pain relief', 'back pain', 'pain therapy', 'spasms', 'pain episodes', 'stabbing pain', 'pain flare-up', 'widespread pain', 'joint pain', 'radiating pain', 'pain', 'painful swelling', 'acute pain', 'aching joints', 'severe pain', 'nagging pain', 'headache', 'hurt'},

'insomnia': {'insomnia', 'frequent waking', 'sleepless', 'sleep problems', 'sleep disorder', 'trouble sleeping', 'sleep issues', "can't sleep", 'sleep disturbances', 'wakeful nights', 'sleep disruption', 'insomnia', 'nighttime awakenings', 'restless sleep', 'sleep', 'wide awake'},

'anxiety': {'anxiety', 'nervousness', 'stress', 'nervous', 'tense', 'worrying', 'anxious feelings', 'overthinking', 'anxious', 'restlessness'},

'cognitive impairment': {'cognitive impairment', 'mental fog', 'memory loss', 'brain fog', 'confusion', 'cognitive decline', 'forgetfulness', 'difficulty concentrating'},

'fatigue': {'fatigue', 'worn out', 'sleepy', 'tiredness', 'exhaustion', 'lack of energy', 'low energy', 'cranky', 'weary', 'exhausted', 'bed', 'knocked out'},

 # motor symptoms

'bradykinesia': {'bradykinesia', 'sluggish motion', 'movement slowness', 'slow movement', 'delayed movements', 'reduced speed of movement', 'slow to move'},

'rigidity': {'rigidity', 'muscular rigidity', 'tight muscles', 'hard to move', 'inflexible joints', 'rigid muscles', 'stiff muscles', 'muscle stiffness'},

'tremors': {'tremors', 'tremor', 'trembling hands', 'hand tremble', 'shaking', 'shaky movements', 'body shakes', 'involuntary shaking'},

'bulbar_dysfunction': {'bulbar dysfunction'},

'dysarthria': {'slurred speech', 'unclear speech', 'difficulty speaking', 'speaking problems', 'mumbling', 'slow speech', 'hard to understand', 'speech impairment', 'struggling to talk', 'garbled speech', 'weak voice', 'speaking fatigue', 'difficulty articulating', 'speech difficulties', 'trouble pronouncing', 'stuttering', 'slurring words', 'hard to speak', 'speech disorder', 'jumbled speech', 'strained voice', 'nasal speech', 'speech changes', 'choppy speech', 'talking difficulties', 'incoherent speech', 'hard to verbalize', 'stammering', 'voice changes', 'altered speech', 'speech problem'},

'dysphagia': {'trouble swallowing', 'difficulty swallowing', 'swallowing problems', 'hard to swallow', 'painful swallowing', 'choking on food', 'gagging when eating', 'swallowing difficulty', 'food getting stuck', "can't swallow", 'throat discomfort', 'swallowing pain', 'difficulty eating', 'swallowing discomfort', 'feeling of food stuck', 'coughing when eating', 'hard swallowing', 'swallowing trouble', 'swallowing disorder', 'issues with swallowing', 'food aspiration', 'difficulty chewing', 'pain when swallowing', 'hard to eat', 'eating

difficulties', 'food sticking in throat', 'sore throat while eating', 'fear of choking', 'trouble eating', 'chewing problems'},

'on-off periods': {'medication wearing off', 'symptom fluctuations', 'off periods', 'on periods', 'dose wearing off', 'motor fluctuations', 'levodopa wearing off', 'medication not working', 'medication off time', 'increased symptoms', 'unpredictable symptom control', 'erratic symptom relief', 'medication cycle fluctuations', 'dyskinesia', 'medication on time', 'inconsistent medication effect', 'motor symptom variability', 'levodopa-induced dyskinesia', 'deteriorating medication effect', 'fluctuating response', 'wearing-off phenomenon', 'drug-induced motor complications', 'medication-related motor issues', 'variable symptom control', 'periods of poor mobility', 'good on time', 'bad off time', 'medication response fluctuations', 'peak dose dyskinesia', 'delayed on', 'sudden off', 'dose failure', 'end-of-dose wearing off', 'early morning off', 'medication-related mobility changes'},

'postural instability': {'postural instability', 'difficulty standing upright', 'balance problems', 'unsteady standing', 'instability standing', 'poor balance', 'falling easily', 'balance issues', 'loss of balance', 'unsteady', 'balance problems', 'imbalance', 'dizziness', 'leaning', 'staggering', 'difficulty standing', 'standing issues', 'balance difficulties', 'wobbling', 'swaying', ‘balance’},

'falls': {'falls', 'falling', 'collapsed', 'lost balance', 'serious fall', 'fall scare', 'wiped out', 'toppled over', 'hurt myself falling', 'tumbled', 'slip', 'tripping', 'stumbling', 'dizziness', 'unsteady',

'loss of balance', 'leaning', 'imbalance', 'grabbing', 'bruise', 'catch', 'caught', 'took a spill', 'sudden fall', 'lost my balance', 'staggered', 'caught myself falling', 'knocked down', 'hurt', 'dropped', 'stumbled', 'ground', 'terrain', 'trip', 'near fall', 'pitched forward', 'help', 'rolled over', 'stairs', 'shuffle', 'accident', 'almost fell', 'plummeted', 'stumble', 'fallen', 'collapse', 'recovery', 'floor', 'surface', 'seat', 'face plant', 'frequent falling', 'shower', 'hit the ground', 'walking', 'fell', 'slid', 'slipped', 'crashed', 'toppled', 'slipping', 'falling down', 'railing', 'injury', 'tumble down', 'care', 'unexpected fall'},

'gait difficulties': {'gait difficulties', 'gait disturbance', 'unsteady walk', 'abnormal walk', 'walking difficulty', 'shuffling steps', 'trouble walking', 'walking issues', 'mobility issues', 'shuffling gait', 'freezing of gait', 'gait freezing', 'shuffling', 'dragging feet', 'walking impairment', 'uneven gait', 'irregular gait', 'walking instability', 'festinating gait', 'slow gait', 'gait asymmetry', ‘unpredictability of my legs’},

'assistive device use': {'mobility aids', 'assistive devices', 'walking aids', 'use of cane', 'using walker', 'wheelchair use', 'mobility scooter', 'assistive walking devices', 'supportive devices', 'orthotic devices', 'braces', 'walking stick', 'rollator', 'grab bars', 'handrails', 'adaptive equipment', ‘adaptive chair’, ‘scooter’},

'freezing of gait': {'freezing of gait',"can't move feet", 'gait freeze', 'freezing', 'start hesitation', 'immobilized', 'momentary freeze', "can't step", 'stuck in place', 'movement hesitation', 'freezing episode', "can't lift feet", 'sudden stop', 'motor block', 'frozen gait', 'walking

freeze', 'feet glued', "legs won't move", "can't move", 'temporary paralysis', 'sudden stop walking'},

'dyskinesia': {'dyskinesia', 'involuntary movements', 'uncontrolled movements', 'motor restlessness', 'chorea', 'athetosis', 'muscle twitching', 'jerking movements', 'writhing movements', 'fidgeting', 'restless movements', 'abnormal posturing', 'levodopa-induced dyskinesia', 'drug-induced movement disorder', 'hyperkinesia', 'muscle rigidity', 'unintended muscle movements', 'muscular jerks', 'fluctuating movements', 'jerky movements', 'spontaneous movements', 'unpredictable movements'},

'dystonia': {'dystonia', 'muscular spasms', 'muscle attacks', 'sustained muscle contractivities', 'involuntary muscular contractivities', 'twisting movements', 'abnormal postures', 'muscle rigidity', 'muscular tension', 'muscle stiffness', 'muscle twisting', 'body distortion', 'abnormal muscle tone', 'sustained postures', 'twisting postures', 'repetitive movements', 'abnormal body positions', 'neck spasms', "writer's cramp", 'torticollis', 'muscle cramping', 'limb dystonia', 'focal dystonia', 'generalized dystonia', 'task-specific dystonia', ‘muscle attack’, ‘attack’, “legs wouldn't work”},

'related disorders': {'dystonia', 'drd', 'td', 'essential tremor', 'multiple system atrophy', 'progressive supranuclear palsy', 'ataxia', 'lewy body dementia', 'frontotemporal dementia', 'normal pressure hydrocephalus', 'vascular parkinsonism', 'drug-induced parkinsonism', 'secondary parkinsonism', 'spinocerebellar ataxia', 'myoclonus', 'restless legs syndrome', 'tardive dyskinesia', 'parkinsonism', 'parkinson-plus syndrome', 'bradykinesia', 'postural

instability', 'shy-drager syndrome', 'benign essential tremor', 'parkinsonian gait', 'akathisia', 'psychogenic movement disorder', 'neuroleptic malignant syndrome', 'olivopontocerebellar atrophy', 'striatonigral degeneration', 'paraneoplastic syndromes'},

 # non-motor symptoms

'autonomic dysfunction': {'autonomic dysfunction', 'autonomic issues', 'irregular heartbeat', 'dysautonomia', 'temperature regulation problems'},

'orthostatic hypotension': {'orthostatic hypotension', 'blood pressure issues', 'postural hypotension', 'light-headedness on standing', 'fainting spells', 'low blood pressure', 'sudden dizziness', 'dizzy standing'},

'altered sweating': {'excessive sweating', 'sweating too much', 'hyperhidrosis', 'increased sweating', 'sweat attacks', 'sweat profusely', 'heavy sweating', 'drenching sweats', 'night sweats', 'sweaty palms', 'clammy skin', 'sweating episodes', 'anhidrosis', 'lack of sweat', 'reduced sweating', 'dry skin', 'non-sweating', 'difficulty sweating', 'sweating abnormalities', 'sweating disorder', 'heat intolerance', 'sweating irregularities', 'sudden sweating', 'sweat excessively', 'sweating imbalance', 'overactive sweat glands', 'underactive sweat glands', 'sweating fluctuation', 'sweating disturbance', 'sweating dysfunction', 'sweat gland issues', 'sweating problem', 'sweating difficulty', 'abnormal perspiration', 'uncontrolled sweating', 'profuse perspiration', 'excessive perspiration', 'no sweat', 'lack of perspiration', 'failed sweat response'},

'psychosis': {'psychosis', 'psychotic symptoms', 'reality distortion', 'delusions', 'hallucinations', 'seeing things', 'irrational thoughts', 'hallucination', 'hearing voices', 'paranoia', 'delusional thinking', 'visual hallucinations', 'auditory hallucinations', 'paranoid delusions', 'false beliefs', 'distorted reality', 'psychotic episode', 'schizophrenia-like symptoms', 'psychotic break', 'unreal perceptions', 'psychotic disorder', 'paranoid thinking', 'disorganized thinking', 'psychotic depression', 'psychotic behavior', 'bizarre delusions', 'grandiose delusions', 'persecutory delusions'},

'mental health': {'emotional challenges', 'mental health', 'mental toll', 'emotional well-being', 'psychological distress', 'mental health issues', 'emotional distress', 'mental strain', 'psychological issues', 'emotional health', 'mental health struggles', 'anxiety', 'depression', 'stress', 'mental fatigue', 'emotional strain', 'psychological well-being', 'mental health concerns', 'emotional problems', 'mental health care', 'mental health support', 'psychological support', 'emotional support', 'psychological health', 'mental resilience', 'mental well-being'},

'cognitive decline': {'cognitive decline', 'forgetful', 'cognitive deterioration', 'thinking issues', 'memory loss', 'brain fog', 'losing cognition', 'mental decline', 'confused', 'memory lapses', 'forgetting things', 'short-term memory loss', 'forgetfulness', 'memory issues', "can't remember", 'slow to respond', 'delayed cognition', 'brain slowing', 'sluggish thought', 'mental slowness', 'thinking delay', 'dementia', 'major forgetfulness', 'significant cognitive decline', 'advanced memory issues', 'severe memory loss', 'cognitive changes', 'mental confusion', 'disorientation', 'cognitive impairment', 'mental fog', 'cognitive difficulties',

'mental deterioration', 'declining memory', 'cognitive loss', 'mental fuzziness', 'cognitive dysfunction', 'memory deterioration', 'mental fogging', 'cognitive decline in aging', 'memory decline', 'cognitive slowing', 'cognitive struggles'},

'addiction issues': {'compulsive', 'impulse control disorder', 'overeating', 'gambling', 'sex addiction', 'substance abuse', 'drug addiction', 'alcohol addiction', 'addictive behavior', 'compulsive behavior', 'addictive tendencies', 'binge eating', 'problem gambling', 'sexual compulsivity', 'drug abuse', 'alcohol abuse', 'chemical dependency', 'addictive habits', 'compulsive gambling', 'compulsive eating', 'addiction problems', 'habitual overeating', 'pathological gambling', 'substance dependence', 'excessive gambling', 'excessive sexual behavior', 'drug dependency', 'alcoholism', 'opioid addiction', 'prescription drug abuse', 'narcotic abuse', 'compulsive shopping', 'internet addiction', 'addictive personality'} ,

'drooling': {'drooling', 'excessive saliva', 'salivating', 'sialorrhea', 'dribbling saliva', 'constant drooling', 'mouth drooling', 'uncontrolled saliva', 'drool', 'spitting', 'saliva control problems', 'saliva management', 'drooling issues', 'drooling at night', 'saliva accumulation', 'salivary control', 'drooling problems'},

 # treatment

'medication': { 'medication', 'meds', 'entacapone', 'inbrija', 'rytary', 'anticholinergics', 'duopa', 'comt inhibitors', 'dopamine agonists', 'gocovri', 'opicapone', 'mao-b inhibitors', 'ongentys', 'istradefylline', 'tasmar', 'artane', 'nourianz', 'benztropine', 'xadago', 'ropinirole', 'mirapex', 'rotigotine', 'trihexyphenidyl', 'eldepryl', 'amantadine', 'neupro', 'comtan', 'pramipexole',

'rasagiline', 'l-dopa', 'symmetrel', 'tolcapone', 'sinemet', 'carbidopa', 'selegiline', 'safinamide', 'levodopa', 'osmolex er', 'zelapar', 'requip', 'cogentin', 'started taking', 'levadopa', 'carbodopa', 'new meds', 'l-dopa', 'impax', 'mao b inhibitor', 'drug', 'levodopa', 'azilect', 'baclofen', 'nuplazid', 'clonazepam', 'apokyn', 'dopamine', ‘pill’, ‘med’},

'treatment': {'treatment', 'doctor', 'physician', 'therapist', 'neurologist', 'therapy', 'occupational therapist', "physician's assistant", 'physical therapist', 'treatments', 'speech therapist', 'hospital', 'healing', 'dr', 'pt', 'exercise physiologist', 'speech therapy', 'therapy sessions', 'physiotherapy', 'alternative treatment', 'neurologist appointment', 'treatment options', 'symptom management', 'treatment side effects', 'specialist visit', 'holistic approaches', 'disease stage', 'md', 'mph', 'dpt', 'doctor', 'symptom', 'speech issues', 'lsvt', ‘rehab’, ‘nursing facility’, ‘nurse’, ‘neuro’, ‘doc’, ‘doctors’},

'dbs': {'dbs', 'deep brain stimulation', 'dbs surgery', "brain surgery", 'neurosurgery', 'brain stimulation therapy', 'dbs therapy', "parkinson's surgery", 'neurostimulator', 'dbs procedure', 'dbs treatment', "surgical treatment for parkinson's", 'brain stimulation treatment', 'dbs implant', 'neurological surgery', 'dbs device', 'brain pacemaker', "surgical options for parkinson's", 'electrical stimulation brain', 'dbs benefits', 'dbs risks', 'brain stimulation surgery', 'dbs outcomes', 'neurosurgical procedure', 'implanting dbs device', ‘surgery’},

'exercise' : {'exercise routine', 'yoga', 'exercise', 'boxing', 'dance', 'ride', 'bike', 'tango', 'bicycling', 'walk', 'cycling', 'elliptical', 'tai chi', 'zumba', 'gym', 'workout', 'running', 'jogging',

'swimming', 'aerobics', 'cardio workout', 'strength training', 'weight lifting', 'fitness class', 'pilates', 'spinning', 'kickboxing', 'hiking', 'crossfit', 'bodybuilding', 'marathon training', 'trail running', 'rock climbing', 'rowing', 'spin class', 'boot camp', 'stretching', 'calisthenics', 'power walking', 'kettlebell workout', 'circuit training', 'hiit', 'interval training', 'aqua aerobics', 'water aerobics', 'personal training', 'physical fitness', 'sports', 'badminton', 'tennis', 'basketball', 'soccer', 'football', 'volleyball', 'skiing', 'snowboarding', 'skateboarding', 'surfing', 'mountain biking', 'barre', 'rollerblading', 'skating', 'sprinting', 'parcour', 'parkour', 'outdoor activities', 'functional training', 'group fitness', 'balance exercises', 'mobility exercises', 'muscle building', 'endurance training', '5k', '10k', 'race', ‘move’, ‘moving’, ‘interval training’},

 # organizations

'organizations': {'fundraising', "parkinson's foundation", "parkinson's foundation symposium", 'michael j. fox', 'fox', 'foundation', 'apda', 'neuro challenge', 'winning round foundation', 'wilkins parkinson', 'brian grant foundation'},

 # other

'quality of life': {'diet changes', 'care routine', 'going natural', 'daily challenges', 'quality of life', 'new normal', 'disease milestones', 'day-to-day', 'recent milestone', 'disease progression', 'work adjustments', 'driving issues', 'social isolation', 'home adjustments', 'illness impact', 'living with disease', 'disability'},

'community': {'support group', 'community support', 'peer support', 'patient forum', 'community involvement', 'social support', 'support network', ‘support’, ‘moving day’},

'diagnosis': {'diagnosis', "late-onset parkinson's", 'yopd',"early-onset parkinson's", "diagnosed with parkinson's", "parkinson's diagnosis", 'confirming diagnosis', 'diagnostic journey', ‘diagnosis’},

'caregiver terms': {'caregiver support', 'caregiving challenges', 'spousal caregiver', 'family caregiver', 'caregiver journey', 'caregiver experience', 'caring for spouse', 'caring for parent', 'caregiver'},

'research': {'clinical trial', 'medical research', 'research study', "parkinson's research", 'clinical study', 'research updates', 'research breakthroughs', 'medical trial', ‘trial’ },

'advocacy': {'patient advocacy', 'raising awareness', "parkinson's awareness", 'health advocacy', 'advocacy efforts', 'advocating for patients', 'disease awareness', 'community advocacy'},

‘other': {'challenge', "fight against parkinson's", 'travel concerns', 'disease', 'handicap', 'overcoming challenges', 'adapting to illness', 'personal battle', 'living with handicap', ‘illness’, 'support dog', 'support animal', ‘cure’, ‘helping’, ‘help’},

'parkinson': {'parkinson', 'pd', 'pwp', ‘parkies’, ‘parkie’},

'gene': {'lrrk2', 'snca', 'park1', 'park8', 'vps35', 'park17', 'prkn', 'park2', 'pink1', 'park6', 'dj1', 'park7',

'gba'}

}

Chapter 6: General Discussion

Reflecting on the journey of my dissertation, I realized that the core motivations driving my work stem from an enduring fascination with knowledge-based discovery. We are surrounded by a sea of data, more than we could ever hope to navigate. My dissertation represents a foray into harnessing this vast sea, aiming to organize it for insight and develop tools that enable the extraction of knowledge at scales beyond the reach of individual researchers or even entire teams.

In this chapter, I briefly revisit the key findings and insights from each data chapter. Subsequently, I will explore future research directions motivated by these studies.

Chapter Summaries

The initial data chapter of my dissertation, a systematic review, embarked on a manual exploration to reveal the persistent underrepresentation of women in clinical research on PD. This endeavor, involving a year and a half of labor and the concerted efforts of six additional research assistants, provided only a glimpse into a few hundred papers spanning select years of PD research. While insightful, the resource-intensive nature of this approach limits its replicability and scalability, despite the importance of the insights gained.

To mitigate the resource-intensive nature of manual data extraction, the next chapter of my dissertation sought to automate this data extraction process. Utilizing the GPT-4-Turbo model, we demonstrated that it is indeed possible to extract unstructured target information in a structured format with high accuracy from scientific publications. Insights regarding the proportional representation of women in PD research derived from model predictions mirrored those obtained through painstaking manual analysis. This breakthrough suggests a path to scaling

up the monitoring of sex inclusion practices in ongoing research, also offering a way to bridge the gaps between the time periods to which we were initially constrained by practicality.

The third data chapter introduced a classifier designed to dissect a common PD complication—falls—with nuance. By differentiating falls caused by perturbations to one's center of mass from those resulting from other issues, we aim to refine our understanding and prevention of this significant and hazardous complication.

In the final data chapter, we turned to Facebook to explore the extent to which individuals with PD and their caregivers share medically relevant information on the platform. Previous literature hinted at this possibility given the existence of online PD-specific groups and forums (Al-Busaidi, 2017; Al-Busaidi & Alamri, 2020; Bayen et al., 2021). We found that people with PD and their caregivers share medically relevant information on their Facebook accounts, including symptoms, indications of disease progression, and rare complications of the disease.

Future Research

Collectively, my dissertation lays the groundwork for new methodologies to extract, categorize, and analyze information about individuals affected by PD. From analyzing published research to categorizing symptoms with nuance and extracting firsthand accounts of lived experiences from social media, these efforts point to the untapped potential of text data in PD research. In essence, this work establishes the foundation for future research to navigate the vast information landscape of PD, transforming unstructured data into valuable insights that can enhance our understanding and management of the disease. Below are project proposals inspired by this research.

Exploring Historical Inclusion of Women in Research Involving Age-related Conditions

Despite the underrepresentation of women in PD clinical studies, outright exclusion has been rare, even during periods when excluding women from research was more common. This leads us to hypothesize that the historical inclusion of women in PD research is due to most female PD patients being beyond their childbearing years. To test this hypothesis, one could develop a NLP pipeline, like we did in Chapter 3, to analyze sex/gender-related data from clinical research articles on other predominantly geriatric conditions, like Alzheimer's disease.

Further, alternative resources might also be viable. For instance, ClinicalTrials.gov, a registry mandated by the Food and Drug Modernization Act of 1997 and made public in 2000, could be a valuable tool for this analysis. Shah and Saliba (2023) used this resource to investigate gender and ethnic disparities in teledermatology clinical trial participants. ClinicalTrials.gov provides comprehensive information on clinical trials, including disease focus, interventions, eligibility criteria, and in some cases, results and participant demographics. Initially, the registry was exclusive to NIH-funded drug trials, but over the years, the scope has broadened significantly. Requirements for registration have expanded, and the database now includes a vast array of over 400,000 trials, observational studies, and expanded access studies from 221 countries, making it an invaluable resource for analyzing clinical trial trends over time.

Expanding Automated Extraction of Sex-Related Data in Scientific Research

Further studies could extend the applicability of the sex-related data extraction pipeline beyond the realm of PD research to include diverse fields, evaluating its effectiveness in generalizing across various types of scientific literature. An interesting avenue for exploration could involve applying the model to extract sex-related data from research on non-human subjects, providing insights into sex inclusion across a wider spectrum of scientific inquiry.

Additionally, there is a critical need to explore the model's capacity for extracting and analyzing more nuanced variables, such as the methods and outcomes of sex-based analyses within studies. It is important not only to determine whether such analyses are conducted but also to assess their methodological rigor and the validity of their conclusions in relation to sex differences. This deeper level of analysis could uncover insights into the quality of sex-based research at scale and its implications, necessitating a blend of NLP capabilities with domain-specific knowledge. There is evidence in the literature that these models struggle with these higher level tasks (Foppiano et al., 2024), highlighting the need for more research on domain-specific LLMs.

Combining Data from Social Media and Electronic Health Records to Better Understand PD

Future studies should explore the alignment between medical information shared on Facebook and documented in electronic health records (EHRs) to highlight the clinical relevance of social media data and differences between these data sources.

EHRs are rich in both structured and unstructured data provided by clinicians. Social media platforms like Facebook offer rich sources of unstructured, patient-generated data. Integrating insights from EHRs and social media could enable large-scale mapping of PD phenotypes, aiding in understanding the variability in disease presentation and progression. For example, NLP pipelines could sift through large databases to identify and analyze rare cases, such as the effects of pregnancy on young women with PD. By integrating insights from thousands of individuals across various data sources, clinicians could better understand these rare occurrences and improve patient care at a population level.

For scalability, this process requires automation. LLMs, particularly those hosted on HIPAA-compliant platforms like Azure (Adams, 2023) or in-house LLMs, show great potential

for this task. Developing an automated pipeline could provide scalable insights into premorbid symptoms, disease progression, and rare complications. Such a pipeline could streamline analyses similar to those conducted by Mammen et al. (2023), who manually mapped personal symptoms and their impacts. Automation assumes that the struggles people share online, deemed significant enough to post about, can be systematically analyzed and quantified.

Additionally, it is crucial to recognize the origins and biases within the data. Social media reflects varied levels of technological access and engagement, potentially skewing the data pool. Digital literacy among people with PD varies with socioeconomic status and can decline with disease progression (Esper et al., 2024). Similarly, disparities in research representation and healthcare access can bias the insights gained from big data. Ensuring that the diversity of experiences among those living with PD is accurately captured and reflected in research outcomes is essential.

Conclusion

The integration of NLP into clinical care promises to revolutionize our understanding and management of PD. By harnessing the vast amounts of text-based data available, from scientific publications to patient narratives on social media, NLP can provide unprecedented insights into PD. This, in turn, could lead to more personalized, effective, and anticipatory care for individuals with PD. Realizing this potential requires not only technological advancements but also a steadfast commitment to navigating the ethical and logistical challenges associated with data privacy and security.

References

Adams, K. (2023). Epic to integrate GPT-4 into its EHR through expanded Microsoft partnership. *MedCity News*. https://medcitynews.com/2023/04/epic-to-integrate-gpt-4-into-its-ehr-through-expanded-microsoft-partnership/

Al-Busaidi, I. S. (2017). Qualitative analysis of Parkinson's disease information on social media: The case of YouTube™. *European Association for Predictive, Preventive and Personalised Medicine*, *8*, 273–277. https://doi.org/10.1007/s13167-017-0113-7

Al-Busaidi, I. S., & Alamri, Y. (2020). Parkinson's disease and social media. In *Diagnosis and Management in Parkinson's Disease* (pp. 125–138). Academic Press. https://doi.org/10.1016/B978-0-12-815946-0.00008-9

Bayen, S., Carpentier, C., Baran, J., Cottencin, O., Defebvre, L., Moreau, C., Devos, D., & Messaadi, N. (2021). Parkinson's disease: Content analysis of patient online discussion forums. A prospective observational study using Netnography. *Patient Education and Counseling*, *104*(8), 2060–2066. https://doi.org/10.1016/j.pec.2021.01.028

Esper, C. D., Valdovinos, B. Y., & Schneider, R. B. (2024). The importance of digital health literacy in an evolving Parkinson's disease care system. *Journal of Parkinson's Disease*, 1–9. https://doi.org/10.3233/JPD-230229

Foppiano, L., Lambard, G., Amagasa, T., & Ishii, M. (2024). *Mining experimental data from materials science literature with large language models* (No. arXiv:2401.11052). arXiv. http://arxiv.org/abs/2401.11052

Mammen, J. R., Speck, R. M., Stebbins, G. M., Müller, M. L. T. M., Yang, P. T., Campbell, M., Cosman, J., Crawford, J. E., Dam, T., Hellsten, J., Jensen-Roberts, S., Kostrzebski, M.,

Simuni, T., Barowicz, K. W., Cedarbaum, J. M., Dorsey, E. R., Stephenson, D., & Adams, J. L. (2023). Mapping relevance of digital measures to meaningful symptoms and impacts in early Parkinson's disease. *Journal of Parkinson's Disease*, *13*(4), 589–607. https://doi.org/10.3233/JPD-225122

Shah, A., & Saliba, E. (2023). Gender and ethnic disparities in teledermatology clinical trial participants: Cross-sectional analysis of ClinicalTrials.gov-registered trials. *JMIR Dermatology*, *6*, e46031. https://doi.org/10.2196/46031

Appendix A: Sex and Gender in Biomedical Research

The Committee on Measuring Sex, Gender Identity, and Sexual Orientation (2022) defines sex as "a multidimensional construct based on a cluster of anatomical and physiological traits, including external genitalia, secondary sex characteristics, gonads, chromosomes, and hormones." Gender, in contrast, is defined as "a multidimensional construct that intertwines gender identity—a core element of an individual's identity—with gender expression, which is how a person communicates their gender to others through behavior and appearance, such as hairstyle and clothing, as well as cultural expectations regarding social status, characteristics, and behavior associated with sex traits." These definitions highlight the complexity and distinctiveness of each construct, yet their close interrelationship often makes them challenging to distinguish in experimental settings (Eliot et al., 2023). In such instances, using the combined term 'sex/gender' is suitable.

Frequently, researchers regard sex and gender as binary, immutable, and interchangeable constructs (Committee on Measuring Sex, Gender Identity, and Sexual Orientation et al., 2022; Hammarström & Annandale, 2012). This oversimplification may have tangible impacts, such as inaccurate research findings as well as the use of these scientific articles to support laws that advance sex essentialist beliefs (Richardson, 2022). Sex essentialism asserts that sex fundamentally determines immutable traits, behaviors, and roles. This perspective overlooks the diversity and intricacy of sex- and gender-related biology (Richardson, 2022).

The NIH's Sex as a Biological Variable policy (NOT‑OD-15-102) promotes a binary view of sex by mandating the inclusion of materials from “both” (male and female) sexes in research, thereby endorsing a sex essentialist approach (Richardson, 2022). This approach reduces sex to a binary classification, ignoring biological diversity, including intersex conditions

and the variability of sex traits within and across groups (Richardson, 2022). Critics argue that this policy oversimplifies the nuanced relationship between biology and health outcomes, failing to account for the full spectrum of sex-related differences (Committee on Measuring Sex, Gender Identity, and Sexual Orientation et al., 2022; Hammarström & Annandale, 2012; Richardson, 2022).

Moreover, this approach does not account for the principles of sex contextualism, which advocates for a nuanced and context-specific understanding of sex in biomedical research (Richardson, 2022). Sex contextualism emphasizes tailoring the relevance and operationalization of sex-related variables to a study's specific aims and scope, promoting a shift towards more inclusive and precise representations of sex and gender diversity (Eliot et al., 2023; Richardson, 2022). Through the lens of sex contextualism, sex as a study variable serves as an interim proxy for other, not yet pinpointed factors, such as hormones, which more accurately account for variations than sex itself (Maney, 2016; Pape et al., 2024).

Implementing sex contextualism in research, alongside the use of suitable statistical analyses, involves two key steps: 1) clearly defining how sex is measured (for instance, through self-reporting), and 2) identifying and assessing variables like physiological, environmental, societal, or behavioral factors that may covary with sex, aiding in the explanation of observed differences (Eliot et al., 2023). Furthermore, there's a movement towards using multi-dimensional measures for constructs like sex and gender, moving away from simplistic, single-variable approaches. For instance, Nielsen et al. (2021) introduced the Stanford Gender-related Variables for Health Research method. This methodology allows for a comprehensive evaluation of gender through multiple dimensions, facilitating a more nuanced exploration of gender's effects on health beyond a single, categorical framework.

Ultimately, redefining and operationalizing sex and gender in biomedical research demands moving beyond traditional binary frameworks towards a more inclusive and accurate depiction of human diversity. Such an approach not only improves the validity of research findings but also ensures that medical interventions are equitable and responsive to the diverse needs of all individuals.

References

Committee on Measuring Sex, Gender Identity, and Sexual Orientation, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, & National Academies of Sciences, Engineering, and Medicine. (2022). *Measuring Sex, Gender Identity, and Sexual Orientation* (N. Bates, M. Chin, & T. Becker, Eds.; p. 26424). National Academies Press. https://doi.org/10.17226/26424

Eliot, L., Beery, A., Jacobs, E., LeBlanc, H., Maney, D., & McCarthy, M. (2023). Why and How to Account for Sex and Gender in Brain and Behavioral Research. *The Journal of Neuroscience*, *43*, 6344–6356. https://doi.org/10.1523/JNEUROSCI.0020-23.2023

Hammarström, A., & Annandale, E. (2012). A conceptual muddle: An empirical analysis of the use of "sex" and "gender" in "gender-specific medicine" journals. *PloS One*, *7*(4). https://doi.org/10.1371/journal.pone.0034193

Maney, D. L. (2016). Perils and pitfalls of reporting sex differences. *Phil. Trans. R. Soc. B*, *371*, 20150119. https://doi.org/10.1098/rstb.2015.0119

Nielsen, M. W., Stefanick, M. L., Peragine, D., Neilands, T. B., Ioannidis, J. P. A., Pilote, L., Prochaska, J. J., Cullen, M. R., Einstein, G., Klinge, I., LeBlanc, H., Paik, H. Y., & Schiebinger, L. (2021). Gender-related variables for health research. *Biology of Sex Differences*, *12*(1), 23. https://doi.org/10.1186/s13293-021-00366-3

Pape, M., Miyagi, M., Ritz, S. A., Boulicault, M., Richardson, S. S., & Maney, D. L. (2024). Sex contextualism in laboratory research: Enhancing rigor and precision in the study of sex-related variables. *Cell*.

Richardson, S. S. (2022). Sex Contextualism. *Philosophy, Theory, and Practice in Biology*, *14*(0). https://doi.org/10.3998/ptpbio.2096